POLITECNICO DI MILANO
Scuola di Ingegneria Industriale e dell'Informazione
Mathematical Engineering - Statistical Learning

Weighted functional data analysis for partially observed seismic data: An application to ground motion modelling in Italy

Supervisor: Prof. Alessandra Menafoglio

Co-Supervisors: Dr. Riccardo Peli
Dr. Sara Sgobba
Dr. Giovanni Lanzano

Candidate:
Teresa Bortolotti
ID 920432

Academic Year 2020-2021

Abstract

Motivated by the crucial implications of Ground Motion Prediction Equations in terms of seismic hazard analysis and civil protection planning, this work extends to a functional framework the model proposed by Lanzano, Luzi, Pacor, et al. (2019) in the Italian context, for the estimation of ground motion conditionally on a given seismic scenario. In particular, from the inherent characteristic of seismic data to be incomplete over the domain, comes the necessity to develop a novel methodology for the analysis of partially observed functional data. The approach consists in combining pre-existing techniques of data reconstruction with the definition of observation-specific weights. The latter enter the estimation process by defining criteria that give less weight to the errors in the reconstructed parts of the curves, and full weight to those made on the observed values. This work extends the classical methods of *smoothing* and *function-on-scalar regression* to this weighted functional approach. The entire analysis results in a functional model that is effective in capturing the seismological features underlying its formulation, and in producing results that are physically explainable. This model is complementary to the functional geostatistical model proposed by Menafoglio et al. (2020), since the two, combined, allow one to obtain maps of ground shaking that are inserted in a fully functional context.

Keywords: weighted functional analysis, partially observed functional data, ground motion prediction equations

Sommario

Motivato dai risvolti applicativi dei modelli di previsione del movimento del suolo (Ground Motion Prediction Equations) in termini di analisi del rischio sismico e di protezione civile, questo lavoro estende a un contesto di analisi funzionale dei dati il modello proposto da Lanzano, Luzi, Pacor, et al. (2019) per l'Italia, che stima lo scuotimento del terreno condizionatamente alle caratteristiche di un dato evento sismico. In particolare, poiché i dati del caso studio risultano incompleti nel proprio dominio, è proposto un nuovo approccio per l'analisi dei dati funzionali parzialmente osservati. Tale approccio consiste nel combinare le tecniche di ricostruzione dei dati incompleti, presenti in letteratura, con l'introduzione di pesi che siano associati ad ogni osservazione. Questi ultimi intervengono nei processi di stima assegnando un peso minore agli errori commessi sulle parti ricostruite di una curva, e peso pieno agli errori commessi sulle parti di curva effettivamente osservate. Il lavoro estende all'approccio funzionale pesato i metodi classici di *smoothing* e di *regressione funzionale a predittori scalari*. L'intera analisi ha come risultato un modello funzionale efficace nel cogliere le caratteristiche sismologiche alla base della sua formulazione, e capace di produrre risultati interpretabili a livello geofisico. Questo modello è complementare al modello geostatistico funzionale proposto da Menafoglio et al. (2020), dal momento che i due, combinati insieme, consentono di ottenere delle mappe di scuotimento del terreno che in questo modo risultano inserite in un contesto completamente funzionale.

Parole chiave: analisi funzionale pesata, dati funzionali parzialmente osservati, equazioni di previsione del movimento del suolo

Contents

List of Figures

List of Tables

Introduction

The present work is in the context of the assessment of seismic hazard, related to the occurrence of an earthquake and to the ground shaking that the release of elastic energy causes. Through the introduction of intensity measures to quantify ground motion, this area of research, that goes by the name of Probabilistic Seismic Hazard Analysis, aims to determine the probability that a certain risk threshold will be exceeded. In this field, a practical tool for the prediction of ground motion at a site is provided by Ground Motion Prediction Equations (GMPE), that estimate the median value of an intensity measure and the associated uncertainty, conditionally on a set of seismic parameters, including the magnitude of the event, the distance of the site from the epicentre, the faulting mechanism and the characteristics of soil at the site.

As a building can effectively be assimilated to a single-degree-of-freedom oscillator with a natural period of vibration, it is convenient for the assessment of the seismic risk related to architectural damages to quantify ground motion in terms of measures of intensity that are period-dependent. From this inherent characteristic of the intensity measures comes the threefold possibility of embedding the model in a univariate, multivariate or functional setting. In the literature of GMPEs, many ground motion models are proposed in univariate (Bindi et al., 2011, Lanzano, Luzi, Pacor, et al., 2019, Kotha, Bindi, and Cotton, 2016, Boore et al., 2014) and multivariate (Worden et al., 2018, Huang and Galasso, 2019) formulations. Conversely, little work has been put in the development of functional ground motion models (Menafoglio et al., 2020). Compared to univariate or multivariate analysis, a functional approach exploits the natural dependency of the intensity measures on the vibration period, and allows one to shift the focus from a period-specific estimate to its profile over the whole period domain. Additionally, it has the double advantage of considering the correlations among different periods, overcoming the intrinsic limitations of multivariate approaches that suffer from the curse of dimensionality.

This work stems from the need to provide an extension to the framework of Functional Data Analysis (FDA, Ramsay and Silverman, 2005) of the scalar ground

motion model for the median, proposed in Lanzano, Luzi, Pacor, et al. (2019) and hereafter referred to as ITA18. More precisely, a functional estimate for the median is intended to combine with the functional geostatistical model for the residuals of Menafoglio et al. (2020), to allow for a fully functional approach to address the analysis of soil motion. Indeed, the median estimate would complement the *event*- and *site*-related systematic corrective terms for the residuals, identified in Menafoglio et al. (2020), for the formulation of a functional model that includes spatial dependence in the predictions. Eventually, these methods jointly provide a convenient tool for the construction of reliable, period-continuous seismic shaking maps.

Classical techniques of FDA develop under the assumption that the observations are recorded over a common domain. Specifically for the case study analysed in this work, data have the peculiarity of being manually processed, with the result that a non-negligible number of intensity measure profiles are recorded only partly, and not over the whole period domain. Since many methodologies of classical FDA fail for observations of this kind, and the removal of incomplete observations would imply an undesirable loss of information, the necessity arises to develop convenient strategies for the handling of partially observed functional data. The literature on this topic proposes to reconstruct the incomplete data, using the estimates of mean and covariance defined specifically for this context (Kraus, 2015, Kneip and Liebl, 2020). Doing so, the idea is to maximize the information provided by the recorded data, by inferring the values of the missing observations.

The aim of this work is to provide a functional extension of the model for the median proposed in Lanzano, Luzi, Pacor, et al. (2019), by embedding the problem in the framework of partially observed functional data. The benefits of such novel approach are manifold. First, the choice of a suitable functional space enables us to capture the smoothness and the regularity underlying the discrete observations recorded along the grid of period points. Second, the reconstruction of incomplete data avoids the loss of information that the removal of missing values would involve. Third, the extension over the whole domain provides spectra of ground motion for a wider range of vibration periods, consisting in an effective tool for probabilistic seismic hazard assessment.

The intent is pursued by proposing a novel methodology for the handling of incomplete data, that couples each reconstructed curve with a weight function, taking value 1 where the datum is observed and decreasing to 0 the further the reconstruction gets from the last recorded value. The rationale of the method consists in associating less confidence to the parts of a curve that undergo reconstruction, with respect to those that are observed originally. The classical techniques of *penalized*

smoothing and *penalized function-on-scalar regression* are extended to include the weights, which enter the estimation process by defining criteria that give less weight to the errors committed on the reconstructed values of the curves, and full weight to those committed on the observed values. A methodology that quantifies the pointwise variance associated to the estimates is developed in this weighted functional framework, and a bootstrap procedure is outlined for the construction of simultaneous confidence bands for the regression coefficients estimates. The soundness of the weighted methodologies is tested on the data of our case study, by assessing whether the introduction of the weights is effective in stabilizing the results with respect to the adopted reconstruction method. The reconstruction methods over which the technique is tested are those present in the literature, and mentioned above in this introduction. Eventually, the proposed methodologies are applied to data of the Engineering Strong-Motion (ESM) database for Italy (Lanzano, Sgobba, Luzi, et al., 2018), over which ITA18 is calibrated. The results are commented by combining diagnostic techniques with the seismological interpretation of the phenomenon under study.

The dissertation is developed in five chapters, the general structure of which is as follows:

Chapter 1: *A review of Ground Motion Prediction Equations.* This chapter introduces the reader to Ground Motion Prediction Equations. First, it provides an overview of the *state-of-the-art* of ground motion models and defines the context from which this work arises. Then, it presents the quantities involved in the formulation of the equations, as categorized in *source-*, *path-* and *site-*related parameters, and clarifies their seismological meaning. At the end of the chapter, the functional form of the model of Lanzano, Luzi, Pacor, et al., 2019 is introduced and commented.

Chapter 2: *Introduction to Functional Data Analysis.* This chapter consists of two sections. The first reviews the fundamental notions of Functional Data Analysis and provides the formal definitions of functional random variable and functional datum. Separable Hilbert spaces are identified as suitable spaces in which to embed the analysis, and their basic properties are presented. Then, the definitions of mean and covariance operators are introduced, both in a general separable Hilbert space and with a focus on the L^2 embedding. The second section is devoted to the illustration of the *state-of-the-art* techniques employed for the analysis of partially observed functional data. In particular, three methodologies for the reconstruction of incomplete data are presented.

Chapter 3: *Weighted analysis for functional data.* This chapter gives the theoretical and methodological foundations to extend the techniques of *penalized smoothing* and *penalized function-on-scalar regression* to context of weighted functional analysis. The procedure that quantifies the variability associated to the regression coefficients estimates is extended to this context. Finally, a bootstrap procedure is outlined for the construction of simultaneous confidence bands for the functional regression coefficients.

Chapter 4: *Case study: A functional Ground Motion Model for Italy.* This chapter is devoted to a preliminary analysis of the case study. It includes the formulation of the functional extension of ITA18, and an in-depth exploration of the dataset of calibration. Then, the overall workflow is outlined, by completing the methods proposed in Chapter 3 with some crucial intermediate steps of the analysis, concerning the calibration of the penalization parameters for smoothing and regression. Finally, the soundness of the entire procedure is validated through a comparison with *state-of-the-art* techniques for the reconstruction of incomplete data.

Chapter 5: *Case study: Results and diagnostic.* Here, the results of the conducted analysis are discussed. The classical techniques of model diagnostic are combined with a seismological interpretation of the quantities involved in the model, based on the literature on the topic. First, we comment on the estimates of the regression coefficients and conduct an analysis of *goodness-of-fit*. Then, the focus moves on the ground motion predictions and on a sensitivity analysis on the hyperparameters of the model. Finally, the novel functional model is compared to its scalar counterpart, ITA18.

Appendix A: *Theoretical results.* This appendix reports two theoretical arguments that lead to some fundamental results discussed in Chapter 3. Appendix A1 shows the calculations leading to a closed form of the Weighted Functional Penalized Least Squares criterion, employed for the fitting of the regression model. Appendix A2 deploys the educated reasoning behind the estimation of the degrees-of-freedom of the regression model.

Appendix B: *Additional figures.* Here are some additional figures, which aim to complement the information used for the comments on results and diagnostics in Chapter 5.

Appendix C: *Codes*. This appendix briefly explains the structure of the GitHub repository, containing the main `R` scripts and `R` functions that have been implemented for the application of the methodologies proposed in this work.

Chapter 1

A review of Ground Motion Prediction Equations

1.1 Introduction to GMPEs

Engineering seismology serves as linkage between geology and engineering, and arises from the need to exploit the knowledge that comes from earth science for an anti-seismic design of structures. The engineer deals with the trade-off between a costly seismic-resistant design and the risks of economic loss that may occur after an earthquake, *e.g.* in the form of architectural damage in buildings. Therefore, it is crucial for a well-balanced, earthquake-proof planning to assess the seismic hazard and the risk that the motion of the ground exceeds a certain threshold.

As a building can be assimilated to an oscillator with one degree-of-freedom and a certain natural period of oscillation, it is reasonable to quantify the level of ground motion in terms of intensity measures that are defined over a period domain. One common approach to estimate such measures employs Ground-Motion Prediction Equations (GMPE), that infer the median value of ground shaking and its associated uncertainty, conditionally on some parameters, like the magnitude of the earthquake, the style of faulting of the crust, and the epicentral distance of the site.

GMPEs are formulated as sum of a median term and a standard deviation in the form

$$Y = \mu(X, \beta) + \sigma,$$

where Y is typically a logarithmic transformation of the measure of intensity of the ground motion, μ is the median term described by a functional form that provides a simplified representation of the underlying seismological phenomenon, and σ is

the standard deviation. In the estimation for μ, which is going to be discussed later in the chapter, X is the collection of all parameters that are descriptive of the earthquake and of the site conditions, and β is the vector containing the coefficients of the functional form.

It is common finding in the literature that the magnitude of the standard deviation heavily affects the outcomes of probabilistic seismic hazard analysis (Al-Atik et al., 2010). Therefore, a great deal of effort is put in finding a formulation of GMPEs that is associated to lower values of standard deviation. One way to achieve this reduction is through the formulation of non-ergodic models.
Historically, the scarcity of available data has led to the widespread use of models formulated under the ergodic assumption. The latter states that the spatial averages over a single realization of the ground motion parameter of interest converge to the average value over time of the random field at any site (Anderson and Brune, 1999). In more recent years, a greater availability of data provided the conditions to drop this assumption, in the form of a decomposition of the residuals of the model into systematic *source*-, *path*- and *site*-specific effects. In particular, multiple recordings of a number of different earthquakes measured at many different sites made it possible to compute such effects and include them as systematic corrections to the median term in the GMPE. In this new framework, many studies (P.-S. Lin et al., 2011, Anderson and Uchiyama, 2011, Stafford, 2014) have shown how the drop of the ergodic assumption in the formulation of a GMPE implies a reduction of the standard deviation.

In addition to this, recent works have exploited the spatial correlation of the residuals of the model, with the aim of making inference on ground motion values at locations where observations are not available. As the majority of studies adopts univariate (Lanzano, Luzi, Pacor, et al., 2019, Sgobba, Lanzano, et al., 2019) or multivariate (Worden et al., 2018) geostatistical approaches to the analysis of spatial correlation, little to no work has been put to develop a geostatistical approach from the perspective of functional data analysis. The work of Menafoglio et al. (2020) fits into this context. In fact, the authors provide a functional extension of the work of Sgobba, Lanzano, et al. (2019), by handling the *source*-, *path*- and *site*-specific effects as period-dependent functions that vary over space and are spatially correlated. The functional datum $Y_s \in H$ is assumed to be a realization of a stationary Gaussian random field $\{Y_s, s \in \mathcal{D}\}$, where $\mathcal{D}$ is the domain over which the observations are sampled. In order to complete the aforementioned work and build a non-ergodic geostatistical model that is fully functional, this thesis provides a functional extension of the Lanzano, Luzi, Pacor, et al. (2019) scalar regression model

for the median μ, hereafter referred to as ITA18. Instead of separately fit a number N of models for the different ordinates of intensity measure $(Y(T_1), \ldots, Y(T_N))$, we deal with the function representing its profile $\{Y(T), T \in \mathcal{T}\}$.

Before going into details in the explanation of the functional form for the median, we list the dependent and independent parameters that appear in the ITA18 model, introducing their definition and geological meaning.

1.2 Ground motion parameters

GMPEs are formulated to predict the value of some intensity measure (IM) of interest, that quantifies the level of ground shaking at a given site. Of all intensity measures, *peak ground acceleration* (PGA) and *spectral acceleration* (SA) are most often used to describe ground motion. The definition of these quantities requires the preliminary introduction of the *response spectrum*.

The *response spectrum* is a plot that shows the peak acceleration response of a damped linear oscillator, when stressed by a seismic forcer, as function of its natural vibration period T.

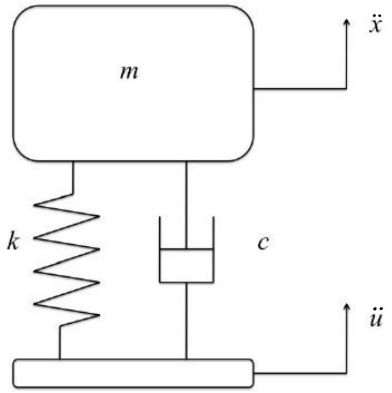


Figure 1.1: Oscillator that models the response of a structure in presence of ground shaking. $\ddot{u}$ and $\ddot{x}$ are the acceleration of the soil and of the mass m, respectively. k is the stiffness of the spring and c is the coefficient of viscous damping of the system.

A damped one dimensional oscillator may be represented by the single-degree-of-freedom system illustrated in Figure 1.1, attached to a base that is moving with displacement u. In the figure, c is the viscous damping of the system and k the stiffness of the spring. The motion x of the mass is described by the equation

$$m\ddot{x} + c(\dot{x} - \dot{u}) + k(x - u) = 0.$$

If divided by m, the equation above becomes

$$\ddot{x} + 2\xi\omega_0\dot{x} + \omega_0^2 x = 2\xi\omega_0\dot{u} + \omega_0^2 u, \tag{1.1}$$

where ω_0 is the undamped natural frequency of the system, ξ the damping ratio and they are defined as

$$\begin{aligned} \omega_0^2 &= \frac{k}{m}, \\ \xi &= \frac{c}{2\sqrt{km}} \end{aligned}.$$

The natural period of the system is given by

$$T = \frac{2\pi}{\omega_0}.$$

Hence, a mathematical definition of the response spectrum, as function of the natural oscillation period T of the system, is in the form

$$RS(T) = \max_t |\ddot{x}_T(t)|,$$

x_T being the solution of (1.1) where we set $\omega_0 = \frac{2\pi}{T}$.

We are now ready to provide a definition of the ground motion parameters.
Spectral acceleration at period T, $SA(T)$, is defined as the value of the response spectrum at T, namely

$$SA(T) = RS(T).$$

In Douglas (2003), the author defines the *peak ground acceleration* as the amplitude of the largest peak acceleration recorded on an accelerogram at a site during a particular earthquake. Mathematically, it is defined as

$$PGA = \max_t |\ddot{u}(t)|.$$

In the argument below, we highlight the link between SA and PGA.
Let x_T be the solution of equation (1.1), describing the absolute displacement x of the mass. We may introduce the relative displacement s_T of the mass as

$$s_T = x_T - u.$$

For systems that are very rigid, *i.e.* with natural oscillation period that tends to 0, the relative displacement, velocity and acceleration of the mass are very small.

Namely, as $T \to 0$,

$$|s_T(t)| \to 0, \qquad |\dot{s}_T(t)| \to 0, \qquad |\ddot{s}_T(t)| \to 0.$$

Hence,

$$\lim_{T\to 0} SA(T) = \lim_{T\to 0} \max_t |\ddot{x}_T(t)| = \lim_{T\to 0} \max_t |\ddot{s}_T(t) + \ddot{u}(t)| = \max_t |\ddot{u}(t)| = PGA.$$

It is now clear that we can think of the PGA as the acceleration endured by a particle placed on the ground, or by a stiff structure that has no relative movement with respect to the ground, while $SA(T)$ is the acceleration experienced by a building, as approximated by a unidimensional oscillator with natural period T.

1.3 Introduction to ITA18

As mentioned above in the chapter, this work focuses on the functional extension of ITA18, the scalar ground motion model for shallow crustal earthquakes in Italy, proposed in Lanzano, Luzi, Pacor, et al. (2019). In their work, the authors revise and solve the shortcomings of the ground motion model introduced in Bindi et al. (2011). In particular, the seismic sequences happened in Italy in 2012 and 2016–2017 allowed for an enlargement of the dataset to a greater magnitude range and to vibration periods up to 10 s. Moreover, as additional information to the data used for ITA10, seismic events are now associated to their fault mechanism and the a parameter characterizing site conditions (V_{S30}) is introduced. Taking advantage of the availability of a richer dataset, ITA18 is calibrated on a discrete set of 37 periods $(T_1, \ldots, T_{37})$, and resorts to a linear ordinary least-squares regression to separately fit N models for the different ordinates of ground motion intensity measure $(Y(T_1), \ldots, Y(T_{37}))$. For the median prediction of each $Y_j = Y(T_j)$, the functional form is

$$\log_{10} Y_j = a + F_M(M_w, \mathrm{SoF}) + F_D(M_w, R) + F_S(V_{S30}) + \epsilon. \tag{1.2}$$

Above, a is the offset, ϵ is the error associated with the median prediction and $F_M(M_w, \mathrm{SoF})$, $F_D(M_w, R)$, $F_S(V_{S30})$ are the *source*, *path* and *site* terms respectively. The latter are specified as follows:

$$F_M(M_w) = \begin{cases} b_1(M_w - M_h) & M_w \leq M_h \\ b_2(M_w - M_h) & M_w > M_h \end{cases},$$
$$F_M(\text{SoF}) = f_j \text{SoF}_j,$$

$$F_D(M_w, R) = [c_1(M_w - M_{\text{ref}}) + c_2] \log_{10}(R) + c_3 R,$$
$$F_S(V_{S30}) = \log\left(\frac{V_0}{800}\right),$$

where M_h is the hinge magnitude, M_{ref} is the reference magnitude and V_0 is defined as $V_0 = V_{S30}$ is $V_{S30 \leq 1500\ m/s}$, $V_0 = 1500m\ s$ otherwise. An in-depth discussion over this formulation and its seismological meaning is taken on in Chapter 4 and Chapter 5. For now, formulation (1.2) stands as benchmark for the extension of ITA18 in an infinite dimensional setting.

In this context, it is necessary to discuss some additional details about the response variable of the model. Most accelerograms report the values of three mutually orthogonal components of the intensity measure, *i.e.* two horizontal and one vertical. On the other hand, the interest of many ground motion models – and ITA18 falls among these – is on predicting values of the intensity measure that are independent of the *in situ* orientation of the recordings. Among all orientation-independent intensity measures introduced in the literature, ITA18 considers the one discussed in Boore (2010) and denoted RotD50, which is computed as the median of the distribution of the intensity measures, obtained from the combination of the two horizontal components across all non-redundant azimuths (Lanzano, Luzi, Pacor, et al., 2019).

Additionally, it is convenient to open a parenthesis on an important feature of the data of our case study, which will be taken up again in Chapter 4, in the sections devoted to dataset exploration (Section 4.2.1).
The recording of soil motion at accelerometric station makes use of high-pass filters. Hence, since the processing is handled manually, high-pass frequencies may differ from site to site, and from component to component of spectral acceleration. This latter peculiarity of our data implies the need to identify, for each observation, the periods at which the recording is not valid. The idea behind the procedure is trivial. Each observation of RotD50 is coupled to the high-pass frequency values of the two filters that recorded the horizontal components of SA. If we refer to these two values as u_h and v_h, then the observation of RotD50 at period T is valid if $1/T$ is greater than both u_h and v_h. If an observation is not valid, then it is considered as a missing value.

It is crucial to point out that this preprocessing generates a number of missing values that varies with the registration periods. Figure 1.2 shows for each registration period T the percentage of curves that are observed at T. The dashed orange line signs the 75%. The number of unobserved curves is low and stable up to a period of 6 s, and then rapidly increases up to 25% at period 10 s.

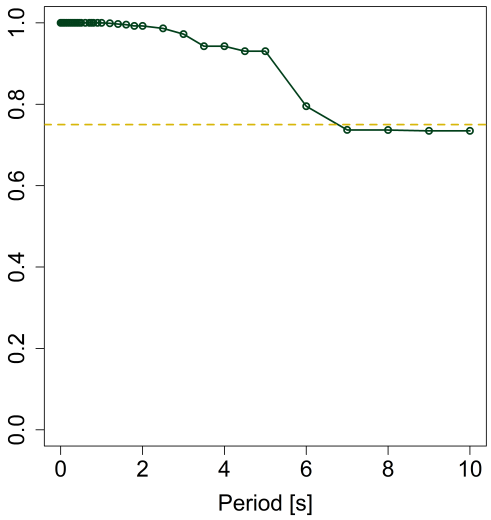


Figure 1.2: Period-specific ratio of the observed records over the total records. At each period T, the green dots represent the percentage of data that are observed at T. The yellow dashed line corresponds to 75 %.

Notice that both discarding the partially observed curves and reducing the analysis over a lower period range would imply a substantial loss of information. Rather, we decide to embed the problem in the context of partially observed functional data, and to reconstruct the records in the unobserved part of their domain. More precisely, the reconstruction is done via extrapolation, that linearly extends the curves from their last valid recording up to 10 s, with a slope equal to the mean slope of all complete curves from that point to 10 s. This peculiarity of our data justifies the embedding of the problem in a very specific functional space, namely in a separable Hilbert space equipped with a weighted norm, that weights differently the observed and the reconstructed parts of an incomplete curve. We further support the choice of such space and provide an introduction to it in Chapter 3.
The extensive discussion about how the ITA18 model is embedded in a period-continuous framework is necessarily preceded by the description of the theoretical and statistical tools that are going to be used for this purpose. The argument is developed in Chapter 2 and Chapter 3. Here, we continue the introduction to the context of ground motion prediction equations, by moving the focus to the seismo-

logical meaning of the parameters that enter *source*-, *path*- and *site*-related terms in formulation (1.2).

1.4 Source, path and site parameters

In the literature, *source*-, *path*- and *site*-related parameters are identified as the three categories of quantities on which the measures of ground motion depend. This split is rather simplistic, both because a separation between the parts is not clearly identifiable, and because it does not reflect how parameters belonging to different groups affect each other. Nonetheless, this representation is commonly adopted, since it reveals intelligible dependencies between variables that are descriptive of the phenomenon and its consequences.

1.4.1 *Source*

The source of the earthquake is characterized by two features: magnitude and mechanism of faulting.
Regarding magnitude, Lanzano, Luzi, Pacor, et al. (2019) along with the majority of publications adopt as measure the moment magnitude M_w, which quantifies the size of an earthquake based on the amount of energy released. Its value is frequently used as the most adequate criterion for ranking seismic events by their strength. The advantage in adopting such parameter is twofold. On the one hand, it has a very clear physical and seismological meaning. On the other hand, it does not saturate for large magnitudes, meaning that it is able to provide a good measure of the energy released over the entire magnitude range (Douglas, 2003).
As for the faulting mechanism, it can be of three types depending on the sense of slip of the rock layers along the fracture plane: *(i)* strike-slip fault (SS), *(ii)* normal fault (NF), *(iii)* thrust (reverse) fault (TF). Even if it is generally found that the style-of-faulting has little impact on the standard deviation of a GMPE, it is still considered as a useful parameter to be included, since the fault mechanism of a future earthquake can easily be identified through techniques of remote sensing and landscape interpretation (Bommer, Douglas, and Strasser, 2003). To the end of producing a more accurate estimate of the seismic hazard, a GMPE that accounts for the style-of-faulting and the identification of the rupture mechanism of a future earthquake efficiently combine to become a practical and convenient tool for seismic hazard analysis.

1.4.2 *Path*

The parameter that characterizes the path term is the distance from the source of the earthquake to the accelerometric station. Such distance may be defined in many alternative ways, but the one used in the ITA18 formulation and that we report here is the *Joyner-Boore* distance (or *surface projection* distance).
Joyner-Boore distance is defined as the distance to the projection on the surface of the rupture plane of the fault.
Distance d is typically not used such as it is, since it leads to improbably high predictions of ground motion values. Alternately, the pure value d is corrected by a term h that represents the depth of the location of the source, and the quantity $R = \sqrt{d^2 + h^2}$ is preferred to d.

1.4.3 *Site*

The necessity of characterizing the local site conditions where ground motion is measured stems from the heavy impact that these conditions may have on the records. Two approaches that integrate GMPEs with site effects are predominantly adopted in the literature.
One resorts to the EC8 site classification, which defines a number of soil categories and associates to each category a multiplicative factor, that enters the equation as a new categorical variable (Eurocode8, 2003). It is worth noticing that, as the boundaries between soil categories are blurry, such method presents an intrinsic subjectivity issue.
A second approach, adopted in ITA18, introduces the *near-surface shear-wave velocity* (VS) as a measure of the characteristics of the soil. The use of this parameter has the double advantage of having a physical sense and of being defined regardless of any soil categorization. Particularly in the GMPE that we are going to discuss in this work, we use the parameter V_{S30}, which is the average shear-wave velocity computed at a reference depth of 30 m. High values of shear-wave velocity are measured at rock and stiff sites and are expected to correspond to lower amplitudes of the spectral acceleration, while low values are associated to soil sites and imply higher amplitudes of SA (Douglas, 2003).

Chapter 2

Introduction to Functional Data Analysis

In recent years, there has been a fast increase in the availability of high-dimensional data coming from the observation of a phenomenon along a continuous domain, *e.g.* time, space or frequency domains. The ever-growing need of analysing data of this kind, characterized by an inherent functional nature, is at the root of the development of the field of Functional Data Analysis (FDA). In this framework, data are observations of functions collected over a discrete grid domain, and are referred to as *functional data*. It is convenient to think of Functional Data Analysis as a generalization of the univariate analysis, where the datum is not a point in the space of real numbers but a function in a functional space. The benefits of adopting a functional approach are manifold. First, the choice of a suitable functional space enables us to capture the smoothness and the regularity underlying the discrete observations recorded along the grid. Second, the embedding of data in an infinite-dimensional setting solves the shortcomings that classical multivariate methods have when the number of variables is larger than the sample size (Ramsay and Silverman, 2005). Third, it allows to extend the focus to differential properties of data, overcoming the limitations of a point-wise analysis.
Generally, Functional Data Analysis takes its steps under the assumption that all functions are observed on the same domain, but in the reality of practitioners it may happen to deal with data that are observed on different sub-domains. For this reason, it has become necessary to develop techniques that are valid and reliable for the manipulation of data of this kind.

Concerning our case study, the adoption of a functional approach allows us to exploit the intrinsic functional nature of the problem, by moving the focus from

a period-specific value of intensity measure, to its profile over a period domain $\mathcal{T}$. Through the formulation of a functional regression model, we provide predictions for the median of the intensity measure ordinates that are continuous with respect to the vibration period. The dependent variable of the model is assumed to be a point of a convenient functional space, with a geometrical structure that captures its key features and natural regularity. The peculiarity of the data analysed in this work is that their processing is manual. The non-automatic handling of the recordings results in high-pass corner frequencies that differ from datum to datum, generating the problem that a non-negligible number of curves are not observed on the whole domain but on different parts of it. This motivates the need to analyse and reconstruct the incomplete curves, extending them into the unobserved part of the domain.

This chapter is organized as follows. First, we consider the classical case of completely observed data and provide a formal definition of random variable with values in an infinite dimensional space. Then, we introduce the space that is a suitable embedding of our analysis; in such framework, we give the formal definitions of mean and covariance operator and recall their main properties. Secondly, we give a brief overview of the analysis of partially observed functional data, developed through the presentation of the state-of-the-art techniques for the reconstruction of the incomplete observations and the estimation of operators.

2.1 Functional data analysis in separable Hilbert spaces

Suppose to observe the values taken by a random variable $\mathcal{X}(\cdot)$ on an interval $\mathcal{T}$. In particular, assume to observe the values of $\mathcal{X}$ at an increasing sequence of sampling instants $\{t_j\}_{j=1}^N$, $t_j \in \mathcal{T}$. Instead of looking at the data as realizations of N distinct random variables $\{\mathcal{X}(t_j)\}_{j=1}^N$, we think of the collection of values as the observation of a continuous random variable $\{\mathcal{X}(t), t \in \mathcal{T}\}$. Below, we provide a formal definition to this intuition.

2.1.1 Aleatory variable in an infinite dimensional space

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let H be an infinite-dimensional vector space with σ-algebra $\mathcal{H}$. The following definitions, present in Ferraty and Vieu (2006), are given in this context.

Definition 2.1.1 (Functional random variable)
A functional random variable $\mathcal{X}$ is a random element on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in the space H, i.e. $\mathcal{X} : \Omega \to H$

Definition 2.1.2 (Functional datum)
A functional datum X is a realization of a functional random variable, i.e. for $\omega \in \Omega$

$$X = \mathcal{X}(\omega) : \mathcal{T} \to \mathbb{R}$$

Definition 2.1.3 (Functional dataset)
A functional dataset is a collection of functional data.

For the sake of analysis, this work takes H to be a separable Hilbert space, the definition of which is quickly recalled in the next section.

2.1.2 Separable Hilbert spaces

Working in an infinite dimensional space implies the need to appropriately extend the statistical tools used in univariate and multivariate analysis. As the inner product exports to a functional space the Euclidean notions of distance, angle and projection, a Hilbert space may be thought of as the generalization to an infinite-dimensional setting of Euclidean spaces. In this peculiarity of Hilbert spaces lies the choice to adopt them as functional spaces in which to embed our analysis. Following the discourse proposed in Horváth and Kokoszka (2012), we recall the definition of Hilbert space and the fundamental notions and properties of operators in Hilbert spaces [1].

Definition 2.1.4 (Hilbert space)
A Hilbert space H is a space equipped with an inner product $\langle \cdot, \cdot \rangle$, that is complete in the distance induced by the inner product.

Recall that the distance between two elements u and v in H is given by the norm of their difference, *i.e.* $dist(u, v) = ||u - v||$.
Every Hilbert space H is connected to its dual by an isometry. Such connection is guaranteed by the Riesz representation theorem.

Theorem 2.1.1 (Riesz representation theorem)
Let $(H, \langle \cdot, \cdot \rangle)$ be a Hilbert space. For every continuous linear functional $\mathcal{L} \in H^*, \exists!\ y \in H$ such that $\mathcal{L}x = \langle x, y \rangle,\ \forall x \in H$.

Let $(H, \langle \cdot, \cdot \rangle)$ be a separable Hilbert space [2] and let $\mathcal{K}$ be the space of compact linear operators on H.

[1] Later in the chapter, these notions will provide important background knowledge for understanding the discussion on incomplete functional data and the state-of-the-art of reconstruction methods.

[2] A Hilbert space is separable if it contains a dense countable subset.

Definition 2.1.5 (Hilbert-Schmidt operator)
An operator $\Psi \in \mathcal{K}$ is Hilbert-Schmidt if $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$, where $\lambda_j \in \mathbb{R}$ are the coefficients appearing in its singular value decomposition [3].

A symmetric [4] and positive-definite [5] Hilbert-Schmidt operator Ψ in a Hilbert space H admits the decomposition

$$\Psi x = \sum_{j=1}^{\infty} \lambda_j \langle x, \varphi_j \rangle \varphi_j, \qquad x \in H,$$

where (λ_j, φ_j) are the eigenvalue-eigenfunction pairings of Ψ [6].

As claimed in Definition 2.1.4, the inner product $\langle \cdot, \cdot \rangle$ induces on H a norm, and hence a metric. Therefore, the choice of the inner product – *i.e.* the choice of the space H – is a crucial step of the analysis and inherently depends on the kind of investigation to be carried out. With an eye toward our analysis, we are interested in exploiting the properties of the space of square integrable functions on a bounded domain [7]. Henceforward, let H be the $L^2(\mathcal{T})$ space, $\mathcal{T} \in \mathbb{R}$ being a bounded interval. The inner product of two elements $x, y \in L^2(\mathcal{T})$ is defined as

$$\langle x, y \rangle = \int_{\mathcal{T}} x(t) y(t) dt,$$

and the induced norm is

$$||x|| = \left(\int_{\mathcal{T}} x^2(t) dt \right)^{1/2}.$$

Hilbert-Schmidt operators on $L^2(\mathcal{T})$ are introduced through the notion of *kernel*. Let $k(\cdot, \cdot) \in L^2(\mathcal{T} \times \mathcal{T})$ and set

$$\Psi_k x(t) = \int_{\mathcal{T}} k(t, s) x(s) ds.$$

Then Ψ_k is a Hilbert-Schmidt operator on $L^2(\mathcal{T})$ and $k(\cdot, \cdot)$ is called kernel of Ψ_k. For a Hilbert-Schmidt operator on L^2, the representation stated by Mercer's theorem holds (Bosq, 1998).

[3] For all $x \in H$, a compact operator Ψ on a separable Hilbert space admits the singular value decomposition $\Psi x = \sum_{j=1}^{\infty} \lambda_j \langle x, \varphi_j \rangle \psi_j$, where $\{\varphi_j\}$ and $\{\psi_j\}$ are two orthonormal bases in H.

[4] An operator $\Psi \in \mathcal{K}$ is symmetric if $\langle \Psi x, y \rangle = \langle x, \Psi y \rangle$, for all $x, y \in H$.

[5] An operator $\Psi \in \mathcal{K}$ is positive-definite if $\langle \Psi x, x \rangle \geq 0$, for all $x \in H$.

[6] (λ_j, φ_j) are the solutions of $\Psi \varphi_j = \lambda_j \varphi_j$.

[7] The reason behind this choice will be made clear later in the chapter, when we introduce the covariance operator and discuss its properties in such space.

Theorem 2.1.2 (Mercer lemma)
Let Ψ_k is a symmetric positive-definite Hilbert-Schmidt operator on $L^2(\mathcal{T})$ and let k be its associated kernel. Then there exists a sequence (φ_j) of continuous functions and a decreasing sequence (λ_j) of positive numbers such that

$$\int_{\mathcal{T}} k(t,s)\varphi_j(s)ds = \lambda_j\varphi_j(t), \qquad t \in \mathcal{T}, \qquad j \in \mathbb{N}, \tag{2.1}$$

and

$$\int_{\mathcal{T}} \varphi_i(s)\varphi_j(s)ds = \delta_{ij}, \qquad i,j \in \mathbb{N}.$$

Moreover,

$$k(t,s) = \sum_{j=1}^{\infty} \lambda_j\varphi_j(t)\varphi_j(s), \qquad t,s \in \mathcal{T}, \tag{2.2}$$

where the series converges uniformly on $\mathcal{T}$, hence

$$\sum_{j=1}^{\infty} \lambda_j = \int_{\mathcal{T}} k(s,s)ds < \infty.$$

Furthermore, the following result holds.

Theorem 2.1.3 (Karhunen-Loève expansion)
Let $\mathcal{X}$ be a zero-mean square-integrable random function with continuous covariance function c. Then

$$\mathcal{X}(t) = \sum_{j=1}^{\infty} \beta_j\varphi_j(t), \qquad t \in \mathcal{T}, \tag{2.3}$$

where (β_j) is a sequence of real zero-mean random variables such that

$$\mathbb{E}[\beta_i\beta_j] = \lambda_i\delta_{ij}, \qquad i,j \in \mathbb{N},$$

and where the sequence (λ_j, φ_j) is defined in Mercer lemma. The series in (2.3) converges uniformly in $L^2(\mathcal{T})$.

2.1.3 Mean and covariance operator in separable Hilbert spaces

Definitions of mean and covariance operator are introduced below, both in the general framework of separable Hilbert spaces and with a focus on the L^2 embedding. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let H be a separable Hilbert space with σ-algebra $\mathcal{H}$. Let $\langle\cdot,\cdot\rangle_H$ be the inner product of H and $||\cdot||_H$ the induced norm. Let $\mathcal{X} : \Omega \to H$ be a functional random variable.

Definition 2.1.6 (Fréchet mean)
We define the Fréchet mean of $\mathcal{X}$ the (unique) element $\mu \in H$ that solves

$$\arg\inf_{x \in H} \mathbb{E}\left[||\mathcal{X} - x||_H^2\right].$$

Additionally, let $\mathcal{X} : \Omega \to H$ be zero-mean.

Definition 2.1.7 (Covariance operator)
The covariance operator of $\mathcal{X}$ is the operator $\mathcal{C} : H \to H$ defined as

$$\mathcal{C}x = \mathbb{E}\left[\langle \mathcal{X}, x\rangle \mathcal{X}\right], \qquad x \in H.$$

As for the case of real-valued random variables, a functional counterpart of the central limit theorem holds. Here, we report the version of the theorem that is stated and proven in Bosq (1998).

Theorem 2.1.4 (Central limit theorem)
Suppose $\{X_i, i \geq 1\}$ is a sequence of iid mean zero random elements in a separable Hilbert space, such that $\mathbb{E}[||X_i||^2] < \infty$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} Z,$$

where Z is a Gaussian random element with covariance operator

$$C(x) = \mathbb{E}[\langle Z, x\rangle Z] = \mathbb{E}[\langle X_1, x\rangle X_1].$$

In Bosq (1998), the author also states and proves a functional version of the law of large numbers.

Theorem 2.1.5 (Law of large numbers)
Suppose $\{X_i, i \geq 1\}$ is a sequence of iid random elements in a separable Hilbert space such that $\mathbb{E}[||X_i||^2] > \infty$. Then $\mu = \mathbb{E}[X_i]$ is uniquely defined by $\langle \mu, x\rangle = \mathbb{E}[\langle X, x\rangle]$, and

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu$$

.

Now let $\mathcal{X}$ be a random variable that takes values in $L^2(\mathcal{T})$, equipped with the Borel σ-algebra. If $\mathcal{X}$ is an integrable random variable[8], then the Fréchet mean of

[8] $\mathcal{X}$ is said to be integrable if $\mathbb{E}[||\mathcal{X}||_H] < \infty$

$\mathcal{X}$ coincides with the point-wise mean

$$\mathbb{E}[\mathcal{X}(t)] = \mu(t) \qquad \text{a.e. in } \mathcal{T}.$$

If $\mathcal{X}$ is zero mean and square integrable[9], then, for all $x \in L^2(\mathcal{T})$, the covariance operator can be defined as

$$\mathcal{C}x(t) = \int_{\mathcal{T}} c(t,s)x(s)ds, \qquad c(t,s) = \mathbb{E}[\mathcal{X}(t)\mathcal{X}(s)]$$

It is possible to show that the covariance operator in L^2 is symmetric and positive-definite, and consequently that it has non-negative eigenvalues. Additionally, since $c(\cdot,\cdot) \in L^2(\mathcal{T} \times \mathcal{T})$, the covariance operator is Hilbert-Schmidt on $L^2(T)$.

2.1.4 Estimation of mean and covariance operators

Suppose to observe a sample $X_1, \ldots, X_n$ of independent realizations of a functional random variable $\mathcal{X}$. For any choice of the Hilbert space H, the empirical counterparts of the mean operator and of the covariance operator are respectively

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$
$$Sx = \frac{1}{n}\sum_{i=1}^{n} \langle X_i, x\rangle X_i, \qquad x \in H.$$

If $\mathcal{X}$ is assumed to be a random element of $L^2(\mathcal{T})$, then the sample mean coincides with the point-wise sample mean

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t),$$

and the covariance kernel coincides with the point-wise sample covariance

$$\hat{c}(t,s) = \frac{1}{n}\sum_{i=1}^{n}(X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(s)), \tag{2.4}$$

so that the sample covariance operator is evaluated as

$$\hat{\mathcal{C}}x = \frac{1}{n}\sum_{i=1}^{n}\langle X_i - \hat{\mu}, x\rangle(X_i - \hat{\mu}), \qquad \forall x \in L^2. \tag{2.5}$$

[9] $\mathcal{X}$ is said to be square integrable if $\mathbb{E}[||\mathcal{X}||_H^2] < \infty$

Under the assumption that the observations are *iid* in $L^2(\mathcal{T})$, $\hat{\mu}$ is proven to be a consistent estimator of μ. In the same way as in the scalar case, it is possible to show that $\hat{c}$ is a biased estimator for c [10], but that the bias is negligible for large sample sizes. As the use of the sample estimate $\hat{\mu}$ introduces in equations (2.4) and (2.5) an additional – although asymptotically negligible – bias in the estimate for $\mathcal{C}$, it is common practice to assume that $\mathcal{X}$ is zero mean. Following this line, the covariance estimators that we will refer to in this work are defined as

$$\hat{c}(t,s) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)X_i(s),$$
$$\hat{\mathcal{C}}x = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, x\rangle X_i$$

2.2 Analysis of partially observed functional data: State of the art

Many techniques and methodologies of classical Functional Data Analysis fail in cases when observations are registered only on subintervals of the domain. For instance, estimators for the mean and the covariance operator cannot be used as they are generally defined, and hence a functional principal component analysis cannot be performed and it is not possible to identify the Karhunen-Loève expansion as in (2.3). We may think to adopt the strategy of deleting the incomplete curves from the dataset, or to restrict the analysis only to the part of the domain in which the curves are all entirely observed. But either option would imply a loss of information that we are not willing to allow. For this reason, several techniques have been developed to manage partially observed data, *de facto* maximizing the information available. These methodologies are effective for the estimation of the mean, of the covariance operator and of the principal component scores, and for the reconstruction of an incomplete recording in the missing part, by borrowing the information from curves that are observed in the missing part.

In this section we provide an overview of the state-of-the-art of such techniques, referring mainly to the two papers Kraus (2015) and Kneip and Liebl (2020), in which alternatives for handling partially observed data are proposed.

[10] The unbiased estimator being

$$\frac{1}{n-1}\sum_{i=1}^{n}\langle X_i - \hat{\mu}, x\rangle(X_i - \hat{\mu})$$

2.2.1 Reconstruction of partially observed functional data

Let $X_1, \ldots, X_n$ be an iid sample of random variables in $L^2(\mathcal{T})$, with $\mathcal{T} \subset \mathbb{R}$. Assume that $\mathbb{E}\left[||X_i||^4_{L^2}\right] < \infty$. Typically, in functional analysis $X_1, \ldots, X_n$ are assumed to be curves observed on the whole interval $\mathcal{T}$. Now, we assume that the curve X_i is observed on $O_i \subset \mathcal{T}$ and that $X_i(t_j),\ t_j \in O_i,\ j = 1, \ldots, N$ are the observed data for the i-th curve.

Let X be a random function with zero mean and covariance operator $\mathcal{C}$, and assume that X has been observed on O and missing on M [11].

- X^O observed part of X, *i.e.* $X^O(t) = X(t)$ for $t \in O$
- X^M missing part of X, *i.e.* $X^M(t) = X(t)$ for $t \in \mathcal{T} \setminus O$
- $\mathcal{C}_{OO}$ covariance operator of X^O
- $\mathcal{C}_{MO}$ cross-covariance operator of X^M and X^O

We mentioned that one of the interests of partially observed data analysis lies in the reconstruction of the incomplete data, through their extension into the missing part of the domain. In the following section, three reconstruction strategies are briefly described as they appear in the literature: *(i)* estimation of the principal component scores, *(ii)* functional completion with a Hilbert-Schmidt operator, *(iii)* functional completion with a reconstruction operator. Each one of these three methods requires the use of the mean and covariance estimators. There are several methods for evaluating these estimators, and Section 2.3 details the proposals made in the two papers we refer to. For the moment, assume that the estimators exist, are consistent and are usable.

2.2.2 Estimation of principal component scores

A convenient way to reconstruct a functional datum can be through the estimation of its scores, as they appear in the Karhunen-Loève representation (2.3). Through the composition of a functional variable as linear combination of its principal components, it is possible to recover the datum, or - better - a fairly good approximation of it.

[11] Here, O denotes a realization of $\mathcal{O}$. Since we assume that the interval of observation is independent of X, then we can consider the observation period as non-random when we make inference on the curve, and derivations are made conditionally on it.

Notice that when a functional datum X_i is unobserved on M_i, a direct computation of

$$\hat{\beta}_{ij} = \langle X_i - \hat{\mu}, \hat{\varphi}_j \rangle = \langle X_i^{O_i} - \hat{\mu}_{O_i}, \hat{\varphi}_j^{O_i} \rangle + \langle X_i^{M_i} - \hat{\mu}_{M_i}, \hat{\varphi}_j^{M_i} \rangle$$

is impossible since the latter addend $\langle X_i^{M_i} - \hat{\mu}_{M_i}, \hat{\varphi}_j^{M_i} \rangle$ is not available. As proposed in Kraus (2015), this section presents a strategy for an estimation from observed data of $\langle X_i^{M_i} - \hat{\mu}_{M_i}, \hat{\varphi}_{jM_i} \rangle$, looking in the class of *continuous linear functionals*.

The goal is to identify the optimal continuous linear functional $\mathcal{L} \in (L^2(\mathcal{T}))^*$ that minimizes

$$\mathbb{E}\left[(\beta_j^M - \mathcal{L}(X^O))^2\right].$$

By Riesz representation theorem, the minimization problem may be equivalently expressed as

$$\min_{a_j \in L^2(O)} \mathbb{E}\left[(\beta_j^M - \langle a_j, X^O \rangle)^2\right]. \tag{2.6}$$

As the objective functional (2.6) is convex, it is minimized by deriving it in the Fréchet sense and equalizing the derivative to 0. Then, the optimal $\hat{a}_j$ is found by solving the following linear inverse problem

$$\mathcal{C}_{OO} a_j = r_j, \tag{2.7}$$

where $r_j = \mathcal{C}_{OM} \varphi_j^M$ and $\mathcal{C}_{OM} = (C_{MO})^*$ is the adjoint operator of $\mathcal{C}_{MO}$.

Observe that system (2.7) is ill posed. Indeed, as $\mathcal{C}_{OO}$ is a compact operator with infinite dimensional range, $\mathcal{C}_{OO}^{-1}$ is not bounded [12] and small perturbations of r_j may imply large perturbations of $\tilde{a}_j = \mathcal{C}_{OO}^{-1} r_j$. Since r_j is not known and has to be estimated, the instability implies that the solution of the inverse problem may be far from the real solution. This is the reason why it becomes crucial to recover stability of the solution in this setting.
To this aim, the method resorts to a ridge regularization and solves

$$\mathcal{C}_{OO}^{(\alpha)} a_j = r_j, \text{where} \qquad \alpha > 0.$$

[12] Concerning the invertibility of compact operators in infinite-dimensional Banach spaces, we are given with the following result.
Let X be an infinite-dimensional Banach space and let T be a compact operator, $T : X \to X$. Assume that T is invertible, with inverse T^{-1}. If T^{-1} were bounded, then $I = TT^{-1}$ would be compact. This implies that the closed unit ball of X is compact, which is impossible. Indeed, one may find an infinite sequence of points in the unit ball whose pairwise distances are bounded, no subsequence of which is Cauchy.

where $\mathcal{C}_{OO}^{(\alpha)} = \mathcal{C}_{OO} + \alpha\mathcal{I}_O$ and $\mathcal{I}_O$ is the identity operator on $L^2(O)$.
Now $(\mathcal{C}_{OO}^{(\alpha)})^{-1}$ is bounded, and the stabilized problem is formulated as

$$\tilde{a}_j^{(\alpha)} = (\mathcal{C}_{OO}^{(\alpha)})^{-1} r_j \tag{2.8}$$

In order to choose for an appropriate regularization parameter $\alpha > 0$, Kraus resorts to a Generalized Cross-validation (GCV) procedure, optimizing with respect to the predictive performance of the corresponding regularized continuous linear functional.

In the practice, the principal score of the i-th curve, observed on O_i, with respect to the j-th eigenfunction is denoted β_{ij} and is estimated via

$$\hat{\beta}_{ij}^{(\alpha)} = \hat{\beta}_{ij}^{O_i} + \hat{\beta}_{ij}^{M_i},$$

with

$$\hat{\beta}_{ij}^{O_i} = \langle X_i^{O_i} - \hat{\mu}^{O_i}, \hat{\varphi}_j^{O_i} \rangle,$$
$$\hat{\beta}_{ij}^{M_i} = \langle \hat{a}_{ij}^{(\alpha)}, X_{iO_i} - \hat{\mu}^{O_i} \rangle.$$

Above, $\hat{\varphi}_j^{O_i}$ is the restriction on O_i of $\hat{\varphi}_j$ and is obtained through the eigenvalue-eigenfunction decomposition of operator $\hat{\mathcal{C}}$, estimated in turn through $\hat{c}$[13]. The term $\hat{\mu}^{O_i}$ is the restriction to O_i of $\hat{\mu}$, and $\hat{a}_{ij}^{(\alpha)}$ is the solution of the empirical counterpart of problem (2.8), namely

$$\hat{a}_{ij}^{(\alpha)} = \left(\hat{\mathcal{C}}_{O_iO_i}^{(\alpha)}\right)^{-1} \hat{r}_{ij}.$$

In the latter, $\hat{\mathcal{C}}_{O_iO_i}^{(\alpha)} = \hat{\mathcal{C}}_{O_iO_i} + \alpha\mathcal{I}_{O_i}$ is the integral operator on $L^2(O_i)$, with kernel equal to the restriction of $\hat{c}$ on $O_i \times O_i$, and $\hat{r}_{ij} = \hat{\mathcal{C}}_{O_iM_i}\hat{\varphi}_{jM_i}$, with $\hat{\mathcal{C}}_{O_iM_i}$ defined by the restriction of $\hat{c}$ to $O_i \times M_i$.

2.2.3 Functional completion with a Hilbert-Schmidt operator

Another viable approach to data reconstruction consists in the functional completion, exploiting the properties of Hilbert-Schmidt operators on L^2. The method that is briefly explained below was first proposed in Kraus (2015) and aims to recover the

[13]Considering the empirical counterpart of (2.1), the eigenvalue-eigenfunction pairings $(\hat{\lambda}_j, \hat{\varphi}_j)$ are found as solutions of

$$\int \hat{c}(t,s)\hat{\varphi}_j(s)ds = \hat{\lambda}_j\hat{\varphi}_j(t), \qquad t \in O \cup M.$$

whole missing part of a trajectory, looking for the solution in the class of *continuous linear operators* from $L^2(O)$ to $L^2(M)$.

Assume for simplicity that the functional variable is zero-mean. Then the minimization problem to be solved is

$$\min_{\Psi:||\Psi||_{\mathcal{B}}<\infty} \mathbb{E}[||X^M - \Psi(X^O)||^2_{L^2}], \tag{2.9}$$

where $||\cdot||_{\mathcal{B}}$ indicates the operator norm[14]. Solving by differentiating the objective function in the Fréchet sense, it is easy to see that the minimization problem can be re-expressed as

$$\begin{aligned} \Psi\mathcal{C}_{OO} &= \mathcal{C}_{MO}, \\ \tilde{\Psi} &= \mathcal{C}_{MO}\mathcal{C}_{OO}^{-1}. \end{aligned} \tag{2.10}$$

and the best linear prediction of the missing part is in the form

$$X^M = \tilde{\Psi}X^O.$$

Observe that, in order to find the solution of (2.10), we need to assume the existence of a bounded solution, *i.e.* $||\mathcal{C}_{MO}\mathcal{C}_{OO}^{-1}||_2 < \infty$ [15].

As in the case of the prediction of the principal scores, $\mathcal{C}_{OO}$ is a compact operator with infinite dimensional range, hence $\mathcal{C}_{OO}^{-1}$ is not bounded and the problem is ill-posed. Again, the solution is stabilized by resorting to a ridge regularization and a GCV procedure for the choice of $\alpha > 0$, so that the best linear continuous regularized operator is found to be in the form

$$\tilde{\Psi}^{(\alpha)} = \mathcal{C}_{MO}(\mathcal{C}_{OO}^{(\alpha)})^{-1}.$$

In the practice of data reconstruction, the mean and covariance operator are substituted by their empirical counterparts $\hat{\mu}$ and $\hat{\mathcal{C}}$ and the estimated optimal operator is set to be

$$\hat{\Psi}_i^{(\alpha)} = \hat{\mathcal{C}}_{M_iO_i}(\hat{\mathcal{C}}_{O_iO_i}^{(\alpha)})^{-1},$$

[14]Let $\mathcal{B}$ be the space of continuous linear operators on a Hilbert space H. Then we define the norm of an operator Ψ in $\mathcal{B}$ as

$$||\Psi||_{\mathcal{B}} = \sup\{||\Psi(x)|| : ||x|| \leq 1\}.$$

[15]The norm $||\cdot||_2$ of an operator Ψ is defined as

$$||\Psi||_2 = \sup\{||\Psi(x)||_{L^2} : ||x||_{L^2} \leq 1\}.$$

where $\hat{\mathcal{C}}_{M_iO_i}$ and $\hat{\mathcal{C}}^{(\alpha)}_{O_iO_i}$ are defined by the restrictions of $\hat{c}$ to $M_i \times O_i$ and $O_i \times O_i$ respectively. Finally, the best linear prediction of the missing part is found to be

$$\hat{X}_i^{M_i(\alpha)} = \hat{\mu}^{M_i} + \hat{\Psi}_i^{(\alpha)}(X_i^{O_i} - \hat{\mu}^{O_i}).$$

Remark 1 (On the non-optimality of Hilbert-Schmidt operators)
It is crucial to point out that, in order to prove consistency for this estimator and additionally to the assumption of boundedness of the solution, we assume it to be a Hilbert-Schmidt operator. In other terms, we strengthen the assumption of continuity $||\tilde{\mathcal{A}}||_{\mathcal{L}} < \infty$ *with the assumption* $||\tilde{\mathcal{A}}||_2 < \infty$.
However, we show below that a Hilbert-Schmidt operator is generally not the optimal reconstruction operator.
Resorting to the Mercer expansion (2.1), a Hilbert-Schmidt operator Ψ_k *on* $L^2(O)$ *may be equivalently expressed as*

$$\Psi_k(X^O)(t) = \int_O k(t,s)X_O(s)ds, \qquad k \in L^2(M \times O). \tag{2.11}$$

Let $\theta \in \delta M$ *be a boundary point of* M. *Then, we would like a constructor to satisfy the continuity constraint at the boundary, i.e.*

$$X^O(\theta) = \Psi(X^O)(\theta), \qquad \forall \theta \in \delta M,$$

which in other terms asks for the first constructed point to equal the last one observed. The kernel satisfying the latter condition is the Dirac-δ*, which is not an element of* $L^2(O)$. *Since the reconstruction cannot be continuous, it is not optimal to our scopes to identify* Ψ *within the class of linear regression operators. Indeed, there is no reason to believe that the optimal reconstructor operator should satisfy (2.11).*
As an alternative, in Kneip and Liebl (2020) the authors look for the optimal linear reconstructor in the more general class of Reconstruction operators.

2.2.4 Functional completion with a reconstruction operator

With the aim of overcoming the discontinuity issue at the boundary and to identify an optimal operator for functional completion, in Kneip and Liebl (2020) the authors introduce a new class of operators, namely the class of *reconstruction operators*, where to seek for the optimal operator that extends a sample of partially observed functional data. In doing so, they introduce the Reproducing Kernel Hilbert Space, in which it is possible to enunciate a representation theorem for reconstruction operators.

First, we introduce the class of reconstruction operators. Let $X^O \in L^2(O)$ be a zero-mean random function.

Definition 2.2.1 (Reconstruction operators) *We call every linear operator $\mathcal{A} : L^2(O) \to L^2(M)$ a reconstruction operator with respect to a random function X^O if $Var(\mathcal{A}(X^O)(t)) < \infty$ for all $t \in M$.*

The Reproducing Kernel Hilbert Space (RKHS), with reproducing kernel the covariance kernel $c^O(t,s) = \mathbb{E}[X^O(t)X^O(s)]$, has inner product

$$\langle f, g\rangle_H \coloneqq \sum_{j=1}^{\infty} \frac{\langle f, \varphi_{OOj}\rangle_2 \langle g, \varphi_{OOj}\rangle_2}{\lambda_{OOj}}, \qquad \forall f, g \in L^2(O),$$

and induced norm

$$||f||_H = \sqrt{\langle f, f\rangle_H}.$$

Above, $(\varphi_{OOj}, \lambda_{OOj})_{j\geq 1}$ are the eigenvalue-eigenfunction pairings of the covariance operator $\mathcal{C}_{OO}$ defined by c^O [16].
The RKH space $H \coloneqq \{f \in L^2(O) : ||f||_H^2 < \infty\}$ is a Hilbert space. The following theorem provides a representation for the linear reconstruction operators on H.

Theorem 2.2.1 (Representation of reconstruction operators) *Let $\mathcal{A} : L^2(O) \to L^2(M)$ be a reconstruction operator with respect to X^O. Then there exists a unique (deterministic) parameter function $\alpha_t \in H$ such that almost surely*

$$\mathcal{A}(X^O)(t) = \langle \alpha_t, X^O\rangle_H, \qquad t \in M.$$

In this framework, the functional completion problem is solved by identifying the optimal linear reconstruction operator $\mathcal{A} : \mathbb{L}^2(O) \to \mathbb{L}^2(M)$ that minimizes

$$\mathbb{E}[(X^M(t) - \mathcal{A}(X^O)(t))^2], \qquad \forall t \in M. \tag{2.12}$$

Using the Representation theorem, the objective function may be re-expressed as

$$\mathbb{E}[(X^M(t) - \langle \alpha_t, X^O\rangle_H)^2], \qquad \forall u \in M.$$

[16] $(\varphi_{OOj}, \lambda_{OOj})_{j\geq 1}$ are solutions of the Mercer's decomposition (2.1) stated for c^O, namely

$$\int_O c^O(t,s)\varphi_j(s)ds = \lambda_j \varphi_j(t), \qquad t \in O.$$

The proposed linear reconstruction operator is

$$\mathcal{A}(X^O)(t) = \langle c_t, X^O \rangle_H, \qquad t \in M, \tag{2.13}$$

where we denote $c_t(s) = c(t,s) = \mathbb{E}[X^M(t)X^O(s)]$, for $t \in M$ and $s \in O$.
The following theorems show that the proposed operator belongs to the class of reconstruction operators, and that it is optimal.

Theorem 2.2.2
(a) $\mathcal{A}(X_O)(t)$ in (2.13) has a continuous and finite variance function, i.e. $V(\mathcal{A}(X^O)(t)) < \infty$ for all $t \in M$.
(b) $\mathbb{E}[\mathcal{A}(X^O)(t)] = 0$ for all $t \in M$, i.e. the operator is unbiased.

Let us introduce the reconstruction error in the form

$$\mathcal{Z} = X^M - \mathcal{A}(X^O), \qquad \mathcal{Z} \in L^2(M).$$

Then the following theorem shows that the reconstruction error is orthogonal to X^O and that the reconstruction operator proposed in (2.13) is optimal.

Theorem 2.2.3 (Optimal linear reconstructor)
(a) For every $s \in O$ and $t \in M$,

$$\mathbb{E}[X^O(s)\mathcal{Z}(t)] = 0.$$

(b) For any linear operator $\mathcal{A} : L^2(O) \to L^2(M)$ that is a reconstruction operator with respect to X^O

$$\mathbb{E}\left[(X^M(t) - \mathcal{A}(X^O)(t))^2\right] \geq V(\mathcal{Z}(t)), \qquad \forall t \in M.$$

Remark 2 (On the continuity of $\mathcal{A}$)
The definition of $\mathcal{A}(X^O)$ over $O \cup M$ comes as a direct extension of the Karhunen-Loève representation of $X^O \in L^2(O)$

$$X^O = \sum_{j=1}^{\infty} \langle X^O, \varphi_{OOj} \rangle_2 \varphi_{OOj}(t), = \sum_{j=1}^{\infty} \beta_j^O \varphi_{OOj}(t), \qquad t \in O.$$

Starting from this representation, every eigenfunction φ_{OOj} defined over O may be continuously extended to a function $\tilde{\varphi}_{OOj}$. Indeed, by Mercer's equation (2.1) for

φ_{OOj} *we have that*

$$\varphi_{OOj}(t) = \frac{\langle \varphi_{OOj}, c_t^O \rangle_2}{\lambda_{OOj}}, \qquad t \in O. \tag{2.14}$$

Through the introduction of $c_t(s)$ *in (2.14),* φ_{OOj} *may be extrapolated to the missing part of the domain as*

$$\tilde{\varphi}_{OOj}(t) \coloneqq \frac{\langle \varphi_{OOj}, c_t \rangle_2}{\lambda_{OOj}}, \qquad t \in M. \tag{2.15}$$

Now it is easy to see how $\tilde{\varphi}_{OOj}$ *appears in the representation of* $\mathcal{A}(X^O)$*. Indeed,*

$$\mathcal{A}(X^O)(t) \coloneqq \langle c_t, X^O \rangle_H = \sum_{j=1}^{\infty} \frac{\langle c_t, \varphi_{OOj} \rangle_2 \langle X^O, \varphi_{OOj} \rangle_2}{\lambda_{OOj}} =$$

$$= \sum_{j=1}^{\infty} \beta_j^O \tilde{\varphi}_{OOj}(t), \forall t \in M. \tag{2.16}$$

Observe now that the definition of $\mathcal{A} : L^2(O) \to L^2(M)$ *can be extended to an operator* $\mathcal{A} : L^2(O) \to L^2(O \cup M)$*. Indeed,*

$$\mathcal{A}(X^O)(t) \coloneqq \langle c_t, X^O \rangle_H = \sum_{j=1}^{\infty} \frac{\langle c_t, \varphi_{OOj} \rangle_2 \langle X^O, \varphi_{OOj} \rangle_2}{\lambda_{OOj}} =$$

$$= \sum_{j=1}^{\infty} \langle X^O, \varphi_{OOj} \rangle_2 \frac{\langle \sum_{k=1}^{\infty} \lambda_{OOk} \varphi_{OOk}(t) \varphi_{OOk}, \varphi_{OOj} \rangle_2}{\lambda_{OOj}} =$$

$$= \sum_{j=1}^{\infty} \langle X^O, \varphi_{OOj} \rangle_2 \frac{\lambda_{OOj} \varphi_{OOj}(t)}{\lambda_{OOj}} = \sum_{j=1}^{\infty} \langle X^O, \varphi_{OOj} \rangle_2 \varphi_{OOj}(t) = X^O(t), \qquad \forall t \in M.$$

Since $c_t(s) = c(t,s) = \mathbb{E}[X(t)X(s)]$ *is a continuous function on* $O \cup M$*, then the reconstructed function* $\mathcal{A}(X^O)$ *is continuous on* $O \cup M$*, and in particular at any boundary point* $\theta \in \delta M$*.*

The KL representation of the reconstruction operator, as expressed in (2.16), and the definition of the reconstruction error allow us to provide a form of the complete reconstructed function X on $O \cup M$ as

$$X(t) = \begin{cases} \sum_{j=1}^{\infty} \beta_j \varphi_{OOj}(t), & t \in O \\ \sum_{j=1}^{\infty} \beta_j \tilde{\varphi}_{OOj}(t) + \mathcal{Z}(t), & t \in M \end{cases}. \tag{2.17}$$

In the practice of estimation of the missing part of an observed function, consider

the general case in which the observed functions $X_i^{O_i}$ are not zero mean. Consider the truncated version of operator (2.16), which is

$$\mathcal{A}_J(X_i^{O_i})(t) = \mu(t) + \sum_{j=1}^{J} \beta_{ij}^{O_i} \tilde{\phi}_{OOj}(t) = \mu(t) + \sum_{j=1}^{J} \beta_{ij}^{O_i} \frac{\langle \varphi_{OOj}, c_t \rangle_2}{\lambda_{OOj}}.$$

Here, the quantities μ, $\beta_{ij}^{O_i}$, $\tilde{\phi}_{OOj}$, c_t and λ_{OOj} are to be estimated, so that the empirical counterpart of the functional reconstructor is found in the form

$$\hat{\mathcal{A}}_J(X_i^{O_i})(t) = \hat{\mu}(t) + \sum_{j=1}^{J} \hat{\beta}_{ij}^{O_i} \frac{\langle \hat{\varphi}_{OOj}, \hat{c}_t \rangle_2}{\hat{\lambda}_{OOj}}, \qquad t \in O \cup M. \tag{2.18}$$

2.3 Estimation of mean and covariance operators for incomplete functional data

When dealing with incomplete functional data, the estimates introduced in Section 2.1.4 are not valid over the entire domain $\mathcal{T}$, because not all curves are defined in every $t_j \in \mathcal{T}$, $j = 1, \ldots, N$. Therefore, effort has been put to find opportune alternatives for mean and covariance estimates, that maximize the information provided by the samples.

An idea of estimators is identified in Kraus (2015), in which the author proposes that $\mu(t)$ can be estimated as the sample mean of all observed curves in t, namely

$$\hat{\mu}(t) = \frac{J(t)}{\sum_{i=1}^{n} O_i(t)} \sum_{i=1}^{n} O_i(t) X_i(t), \tag{2.19}$$

where

$$O_i(t) = \begin{cases} 1 & \text{if } X_i \text{ is observed in } t \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad J(t) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} O_i(t) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

and the covariance operator $\mathcal{C}$ can be estimated through an estimator for $c(t, s)$, which is the sample covariance of all complete pairs of functional values at t and s, namely

$$\hat{c}(t, s) = \frac{I(t, s)}{\sum_{i=1}^{n} U_i(t, s)} \sum_{i=1}^{n} U_i(t, s) \{X_i(t) - \hat{\mu}_{ts}(t)\} \{X_i(s) - \hat{\mu}_{ts}(s)\}, \tag{2.20}$$

where

$$U_i(s,t) = O_i(s)O_i(t),$$

$$I(s,t) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} U_i(s,t) > 0 \\ 0 & \text{otherwise} \end{cases},$$

$$\hat{\mu}_{ts}(t) = \frac{I(t,s)}{\sum_{i=1}^{n} U_i(t,s)} \sum_{i=1}^{n} U_i(t,s) X_i(t).$$

In the framework introduced in Section 2.2.1, estimators (2.19) and (2.20) are proven to be consistent estimators for $\mu(t)$ and $c(t,s)$ respectively, under the additional assumptions that: *(i)* $O_1, \ldots, O_n$ are independent and identically distributed, *(ii)* for every pair (t,s), the probability that a curve is observed on both t and s is greater than 0.

Another option for mean and covariance estimates consists in techniques that are typically used for the analysis of sparse functional data (Yao, Müller, and Wang, 2005). In Kneip and Liebl (2020), the authors take advantage of the class of Local Linear Kernel (LLK) estimators to provide smoother estimates than the point-wise evaluations proposed in Kraus (2015). In the paper the estimators are introduced as follows.

Let $K(\cdot)$ be a second-order kernel with compact support. Then $\hat{\mu}(t) = \hat{\gamma}_0$, where $(\hat{\gamma}_0, \hat{\gamma}_1)$ solves

$$\underset{\gamma_0,\gamma_1}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{N} [X_i(t_j) - \gamma_0 - \gamma_1(t_j - t)]^2 K(t_j - t),$$

and $\hat{c}(t,s) = \hat{\gamma}_0$, where $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2)$ solves

$$\underset{\gamma_0,\gamma_1,\gamma_2}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{1 \le j,l \le N} [\hat{\mathcal{C}}_{ijl} - \gamma_0 - \gamma_1(t_j - t) - \gamma_2 t_j - s)]^2 K(t_j - t) K(t_l - s),$$

and $\hat{\mathcal{C}}_{ijl} = (X_i(t_j) - \hat{\mu}(t_j))(X_i(t_l) - \hat{\mu}(t_l))$ is the sample covariance estimate.

Again in the framework of Section 2.2.1, the consistency of these estimators is proven in Kneip and Liebl (2020) under the additional assumptions of: *(i)* sufficient smoothness of μ and c, *(ii)* sufficient numerosity of the time instants of recording with respect to the number of observations.

Chapter 3

Weighted analysis for functional data

In the framework of partially observed functional data, the handling of an incomplete curve is typically done via reconstruction, exploiting for instance one of the methodologies presented in Chapter 2. In this chapter, we present a methodology to deal with incomplete functional data that have been reconstructed. The idea behind this method is simple: a reconstruction of the data is performed, and then the reconstructed observation is associated to a weight, which takes value 1 where the data is not reconstructed – i.e., where we have maximum confidence in the value assumed by the curve – and decreases towards 0 the less confidence we have in the reconstructed value. Doing so, the part of the domain which corresponds to the unobserved data counts less in the estimation process. We want to give the theoretical and methodological basis to handle the functional analysis in this context, thus including the weights in the classical methodologies of *smoothing*, *function-on-scalar regression* and *estimation of the variability* of the regression coefficients estimates.

3.1 Weighted norm in L^2

Let $\mathcal{T}$ be an open subset of $\mathbb{R}$ and $w : \mathbb{T} \to [0, \infty)$ be a bounded non-negative function, which we refer to as *weight*.
Now let $f, g \in L^2(\mathcal{T})$ and let w, v be weights associated to f and g respectively. We define the *weighted inner product* in L^2 as

$$\langle f, g \rangle_W = \int_{\mathcal{T}} \sqrt{w(s)} f(s) \sqrt{v(s)} g(s) ds.$$

Observing that $\langle f, g\rangle_W = \langle\sqrt{w}f, \sqrt{v}g\rangle_{L^2(\mathcal{T})}$, it is immediate to verify that the weighted inner product is bilinear, symmetric and positive definite.
The weighted inner product induces a norm on $L^2(\mathcal{T})$. We define the *weighted* L^2 *norm* of f with respect to w as

$$||f||_W = \sqrt{\langle f, f\rangle_W} = \left(\int_{\mathcal{T}} (\sqrt{w(s)}f(s))^2 ds\right)^{1/2}.$$

It is trivial to see that if the L^2 norm of f is finite, then also the weighted L^2 norm of f is finite. Indeed,

$$||f||_W^2 = \int_{\mathcal{T}} (\sqrt{w(s)}f(s))^2 ds \leq \sup_{t\in\mathcal{T}} w(t) \int_{\mathcal{T}} (f(s))^2 ds \leq K||f||_{L^2(\mathcal{T})} < \infty.$$

In the following sections we will resort to the weighted L^2 norm to identify the criteria for smoothing functional data and fitting the regression.

3.2 Smoothing

The smoothing methodology allows one to represent the functional data as immersed in a space generated by a (finite) set of basis functions, and as univocally identified by the coefficients of their projection on the basis. As a result, it becomes possible to link functional analysis to multivariate analysis, and to apply in this context the straightforward extension of methodologies already used in multivariate analysis.
In this section, we illustrate the penalized weighted smoothing technique, that we apply adopting a *roughness penalization method*. Through the introduction of a penalty term, a penalized criterion limits the inherent tendency of the least squares method to perfectly interpolate the points $y_i(t_1), \ldots, y_i(t_N)$, giving rise to a trade-off between the accuracy of the fitting and the regularity of the smoothed function.

We mentioned that we want to adapt the common smoothing technique to the weighted framework where we place ourselves. Actually, a smoothing that weights differently at the different sampling instants is already present in the literature, and is well illustrated in Ramsay and Silverman (2005). However, specifically in our case, the weighting – and consequently the smoothing – is curve-specific and depends on the weighting function coupled to the observation.
Although this peculiarity does not affect the common argument about weighted smoothing, the issue should be kept in mind when one wants to represent the overall

smoothing of observations through a mapping. In this section, first we present the penalized weighted smoothing for a single observation, by closely following Ramsay and Silverman (2005), Chapter 5, and then we deal with the aforementioned problem of building the projection map.

3.2.1 The penalized weighted least squares criterion

Assume $Y_1, \dots, Y_n$ to be independent and identically distributed random variables with values in $L^2(\mathcal{T})$, where $\mathcal{T}$ is an open subset of $\mathbb{R}$. Assume to observe $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$, realizations of $Y_1, \dots, Y_n$. In particular, assume each observation $\boldsymbol{y}_i$ to be a vector of discrete recordings of y_i, evaluated at a sequence of breaks $t_1, \dots, t_T$, where $t_j \in \mathcal{T}$ for all $j = 1, \dots, T$.

For a generic observation $\boldsymbol{y}_i$, the smoothing technique fits the discrete values $y_i(t_1), \dots, y_i(t_T)$ resorting to the model

$$y_i(t_j) = x_i(t_j) + \epsilon_i(t_j), \qquad \forall j = 1, \dots, T,$$

where the smoothed curve x_i is defined on the space generated by a basis $\phi_1, \dots, \phi_L$, so that it may be written in the form

$$x_i(t) = \sum_{l=1}^{L} c_{il}\phi_l(t) = \boldsymbol{c}_i^\intercal \boldsymbol{\phi}(t). \tag{3.1}$$

The basis is common to all smoothed curves $(x_1, \dots, x_n)$. Let $\boldsymbol{c}_i$ be the vector of the coefficients of the linear combination specific of x_i and let $\boldsymbol{\phi}(t) = (\boldsymbol{\phi}_1(t), \dots, \boldsymbol{\phi}_L(t))^\intercal$ be the vector of evaluations of the basis functions in t.
Define the *smoothing error* ϵ_i as a function, observed at the sequence of sampling instants $(t_1, \dots, t_N)$ and taking values

$$\epsilon_i(t_j) = y_i(t_j) - x_i(t_j) = y_i(t_j) - \boldsymbol{c}_i^\intercal \boldsymbol{\phi}(t_j), \qquad \forall j = 1, \dots, T.$$

For each ϵ_i, let w_i be the associated weight as specified in Section 3.1.
The *penalized least squares criterion* works by adding a roughness penalty to the classical error sum of squares. We adopt a natural quantification of roughness, that considers the abrupt changes in the curve by evaluating the square of the L^2 norm of its second derivative, *i.e.*

$$||D^2 x||_{L^2(\mathcal{T})} = \int_{\mathcal{T}} \left[(D^2 x)(s)\right]^2 ds.$$

Summing up, the *penalized weighted least squares criterion* consists in solving

$$\hat{\boldsymbol{c}}_i = \underset{\boldsymbol{c}\in\mathbb{R}^L}{\operatorname{argmin}} \sum_{j=1}^{T} \left(\sqrt{w_i(t_j)}\epsilon_i(t_j) \right)^2 + \zeta ||D^2 x||_{L^2(\mathcal{T})}, \tag{3.2}$$

where ζ is a smoothing parameter that manages the trade-off between the good fitting of data and the roughness of the smooth function: the smaller ζ, the more the estimate is close to the least squares estimate and tends to interpolate the observed points; the greater ζ, the more flat the smooth function tends to be. Such parameter may be calibrated through a generalized cross-validation procedure, that selects ζ corresponding to the minimum of the mean squared error discounted by a measure of increased regularity.

Taking advantage of the result of de Boor (Boor, 2001), we look for the optimal smoothing curve in the class of cubic splines with knots at $t_1, \ldots, t_T$, solving problem (3.2) with respect to the coefficients of (3.1).
The closed form expression of the optimal $\hat{\boldsymbol{c}}$ is derived below, resorting to a matricial representation of (3.2).
First, observe that the penalization term may be re-expressed in matricial form as

$$||D^2 x||_{L^2(\mathcal{T})} = \int_{\mathcal{T}} \left[(D^2 x)(s) \right]^2 ds = \int_{\mathcal{T}} \left[(D^2 \boldsymbol{c}^\intercal \boldsymbol{\phi})(s) \right]^2 ds = \boldsymbol{c}^\intercal R \boldsymbol{c},$$

where $[R]_{lk} = \langle D^2\phi_l, D^2\phi_k \rangle_{L^2(\mathcal{T})}$.
Secondly, observe that the error sum of square may be written as

$$\sum_{j=1}^{T} \left(\sqrt{w_i(t_j)}\epsilon_i(t_j) \right)^2 = \boldsymbol{\epsilon}_i^\intercal W_i \boldsymbol{\epsilon}_i = (\boldsymbol{y}_i - \Phi\boldsymbol{c})^\intercal W_i (\boldsymbol{y}_i - \Phi\boldsymbol{c}).$$

Above, $W_i = \operatorname{diag}(w_i(t_1), \ldots, w_i(t_N))$ is a T-order diagonal matrix which plays the role of weighing differently the different temporal instants, according to the weight which has been assigned to the i-th observation. The term $\Phi \in \mathbb{R}^{(T\times L)}$ is a matrix, whose columns contain the values that the basis functions take at the sampling instants.

Finally, we can equivalently formulate (3.2) as the problem of finding the minimum

of the quadratic form

$$(\boldsymbol{y}_i - \Phi\boldsymbol{c})^\intercal W_i (\boldsymbol{y}_i - \Phi\boldsymbol{c}) + \zeta \boldsymbol{c}^\intercal R\boldsymbol{c}, \tag{3.3}$$

By taking the derivative of (3.3) with respect to $\boldsymbol{c}$ and by setting it to zero, we get

$$-2\Phi^\intercal W_i \boldsymbol{y}_i + 2\Phi^\intercal W_i \Phi\boldsymbol{c} + 2\zeta R\boldsymbol{c} = 0,$$

and it is immediate to see that the optimal coefficient vector for observation i is given by

$$\hat{\boldsymbol{c}}_i = (\Phi^\intercal W_i \Phi + \zeta R)^{-1} \Phi W_i \boldsymbol{y}_i. \tag{3.4}$$

3.2.2 Construction of the smoothing map

Expression (3.4) leads to the introduction of a map S_Φ^i, that for each observation establishes a connection between the vector of recordings $\boldsymbol{y_i}$ and the vector $\hat{\boldsymbol{c}}_i$ of coefficients of the basis expansion. For all $i = 1, \ldots, n$, such mapping is represented by the $L \times T$ matrix

$$S_\Phi^i \coloneqq (\Phi^\intercal W_i \Phi + \zeta R)^{-1} \Phi W_i, \tag{3.5}$$

so that

$$\hat{\boldsymbol{c}}_i = S_\Phi^i \boldsymbol{y}_i.$$

Now let $Y \in \mathbb{R}^{(n \times T)}$ be the matrix containing the values that n observations take in T sampling points, and let $C \in \mathbb{R}^{(n \times L)}$ be the matrix that collects all n optimal coefficient vectors, found following the procedure developed in the previous section. We are interested in building a mapping $\boldsymbol{S}_\Phi$ between Y and C.
Prior to this, it is important to point out that, in the classical framework of weighted smoothing, the weighting is introduced in order to account for the presence of correlations between different time instants. Therefore, it is reasonable that the smooth counterpart of each observation is computed resorting to the same weighting matrix W. This implies that, by definition (3.5), S_Φ is the same for every observation, and that the connection between Y and C is easily found as

$$C = Y S_\Phi^\intercal. \tag{3.6}$$

Notice that, recalling the properties of the Kronecker product[1], the relation above

[1] Let A be a matrix of dimension $k \times l$ and B be a matrix of dimension $m \times n$. Then the Kronecker product $A \otimes B$ generates a matrix C of dimension $km \times ln$ that is given by sub-matrices $a_{ij}B$.

can equivalently be expressed as

$$\text{vec}(C) = \text{vec}(IYS_\Phi^\intercal) = (S_\Phi \otimes I)\text{vec}(Y), \text{[2]} \tag{3.7}$$

where I is the n-dimensional identity matrix.

Moving the focus on our setting again, we observe that the smoothing is performed through an observation-specific mapping S_Φ^i, and hence that a counterpart of representation (3.6) does not exist. On the contrary, it is possible to manually construct a map that extends relation (3.7) to the more general case of observation-specific weighting. Indeed, suppose to have S^i , $i = 1, \dots, n$, different smoothing maps [3]. One may check that the counterpart of $(S_\Phi \otimes I)$ is the $(Ln \times Tn)$-dimensional sparse matrix

$$\boldsymbol{S}_\Phi \coloneqq \begin{pmatrix}
S_{11}^1 & 0 & \dots & 0 & S_{12}^1 & 0 & \dots & 0 & & S_{1T}^1 & 0 & \dots & 0 \\
0 & S_{11}^2 & \dots & 0 & 0 & S_{12}^2 & \dots & 0 & & 0 & S_{1T}^2 & \dots & 0 \\
\vdots & & \ddots & & \vdots & & \ddots & & & \vdots & & \ddots & \\
0 & \dots & 0 & S_{11}^n & 0 & \dots & 0 & S_{12}^n & & 0 & \dots & 0 & S_{1T}^n \\
 & & & & & & & & & & & & \\
 & & & & & & & & & & & & \\
S_{L1}^1 & 0 & \dots & 0 & S_{L2}^1 & 0 & \dots & 0 & & S_{LT}^1 & 0 & \dots & 0 \\
0 & S_{L1}^2 & \dots & 0 & 0 & S_{L2}^2 & \dots & 0 & & 0 & S_{LT}^2 & \dots & 0 \\
\vdots & & \ddots & & \vdots & & \ddots & & & \vdots & & \ddots & \\
0 & \dots & 0 & S_{L1}^n & 0 & \dots & 0 & S_{L2}^n & & 0 & \dots & 0 & S_{LT}^n
\end{pmatrix},$$

so that

$$\text{vec}(C) = \boldsymbol{S}_\Phi \text{vec}(Y). \tag{3.8}$$

This notation will turn out to be particularly useful in Section 3.3.2, where the link between the variability of Y and the variability of C is deployed.

3.3 Functional linear regression

Linear regression is a practical tool that is employed and well known both in univariate and in multivariate statistical analysis, but extensions of the linear regression model are possible also in the context of Functional Data Analysis (Ramsay and Silverman, 2005, Horváth and Kokoszka, 2012).

[2] Given a matrix A of dimension $k \times m$, vec(A) indicates the km-dimensional vector obtained by writing A as a vector column-wise.

[3] Here, we temporarily drop the subscript Φ for clarity of notation.

The main problem with a functional extension of the linear regression is that each regression coefficient β is an infinite dimensional object that has to be estimated from a finite sample (Horváth and Kokoszka, 2012). If no restrictions are imposed on β, then the regression results in an estimate that is inherently unstable, noisy and uninformative, although providing a perfect fitting of the observations. One way to prevent such side effect consists in imposing a roughness penalty on the functional estimate of the coefficient, that counterbalances the pursuit of a good fitting with the estimation of a coefficient that is regular, stable and able to provide useful insights on the phenomenon under analysis.

3.3.1 Function-on-scalar linear regression

As our interest lies in a functional extension of a linear regression with scalar parameters, the most convenient choice falls on the adoption of a *function-on-scalar* linear regression model with independent multivariate covariates $(x_1, \dots, x_q)$ and functional coefficients $(\beta_1(\cdot), \dots, \beta_q(\cdot))$, where the functional errors are assumed to be uncorrelated.
In the framework defined in Section 3.2, the linear regression model is formulated as

$$\boldsymbol{y}(t) = X\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t), \qquad t \in \mathcal{T}. \tag{3.9}$$

Above, $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_q(t))$ denotes the vector of functional coefficients evaluated in t, $X \in \mathbb{R}^{n \times q}$ is the design matrix and $\boldsymbol{y}(t)$ is a n-dimensional vector containing the response functions evaluated in t. Lastly, the error term is a n-dimensional vector of functions ϵ_i, that are assumed to be independent realizations of a stochastic process with zero mean and covariance function $c(t, s)$.

In order to define a criterion to fit the model, we aim to extend to a functional case the penalized least-square principle, resorting to the weighted L^2 norm introduced in Section 3.1. Similarly to the smoothing procedure, we associate the errors $\epsilon_1, \dots, \epsilon_n$ to the weights $w_1, \dots, w_n$ and define the *weighted functional least-square (WFLS)* criterion as the minimization of

$$
\begin{aligned}
WFLS &= \sum_{i=1}^{n} ||\epsilon_i||^2_{L^2(\mathcal{T}),W} \\
&= \sum_{i=1}^{n} \int_{\mathcal{T}} (\sqrt{w_i(s)}\epsilon_i(s))^2 ds \\
&= \int_{\mathcal{T}} \sum_{i=1}^{n} \left[\sqrt{w_i(s)} \left(y_i(s) - \boldsymbol{x}_i^{\intercal} \boldsymbol{\beta}(s) \right) \right]^2 ds \\
&= \int_{\mathcal{T}} \left[\boldsymbol{y}(s) - X\boldsymbol{\beta}(s) \right]^{\intercal} W(s) \left[\boldsymbol{y}(s) - X\boldsymbol{\beta}(s) \right] ds, \qquad (3.10)
\end{aligned}
$$

where we set $W(t) = \text{diag}\left(w_1(t), \ldots, w_n(t)\right)$, for all $t \in \mathcal{T}$, to be the diagonal matrix of the weights evaluated in t.
In order to regularize and stabilize the estimate of the regression coefficients, a penalization term that quantifies the roughness of the coefficient functions is added to the *WFLS*. Suppose to the define a correction to the bumpiness through a linear differential operator L. Then, we call the fitting criterion a *penalized weighted least-square (PWFLS)* and the quantity to be minimized becomes

$$
PWFLS = \int_{\mathcal{T}} \left[\boldsymbol{y}(s) - X\boldsymbol{\beta}(s) \right]^{\intercal} W(s) \left[\boldsymbol{y}(s) - X\boldsymbol{\beta}(s) \right] ds + \lambda \int_{\mathcal{T}} [L\boldsymbol{\beta}(s)]^{\intercal} [L\boldsymbol{\beta}(s)] ds, \quad (3.11)
$$

where λ is a smoothing parameter that can be tuned through a generalized cross-validation procedure.

The general idea for the resolution of the minimization problem consists in moving from an infinite dimensional setting to a multivariate framework, by projecting the observations in the space spanned by the basis functions, as discussed in Section 3.2. The functional coefficients are estimated as elements of a space generated by suitable basis functions. In doing so, the part of the curves that is not captured by the basis functions is assumed to be negligible and is included in the regression error. Having this in mind, in this section we will refer to the response variable and to the functional coefficients as

$$
\begin{aligned}
y_i(t) &= \sum_{l=1}^{L_y} c_{il} \phi_l(t), \qquad i = 1, \ldots, n, \\
\beta_j(t) &= \sum_{l=1}^{L_\beta} b_{jl} \theta_l(t), \qquad j = 1, \ldots, q. \qquad (3.12)
\end{aligned}
$$

The two bases for the observations and the coefficients are typically chosen to be

the same, but we maintain the argument in a more general framework and consider them as distinct.

A new phrasing of model (3.9) follows all the considerations made above, and takes the form

$$\mathrm{C}\boldsymbol{\phi}(t) = X\mathrm{B}\boldsymbol{\theta}(t) + \mathcal{E}(t), \tag{3.13}$$

where C and B are matrices of dimensions $n \times L_y$ and $q \times L_\beta$ respectively, that contain the projection coefficients of $\{y_1, \ldots, y_n\}$ and $\{\beta_1, \ldots, \beta_q\}$ appearing in (3.12). Putting formulation (3.13) in (3.11), one gets

$$\int_{\mathcal{T}} \left[\mathrm{C}\boldsymbol{\phi}(t) - X\mathrm{B}\boldsymbol{\theta}(t)\right]^\intercal W(t) \left[\mathrm{C}\boldsymbol{\phi}(t) - X\mathrm{B}\boldsymbol{\theta}(t)\right] dt + \lambda \int_{\mathcal{T}} \left[L\mathrm{B}\boldsymbol{\theta}(t)\right]^T \left[L\mathrm{B}\boldsymbol{\theta}(t)\right] dt.$$

This formulation of the objective function is the starting point of the calculation, extensively reported in Appendix A.1, that leads to the following equation for matrix B:

$$\left[J + \lambda R \otimes I_q\right] \mathrm{vec}(\mathrm{B}) = \mathrm{vec}\left(\int X^\intercal W(t) \mathrm{C}\boldsymbol{\phi}(t) \boldsymbol{\theta}(t)^\intercal dt\right),$$

where

$$R \coloneqq \int_{\mathcal{T}} \left[L\boldsymbol{\theta}(t)\right] \left[L\boldsymbol{\theta}(t)\right]^T dt,$$
$$J \coloneqq \int \left(\boldsymbol{\theta}(t)\boldsymbol{\theta}(t)^\intercal \otimes X^\intercal W(t) X\right) dt.$$

To sum up, the estimate of the B matrix is given through the vec() operator by

$$\mathrm{vec}(\hat{\mathrm{B}}) = \left[J + \lambda R \otimes I_q\right]^{-1} \mathrm{vec}\left(\int X^\intercal W(t) \mathrm{C}\boldsymbol{\phi}(t) \boldsymbol{\theta}(t)^\intercal dt\right). \tag{3.14}$$

3.3.2 Estimation of the variability of regression coefficients

Whenever an analysis leads to the computation of a point estimation of a parameter, it is statistically relevant to couple the estimate with a quantification of the related uncertainty. This section is devoted to the deployment of an argument that shows how to evaluate the variability associated to the estimate obtained in (3.14). The way to achieve this is through the construction of a mapping that acts as linkage between the raw observations and the estimate $\hat{B}$.

First, let us see how the general idea applies to the simpler case of a single curve y, of which we observe a sample $\boldsymbol{y} = (y(t_1), \ldots, y(t_T))$. Assume that the vector $\boldsymbol{y}$ goes

through a transformation, *e.g.* a smoothing that binds the rough observations of y to a smooth function x. We are interested in investigating the variability of some feature of x, identified by a functional $\mathcal{L}(x)$ as

$$\mathcal{L}(x) = \int \xi(s)x(s)ds.$$

Suppose that our concern lies in a functional

$$\mathcal{L}_t(x) = x(t),$$

represented by a kernel ξ that takes positive values only in a small interval centred in t. Let S_Φ and L be the two linear operators that map $\boldsymbol{y}$ into x and x into $\mathcal{L}_t(x) = x(t)$, respectively. Then let $M = L \circ S_\Phi$ be the composite mapping, so that $\mathcal{L}_t(x) = x(t) = LS_\Phi \boldsymbol{y}$. Then, using the expression for the variance of transformed random vectors, we have that

$$\text{Var}[\hat{x}(t)] = \text{Var}(LS_\Phi \boldsymbol{y}) = LS_\Phi \Sigma_e S_\Phi^\intercal L^\intercal,$$

where Σ_e is the covariance matrix of the sample $\mathbf{y}$.

Let us now extend and adapt this argument to the case in which a sample of functional observations goes through a smoothing and a linear regression, as described in the previous sections. In order to identify the linkage between the observations and the estimates of the regression coefficients, it is useful to view the overall mapping as the composition of: *(i)* the *smoothing map* that associates the observations to the smooth functions, *(ii)* the *regression map* that connects the smooth functions to the matrix of coefficients B, *(iii)* the *basis expansion map* that couples the estimated coefficients with the basis functions for the β's. In this work, all mappings are handled in the same matricial form that has been adopted above to discuss smoothing and regression.

The construction of the *smoothing map* is described in Section 3.2.2, and we refer to it as S_Φ, namely the $(nL \times nT)$-dimensional matrix such that

$$\text{vec}(C) = S_\Phi \text{vec}(Y).$$

The *regression map* connects vec(C) into vec(B) and is found by reformulating what is reported in (3.14). Indeed, exploiting the properties of the Kronecker product,

the relation may be written in the form

$$\text{vec(B)} = [J + \lambda R \otimes I_q]^{-1} \left(\int \boldsymbol{\theta}\boldsymbol{\phi}^\intercal \otimes X^\intercal W \right) \text{vec(C)},$$

and we may identify the mapping as the $(Lq \times Ln)$-dimensional matrix

$$S_\beta \coloneqq [J + \lambda R \otimes I_q]^{-1} \left(\int \boldsymbol{\theta}\boldsymbol{\phi}^\intercal \otimes X^\intercal W \right).$$

The *basis expansion map* carries out the linear combination of the basis function that uniquely define $\hat{\boldsymbol{\beta}}$ from the estimated coefficients.
Recalling the matricial counterpart of (3.12), it is immediately found that

$$\text{vec}(\hat{\boldsymbol{\beta}}) = \text{vec}(\text{B}\Theta^\intercal) = (\Theta \otimes \text{I}_q)\text{vec(B)},$$

where Θ is a $T \times L$ matrix of values of the functions $(\theta_1, \ldots, \theta_L)$ evaluated in the T sampling points. Then the basis expansion mapping is defined as the $(Tq \times Lq)$ matrix

$$S_\Theta \coloneqq \Theta \otimes \text{I}_q.$$

Finally, the complete mapping of Y into $\hat{\boldsymbol{\beta}}$ is given by the composition of all three mappings identified above, and may be expressed in matricial form as

$$\text{Map} \coloneqq S_\Theta S_\beta S_\Phi,$$

and

$$\text{vec}(\hat{\boldsymbol{\beta}}) = S_\Theta S_\beta S_\Phi \text{vec(Y)}. \tag{3.15}$$

Now observe that the variance of the observations is given by the $(Ln \times Ln)$-dimensional matrix

$$\text{Var}[\text{vec(Y)}] = \Sigma_e \otimes \text{I}_n,$$

where Σ_e is the covariance matrix of the residuals $\hat{\epsilon}_i$ from the regression model.

Summing all up, we may express the variability of the regression coefficients in terms of the variability of the observations

$$\text{Var}[\text{vec}(\hat{\boldsymbol{\beta}})] = \text{Var}[S_\Theta S_\beta S_\Phi \text{vec}(Y)] = S_\Theta S_\beta S_\Phi (\Sigma_e \otimes \text{I}_n) S_\Phi^\intercal S_\beta^\intercal S_\Theta^\intercal, \tag{3.16}$$

and we obtain the symmetric covariance matrix, whose diagonal values are the variances of $(\beta_1, \ldots, \beta_q)$, for all $t_j,\ j = 1, \ldots, T$, *i.e.*

$$\left(\begin{array}{c|c|c|c} \mathrm{V}(\hat{\boldsymbol{\beta}}(t_2)) & & & \\ \hline \mathrm{C}(\hat{\boldsymbol{\beta}}(t_1), \hat{\boldsymbol{\beta}}(t_2)) & \mathrm{V}(\hat{\boldsymbol{\beta}}(t_2)) & & \\ \hline \vdots & \vdots & \ddots & \\ \hline \mathrm{C}(\hat{\boldsymbol{\beta}}(t_1), \hat{\boldsymbol{\beta}}(t_T)) & \mathrm{C}(\hat{\boldsymbol{\beta}}(t_2), \hat{\boldsymbol{\beta}}(t_T)) & \ldots & \mathrm{V}(\hat{\boldsymbol{\beta}}(t_T)) \end{array}\right),$$

where the generic $\mathrm{C}(\hat{\boldsymbol{\beta}}(t_i), \hat{\boldsymbol{\beta}}(t_j))$ takes the form

$$\mathrm{C}(\hat{\boldsymbol{\beta}}(t_i), \hat{\boldsymbol{\beta}}(t_j)) = \begin{pmatrix} \mathrm{C}(\hat{\beta}_1(t_i), \hat{\beta}_1(t_j)) & & & \\ \mathrm{C}(\hat{\beta}_1(t_i), \hat{\beta}_2(t_j)) & \mathrm{C}(\hat{\beta}_2(t_i), \hat{\beta}_2(t_j)) & & \\ \vdots & \vdots & \ddots & \\ \mathrm{C}(\hat{\beta}_1(t_i), \hat{\beta}_q(t_j)) & \mathrm{C}(\hat{\beta}_2(t_i), \hat{\beta}_q(t_j)) & & \mathrm{C}(\hat{\beta}_q(t_i), \hat{\beta}_q(t_j)) \end{pmatrix}.$$

3.3.3 Estimation of the point-wise variance of residuals

Consider the regression model in matricial form

$$Y = X\boldsymbol{\beta} + \mathcal{E},$$

where we assumed the errors $\boldsymbol{\epsilon}(t_j) = (\epsilon_1(t_j), \ldots, \epsilon_n(t_j))$ to be such that

$$\mathbb{E}[\boldsymbol{\epsilon}(t_j)] = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\epsilon}(t_j), \boldsymbol{\epsilon}(t_k)) = \sigma_{jk} I.$$

In other words, we assume the errors to be uncorrelated, but evaluations of an error at different time instants to be correlated in principle. For clarity of notation, let $[\Sigma]_{jk} = \sigma_{jk}$ be the $(T \times T)$-dimensional covariance matrix of the errors evaluated at the sampling instants.

Define the residuals of the model in matricial form as

$$\hat{\mathcal{E}} = Y - \hat{Y} = Y - X\hat{\boldsymbol{\beta}}. \tag{3.17}$$

We want to use the information in the residuals to replace the population quan-

tity Σ by a reasonable estimate $\hat{\Sigma}$. If the regression were not penalized, an unbiased estimate for matrix Σ would be

$$\hat{\Sigma} = \frac{1}{n-q}\hat{\mathcal{E}}^\intercal\hat{\mathcal{E}}, \tag{3.18}$$

where $n-q$ are the *degrees of freedom* of the model [4] (Ramsay and Silverman, 2005, Johnson and Wichern, 2018).
As we resort to a penalized criterion, $n-q$ is not a corrected estimate of the degrees of freedom of our model and consequently estimator (3.18) is not unbiased. Because of this we are interested in getting an estimate of the *effective degrees of freedom (edof)* of our PWFLS regression that leads to a reliable estimate for the quantities $(\sigma_{11}, \dots, \sigma_{TT})$.

To achieve this, we first observe that applying the vec() operator to (3.17), one gets

$$\text{vec}(\hat{\mathcal{E}}) = \text{vec}(Y) - \text{vec}(X\hat{\boldsymbol{\beta}}) = \text{vec}(Y) - (I_T \otimes X)\text{vec}(\hat{\boldsymbol{\beta}}), \tag{3.19}$$

and that plugging (3.15) into (3.19) we obtain

$$\text{vec}(\hat{\mathcal{E}}) = (I_{nT} - (I_T \otimes X)P)\text{vec}(Y) = (I_{nT} - H)\text{vec}(Y). \tag{3.20}$$

Above, $P \coloneqq S_\Theta S_\beta S_\Phi$ for notational purposes.
Exploiting the similarity between formulation (3.20) and a classical multivariate regression, we consider the pointwise *residual sum of squares* (RSS) and look for a corrective term that makes $\text{RSS}(t_j)$ a useful estimate for the corresponding σ_{jj}. To clarify, fix the sampling instant t_1. Then the procedure illustrated in Appendix A.2 shows that a useful estimate for σ_{11} may be taken in the form

$$\hat{\sigma}_{11} = \hat{\text{Var}}(\hat{\boldsymbol{\epsilon}}(t_1)) = \frac{1}{\delta_1}\text{RSS}(t_1), \tag{3.21}$$

where $\delta_1 = \text{tr}\left[(I-H)^\intercal A^\intercal A(I-H)\right] = n - 2\text{tr}\left[H_{1:n,1:n}\right] + \text{tr}\left[H_{1:n,\cdot}(H_{1:n,\cdot})^\intercal\right]$.
We point out that estimate (3.21) is unbiased only under the far-fetched hypotheses of uncorrelated time points and of equal variances at the time points. Nonetheless, we consider δ_1 to be a reliable estimate of the *edof* of the model and hence use it as corrective term.

[4] Notice that q counts the intercept term too.

3.4 Bootstrap approach to assess the simultaneous variability of the functional coefficients

It is important to stress that the variability estimated in Equation 3.16 has pointwise validity, meaning that conclusions based on it can only be drawn one-at-a-time. If we are interested in obtaining an estimate that simultaneously quantifies the uncertainty of the $\hat{\beta}$'s over the whole domain, one option is to rely on resampling methods.
Such approaches are based on the use of data to randomly generate additional samples of a population and to investigate its distributional characteristics, avoiding the introduction of any strong assumption on the statistical model. Bootstrapping is one of these methods. First introduced by Efron (1979), the idea behind the method is to obtain an estimate of the distribution of an estimator by repeatedly sampling data with replacement and getting values of the estimator over the empirical distributions of the samples.

In the case of scalar data, the bootstrap method finds a theoretical foundation in the law of large numbers and in the Glivenko-Cantelli theorem, which combine to guarantee uniform convergence of the empirical distribution to the true distribution of the population. If an estimator T of a statistic θ is consistent, *i.e.* $T(\mathcal{F}) = \theta$ where $\mathcal{F}$ is the true distribution underlying data, then we are guaranteed convergence of the empirical distribution of the estimator to its true distribution.

Bootstrap resampling methods and results of asymptotic validity of the bootstrap methodology are extended in the framework of functional data analysis, where the distributional properties of the statistics are particularly problematic to handle (Cuevas, Febrero, and Fraiman, 2004, Politis and Romano, 1994, Cuevas and Fraiman, 2004). In particular, the work of Cuevas and Fraiman (2004) derive a result of bootstrap validity for functional statistics defined from differentiable operators. This result is crucial in justifying the use of a bootstrap approach to get an estimate of the distribution of the functional coefficients, and hence of their overall variability.
Indeed, recall that the operator bringing the functional observations into the functional coefficient estimates takes the form

$$\text{vec}(\hat{\boldsymbol{\beta}}) = S_\Theta S_\beta S_\Phi \text{vec}(\text{Y}).$$

We observe that the vec() operator and the projection maps S_Θ, S_β and S_Φ satisfy

the regularity conditions required by the result of Cuevas and Fraiman (2004), and so does their composition.

The resampling scheme we resort to is the following:

1. Estimate $\hat{\boldsymbol{\beta}}(t)$ of the regression model $\boldsymbol{y}(t) = X\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$,

2. Evaluate the residuals $\hat{\boldsymbol{\epsilon}}(t) = \boldsymbol{y}(t) - X\hat{\boldsymbol{\beta}}(t)$,

3. Randomly generate a bootstrap sample $\hat{\boldsymbol{\epsilon}}^*$ from the empirical distribution of the residuals and define the new pairings $\{(y_1^*, \boldsymbol{x}_1), \ldots, (y_n^*, \boldsymbol{x}_n)\}$ as

$$y_i^*(t) = \boldsymbol{x}_i^\intercal \hat{\boldsymbol{\beta}}(t) + \hat{\epsilon}_i^*(t).$$

4. Estimate $\hat{\boldsymbol{\beta}}^*(t)$ of the regression model $\boldsymbol{y}^*(t) = X\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$,

5. Repeat (3) and (4) for a sufficiently large number of times.

Once the samples of functional coefficients are generated, we visualize their distribution with functional boxplots. Since typically the observations that cross the fences are considered as outliers, we take the amplitude of the fences at the sampling points as useful estimate of the confidence that we have globally on the true coefficients.

It is relevant to point out that the literature presents more refined and compound techniques to simultaneous inference for functional parameters, that is not exploited in this work. Among the methods based on parametric bootstrap, Degras (2011) considers a function-on-scalar regression and proposes a parametric bootstrap method to build simultaneous confidence bands around the estimate of the functional coefficient. An extension of this work is provided by Chang, X. Lin, and Ogden (2017), who propose a wild bootstrap methodology to handle regression with multiple covariates and errors that are non-normal and heterogeneous. The simulation based method of Degras (2017) provides theory, method and implementation of simultaneous confidence bands for the mean, the quantiles, the covariance function and other functional parameters for functional data. Another parametric bootstrap method to simultaneous inference for functional data is proposed by Cao, Yang, and Todem (2012), who propose a spline estimator for the mean function of dense functional data, associated to a simultaneous confidence band which is asymptotically correct. Another approach to simultaneous inference consists in dimensionality reduction

based on functional principal component analysis and to build multivariate confidence ellipses (Goldsmith, Greven, and Crainiceanu, 2013).
The employment of these methodologies may be scope of future work that aims to accurately identify confidence bands for coefficient estimates or for other functional statistics.

Chapter 4

Case study: A functional Ground Motion Model for Italy

This chapter is devoted to the preliminary analysis, the end point of which is the calibration of the regression model. Here, we discuss the extension of the Ground Motion functional form to an infinite dimensional setting, the characteristics of the dataset, the application of the weighted FDA methods outlined in Chapter 3, and some crucial intermediate steps that complete the overall structure of the analysis. Finally, the soundness of the entire procedure is validated through a comparison with *state-of-the-art* techniques for the reconstruction of partially observed functional data.

4.1 Model formulation: functional ITA18

In Section 1.3, we recalled the functional form of the Ground Motion Model proposed in Lanzano, Luzi, Pacor, et al. (2019), henceforward referred to as ITA18. The model predicts the dependent variable Y, *i.e.* the logarithm of the RotD50 intensity measure, separately at 37 different period ordinates $\{T_j\}_{j=1}^{J}$, $T_j \in \mathcal{T} \coloneqq [0, 10s]$ for all j.

Now $\mathcal{Y}(\cdot)$ is assumed to be a functional variable defined over an interval $\mathcal{T}$ of vibration periods. Let Y be functional datum and $Y_1, \ldots, Y_n$ an independent and identically distributed functional dataset.

In this setting, a functional extension of ITA18 is straightforward and takes the form

$$\log_{10} \mathcal{Y} = a + F_M(M_w, \text{SoF}) + F_D(M_w, R) + F_S(V_{S30}) + \epsilon. \tag{4.1}$$

As already mentioned at the end of Chapter 1, $F_M(M_w, SoF), F_D(M_w, R)$ and

$F_S(V_{S30})$ are the *source*-, *path*- and *site*-related terms, respectively.

The *source* is specified as a step-wise linear function

$$F_M(M_w;T) = \begin{cases} b_1(T)(M_w - M_h) & M_w \leq M_h \\ b_2(T)(M_w - M_h) & M_w > M_h \end{cases},$$
$$F_M(\mathrm{SoF};T) = f_j(T)\mathrm{SoF}_j,$$

in which the straight line changes slope at the hinge magnitude M_h. SoF_j are dummy variables that account for the *style-of-faulting*. In particular, we specify $SoF_1 = SS$ strike slip, $SoF_2 = TF$ thrust faulting, and $SoF_3 = NF$ normal faulting. Note that the coefficient f_3, related to the normal faulting, is constrained to zero when the regression is performed.

The *path* term takes the form

$$F_D(M_w, R; T) = [c_1(T)(M_w - M_{\mathrm{ref}}) + c_2(T)] \log_{10}(R) + c_3(T)R,$$

where parameter M_{ref} is the reference magnitude. The first term of this summation accounts for the geometrical spreading of waves from a source, and the second term explains the anelastic attenuation. A dependence on magnitude is introduced in the first component of the geometrical spreading. Recalling what we anticipated in Section 1.4.2, R is a predictor that represents a correction of the pure Joyner-Boore distance, and is defined as $R = \sqrt{d_{JB}^2 + h^2}$, where h is the parameter of pseudodepth measured in kilometres.

Lastly, the *site*-related term has the form

$$F_S(V_{S30}; T) = k(T) \log\left(\frac{V_0}{800}\right),$$

where $V_0 = V_{S30}$ if $V_{S30} \leq 1500\ m/s$, $V_0 = 1500\ m/s$ otherwise. The introduction of an upper bound in the shear-wave velocity is upheld by a poor sampling of sites characterized by very hard rocks. Since little to no information is given in correspondence of such sites, the amplification here is assumed to be independent of the shear-wave velocity. For values of V_{S30} lower than 1500 m/s, the scaling with the spectral acceleration is assumed to be linear.

We stress that, in formulation (4.1), b_1, b_2, f_j for $j = 1, 2, 3$, c_1, c_2, c_3, k are functional coefficients, defined over domain $\mathcal{T}$ and representing the effects at vibration period T of the predictors on the response variable. Therefore, the statistical tool that naturally fits for their estimation is a function-on-scalar linear regression, relating the scalar predictors of magnitude, style of faulting, shear-wave velocity and

distance from the source, to the response function.

Parameters M_h, M_{ref} and h that appear in the functional form are known to be dependent on the spectral periods. For this reason, they are typically estimated through a preliminary step of non-linear regression. As a methodology of this kind is non-trivial when applied to a functional framework, we assume them to be constant and known, and define them as the mean value of their period-wise estimates from ITA18 model. Specifically, these parameters are fixed as

$$M_h = 5.7, \qquad M_{\text{ref}} = 4.5, \qquad h = 5.9 \text{ km}.$$

A substantial simplification of this kind will necessitate a sensitivity analysis, showing how much the a priori choice of parameters impacts the estimates and the variability of the model. We will address this issue later in the following chapter, when we discuss the results and the goodness of the regression model.

4.2 Dataset exploration

For this work, we rely on the same data that have been used for the calibration of ITA18. The dataset is derived from the Engineering Strong Motion (ESM) database (Lanzano, Sgobba, Luzi, et al., 2018, Lanzano, Luzi, Russo, et al., 2018), and includes the records of some small-magnitude events collected in the ITACA archive (Luzi, Pacor, and Puglia, 2017). Additionally, the inclusion of recordings of world-wide events increases the range of magnitudes up to a value of 8, and enriches the dataset with additional scenarios of strike-slip and thrust-faulting mechanisms associated to high-magnitude events. As the great majority of the data concerns the Italian soil (the included global events are the 8% of the total), the functional model is considered to be calibrated specifically for Italy.

4.2.1 Response variable

The functional dependent variable of the model, referred to as RotD50, is defined as the median of the distribution of a quantity, that is the linear combination of the two horizontal components of the spectral acceleration (SA) across all non-redundant azimuths. We are given with 5568 recordings of RotD50 [1], each one observed at 37 periods in $[0, 10\ s]$. Namely, each record is associated to 37 values, representing the measures of the soil acceleration at the registration periods $(T_1, \ldots, T_{37})$. Figure 4.1 provides a representation of the functional dataset.

[1] Disambiguation. For the sake of fluency and in accordance with the literature on this topic, this work will hereafter refer equivalently to spectral acceleration and RotD50.

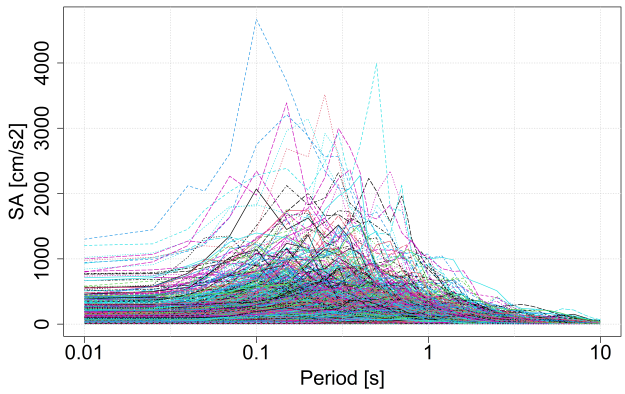
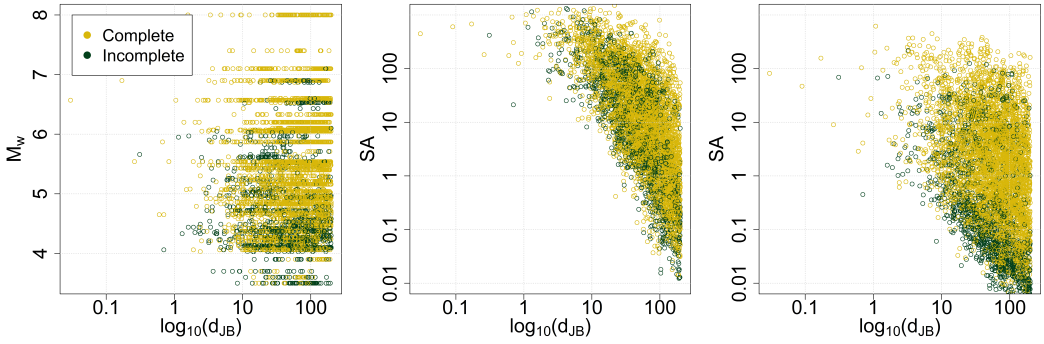


Figure 4.1: Representation of data of intensity measure as functions of the period.

Recall that Figure 1.2 in Section 1.3 showed that the 25% of the records of spectral acceleration is not completely observed on the whole domain. Now, the aim is to go a little deeper into the investigation of the missing values. Let us define *incomplete record* a curve that is not fully observed up to 10 s. In other words, a curve is incomplete if its value is missing at at least one period T in $[0, 10\ s]$. Figure 4.2 provides a visualization of the characteristics of incomplete records with respect to those that are complete. Figure 4.2a suggests that incomplete records tend to correspond to low-magnitude events recorded at medium to long distances, while Figure 4.2c shows how incomplete records correspond to values of soil motion that attenuate more rapidly with distance than complete records, and that this characteristic accentuates as the recording period increases.



(a) Magnitude vs d_{JB}. (b) SA vs d_{JB} at $T = 0.1\ s$. (c) SA vs d_{JB} at $T = 2\ s$.

Figure 4.2: Comparison between complete and incomplete records.

The availability of conspicuous information of this kind is useful in robustly calibrating the model for Italy and in stabilizing the estimates of the regression. For this reason and for the non-negligible ratio of incomplete records with respect to the

total, we are not willing to discard these data. Rather, we propose an alternative strategy, whose discussion is deferred to Section 4.3.1.

4.2.2 Prediction variables

Magnitude (M_w), style-of-faulting (SoF_j), Joyner-Boore distance (d_{JB}) and shear-wave velocity (V_{S30}) are the scalar variables that enter model (4.1). In this section, we are interested in qualitatively exploring their distribution and inner characteristics.

Figure 4.3 displays the frequencies of the faulting mechanisms and the densities of the other variables. The Joyner-Boore distance takes values in the $[0, 200$ km$]$ range and exhibits the greatest amount of recordings around the distance of 25 km from the source. The magnitude has peaks of frequencies in correspondence of about 4.3, 5 and 6 of the Richter scale, then it rapidly lowers at high magnitudes. The peak of observations corresponding to events of magnitudes 6 is due to the presence of global high-magnitude events in the dataset, whereas low-magnitude events are mainly registered in the Italian region. As far as the shear-wave velocity is concerned, it takes values up to 3000 m/s with very few points exceeding 1500 m/s. This stands as further confirmation of the justification provided for the choice of the threshold value, when we commented the functional form of the site term. Lastly, the histogram for the style-of-faulting shows that the majority of the records are associated to normal faulting, and that the less represented focal mechanism is the strike-slip. Such observation is in line with what is commonly known in the literature of Ground Motion models, namely that the normal faulting is the most frequent style-of-faulting in the Italian peninsula (Lanzano, Luzi, Pacor, et al., 2019). In fact, the inclusion in the dataset of worldwide events was also intended to extend the valid range of magnitudes for the less common focal mechanisms in Italy.

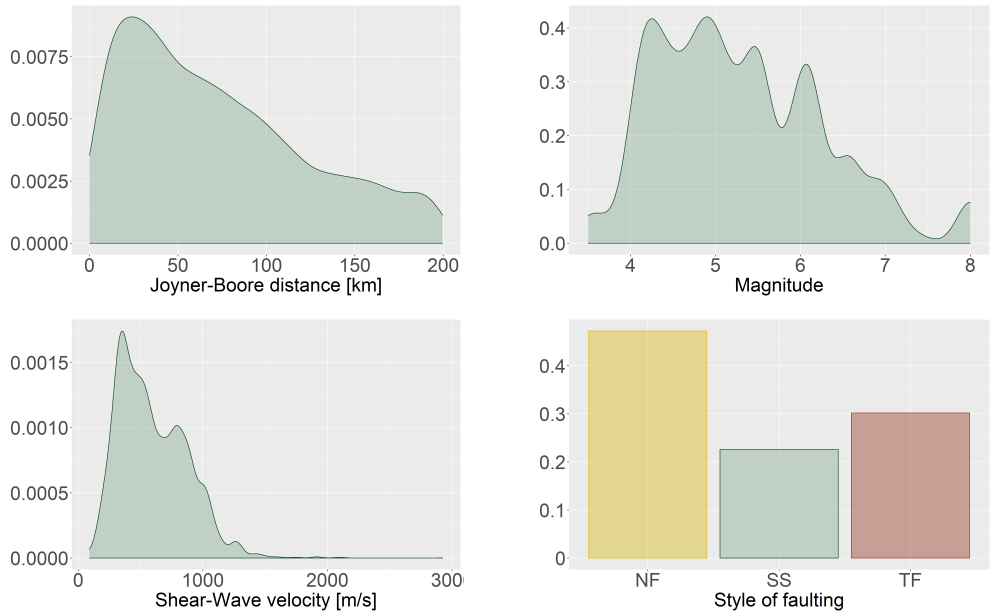


Figure 4.3: Empirical density of the seismological variables. Joyner-Boore distance [km], moment magnitude, shear-wave velocity $[m/s]$, style-of-faulting.

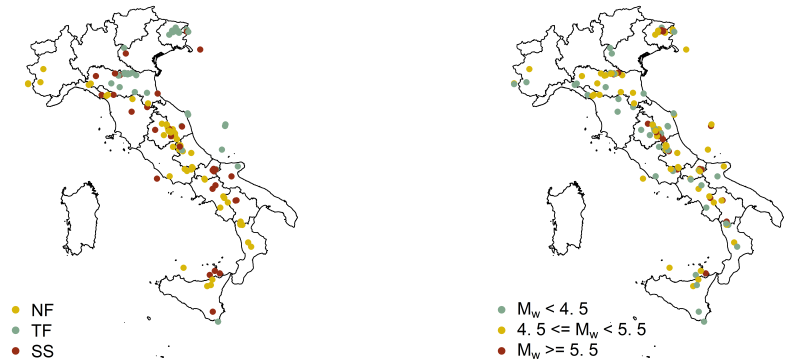
Given the inherent spatial nature of our data, it is appealing to provide a geographical representation of it. In particular, a localization of recordings along the peninsula permits to go deeper into the seismic information provided by the predictors.

The map on the left of Figure 4.4 shows the recordings sites, coloured according to the faulting mechanism of the event. The seismological characteristics of the peninsula are mainly due to the Adriatic plate, which is in contact with the Eurasian, African and Aegean plates[2] (Palano, 2014). The plate moves together with the African plate in direction North-North East, with a small counterclockwise rotation component. This movement causes the coexistence on the Italian territory of extensional, compressional and lateral-shift regimes, that correspond to normal faulting, thrust faulting and strike slip, respectively. The majority of normal-faulting recordings along the Apennines is due to the movement with respect to the African plate, while the thrust faulting in the North East of Italy and the strike-slip all along the peninsula are associated to the rotation of the plate.

The map on the right shows the recordings associated to the magnitude of the events, and in particular the geographical location of sites that recorded the soil motion after high-magnitude events. Among others, these events include the seismic sequences of central Italy in 2016 and 2017, l'Aquila in 2009, and the less recent

[2]The author likes to think of the Adriatic plate as a fish hook, which has the barb in Sicily, the shank along the eastern border of the Apennines and the eye, wider, supporting Northern Italy.

sequences of Umbria and Marche in 1997, Umbria in 1984, Irpinia in 1980 and Friuli in 1976. Recordings of these earthquakes play the crucial role of populating the portion of the dataset associated with critical hazard scenarios, namely ground motions in the vicinity of the earthquake epicentre, *i.e.* at a distance from the source lower than 30 km.



(a) Style-of-faulting (b) Magnitude

Figure 4.4: Geographical localization of the recordings of events along the Italian peninsula. (a) Recordings are coloured according to the style-of-faulting, (b) Recordings are coloured according to three classes of magnitudes: 1. $M_w < 4.5$, 2. $4.5 \leq M_w \leq 5.5$, 3. $M_w > 5.5$.

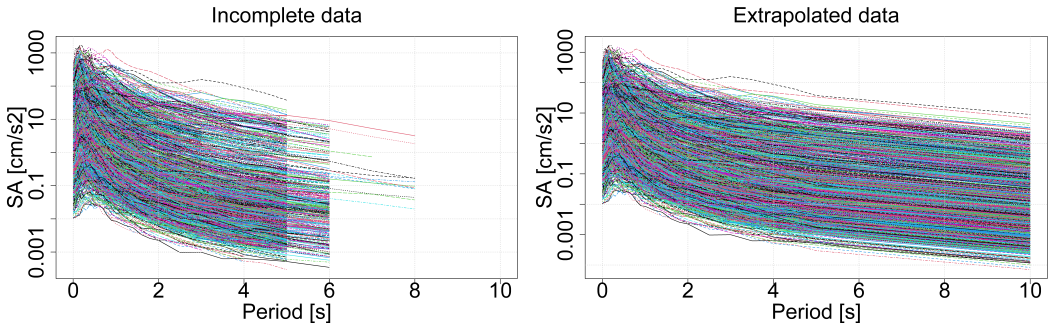
4.3 Workflow: from raw data to regression

In this section, we outline a scheme that leads the raw data to the estimation of the functional coefficients of the regression. A generic FDA workflow involves the smoothing of the raw data and then the estimation of the regression coefficients. The line of this work is the same, but deeply reworked to fit the context of incomplete records. In Chapter 3, we already detailed how smoothing and regression adapt into a weighted functional framework. Here, we meticulously describe also the minor, yet crucial, practical steps of the overall procedure, so as to provide a comprehensive picture of it.

4.3.1 Extrapolation of incomplete records and weights construction

First, it is necessary to handle the presence of incomplete data in the dataset.
The idea is to reconstruct each incomplete record by a simple linear extrapolation

of the curve from its last observed value $\bar{T}$, up to the last recording period $T = 10s$. The slope of the extrapolating line is chosen so as to be equal to the mean, over all complete records, of the slope of the straight line interpolating the values of the complete record at $\bar{T}$ and at $T = 10\ s$. This reconstruction choice is justified by the nature of the functional data, as it is clearly displayed in Figure 4.1. Most of the variability in the curves is concentrated in periods smaller than $1s$, whereas the rest of the domain is characterized by a slow decade of the soil motion to low and little varying values of spectral acceleration. Figure 4.5a displays the incomplete records, Figure 4.5b shows their reconstruction following the above specified procedure.



(a) Incomplete data as functions of period

(b) Extrapolated data as functions of period

In order to properly account for the reconstruction of curves in a subpart of their domain, we are interested in associating less, or more, confidence to observations that are the result, or not, of the extrapolation procedure.

To this aim, each curve i is associated to a specific functional weight w_i. The weight is chosen so as to reflect the reliability that we have on a certain portion of the curve. Where the curve is defined by measured values, the full reliability is represented through a weight that is set to 1. As we move more and more away from the last observed value, the reliability on the extrapolated values is corrected to become continuously smaller. A logistic function is a convenient choice to achieve a decrease in confidence from 1 to small values.

Suppose that observation i has values measured up to a period $T = \bar{T}$. Then we define

$$w_i(T) = \begin{cases} 1\ , & T \leq \bar{T} \\ \frac{1}{1+e^{(T-\mu_i)/s_i}} + \frac{1}{1+e^{(\bar{T}-\mu_i)/s_i}}\ , & T > \bar{T} \end{cases}\ , \tag{4.2}$$

where we set

$$\mu_i = \bar{T} + \xi,$$
$$s = \frac{1}{\mathrm{Var}(SA_{\mathrm{complete}}(\bar{T}))}.$$

The scale parameter s accounts for the rate of decay of the function. The smaller the scale, the more abrupt is the decrease of the weight to 0. Here, a low scale is associated to a great variability of the complete curves in $\bar{T}$, meaning that if a record is interrupted at a period characterized by large variability, then the confidence associated to the reconstruction quickly falls to 0. The location parameter μ identifies the point in which the weight is 0.5. The larger μ, the more far from $\bar{T}$ the confidence on the reconstruction is extended. The fact that the location depends on the last valid period of observation implies that weights are specifically defined according to the observation to which they are coupled. Figure 4.6 shows the weights defined for our partially observed data.

While s is fixed, ξ is an hyperparameter that has to be chosen among a set of values, through a cross-validation procedure that is detailed in the last part of the chapter.

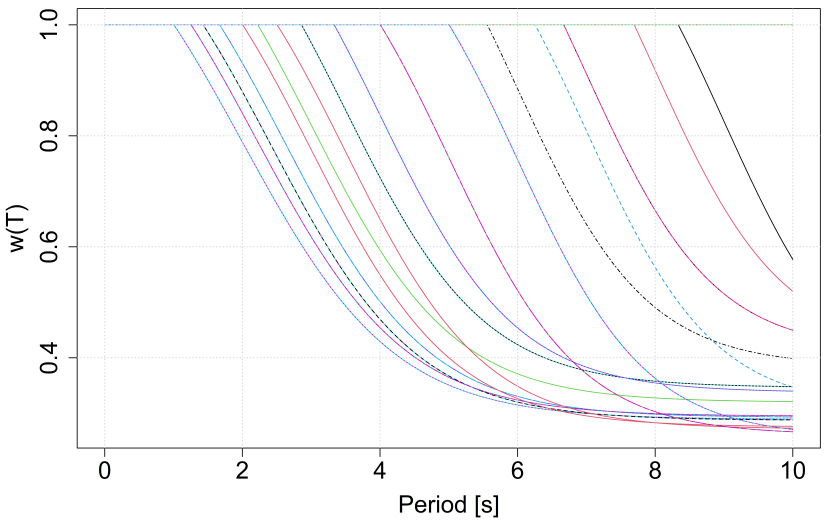


Figure 4.6: Logarithmic weights associated to the incomplete observations of spectral acceleration.

4.3.2 Selection of the penalization parameter

Second to extrapolation and the construction of the logistic weights, the raw data are smoothed on a B-spline basis with knots at the sampling periods, through the

penalized weighted smoothing technique illustrated in Section 3.2. Sticking to such methodology, each functional datum is smoothed separately according to the associated weight. Although the smoothing is performed record-wise, the penalization parameter ζ is selected via a common generalized cross-validation (GCV) procedure, so that the penalization on curvature is the same for every resulting smooth curve. The workflow of the GCV procedure is the following:

1. Pick ζ in a set $(\zeta_1, \ldots, \zeta_J)$ [3],
2. for each functional record i, perform a generalized cross-validation over its point-wise observations and evaluate the $\text{gcv}_i(\zeta)$ statistic, *i.e.* the error sum of squares discounted by a measure of regularity (Ramsay and Silverman, 2005),
3. Define $GCV(\zeta) \coloneqq \sum_{i=1}^{n} \text{gcv}_i(\zeta)$
4. Repeat (1)–(3) for every $(\zeta_1, \ldots, \zeta_J)$, and eventually select ζ_{opt} as the argmin of GCV among all ζ's.

4.3.3 Calibration

As for the smoothing, the regression performed through the penalized weighted functional methodology described in Section 3.3.1 requires the handling of the penalization parameter λ, that forces the regression coefficient functions to be smooth. However, differently to the smoothing that applies to all estimated curves the same penalty to roughness, there is no point in assuming that the functional coefficients should share the same level of regularity.

For the sake of clarity, the discussion in Section 3.3.1 has been developed considering a unique and common penalization parameter λ. As matter of fact, we generalize the functional form (3.14), showing that each regression coefficient function β_j can be associated to its own λ_j. Indeed, let us focus on the term that accounts for the roughness penalty $\lambda R \otimes I_q$. This term simply expands λR for the estimation of each functional coefficient and may equivalently be re-expressed as $R \otimes \lambda I_q$. Now it is immediate to see that if we replace λI_q with a matrix $\Lambda = (\lambda_1, \ldots, \lambda_q)$, then we succeed in introducing different penalty terms.

It is worth underlining that these λ's are q hyperparameters of the model, and that they are all manually specified when the least-squares methodology is exploited. All the more reason, therefore, to appropriately address the issue of developing a

[3] In the specific case of our analysis, after a previous analysis conducted with non-weighted smoothing, we identified $(10^{-8}, 10^{-3})$ as vector of possible values that identified the optimal ζ.

procedure that calibrates the vector $(\lambda_1, \ldots, \lambda_q)$ of penalty parameters. Let us give an idea of the calibration method adopted in this work.

First, we identify a sequence $(\gamma_1, \ldots, \gamma_N)$ of adequate values for the components of $(\lambda_1, \ldots, \lambda_q)$. The sequence is the same for all λ's. Then, since it is not feasible to try all possible dispositions with repetitions of the γ's into a q-length vector, we stick to the hereunder greedy approach:

For each $\gamma \in (\gamma_1, \ldots, \gamma_N)$, fix all penalty parameters $(\lambda_1, \ldots, \lambda_q)$ equal to γ. Hence, compute the corresponding Mean Squared Error (MSE) [4]. Let $\tilde{\gamma}$ be the value providing the lowest MSE. This preliminary step allows us to identify $\tilde{\gamma}$ as a convenient value to initialize the parameters with.
Indeed, initialize all components of $(\lambda_1, \ldots, \lambda_q)$ with $\tilde{\gamma}$.

1. Set $\lambda_1 = \gamma_1$ and let $\boldsymbol{\lambda}^{\gamma_1} = (\gamma_1, \lambda_2, \ldots, \lambda_q)$ be the corresponding vector of penalty parameters.
2. Evaluate the corresponding mean squared error $\mathrm{MSE}_{\lambda_1}(\gamma_1)$.
3. Repeat (1)–(2) for all $\gamma_j \in (\gamma_2, \ldots, \gamma_N)$.
4. Fix $\lambda_1 = \gamma^*$, where γ^* corresponds to the minimum of $(MSE_{\lambda_1}(\gamma_1), \ldots, MSE_{\lambda_1}(\gamma_N))$.
5. Repeat the procedure for all $\lambda_k \in (\lambda_2, \ldots, \lambda_q)$

4.4 Validation of the weighted analysis

The preceding sections of this chapter outline a course of action leading from the raw data to the estimation of the coefficients of the functional regression. To recap, the steps of the analysis are:

1. Reconstruction of the incomplete records

[4] Given a vector of penalty parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)$, the Mean Squared Error is computed resorting to the following cross-validation procedure.
For each training-test partition, fit the model on the training set by penalizing with $\boldsymbol{\lambda}$, then evaluate the fitted values on the test set. Compute the regression error for the j-th partition as

$$\mathrm{MSE}_j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} (y_i(s) - \hat{y}_i(s))^2 ds,$$

where $\mathcal{T}$ is the domain of definition of the curves.
Then the MSE is defined as the mean over all partitions of the MSE_j's.

2. Construction of the weights associated to the records

3. Weighted smoothing with ζ selected via GCV procedure,

4. Calibration of the penalization vector $\boldsymbol{\lambda}$

5. Estimation of the functional coefficients via penalized weighted least squares.

At this stage, we are unable to assess how the naive extrapolation of incomplete records, as described in Section 4.3.1, is an appropriate method for our analysis, nor whether it introduces a bias into the estimates. From these reasons, the need arises to make a comparison between the results obtained with the extrapolation and those attained through other reconstruction methods that are present in the literature and that, to the best of our knowledge, represent the state-of-the art of reconstruction methodologies for partially observed functional data. An evaluation of this kind allows us to understand whether extrapolation, when combined with a weighted functional analysis, is a robust approach to be applied in this case study and other similar frameworks. The theory and the general ideas behind these methodologies have been largely discussed in Section 2.2, and herein only some practical details are emphasized.

4.4.1 Comparative analysis with *state-of-the-art* reconstruction methods

The comparison is conducted among five different reconstruction methods:

- Extrapolation
- Kraus PCS: Estimation of the principal component scores (Section 2.2.2),
- Kraus FCHS: Functional completion with a Hilbert-Schmidt operator (Section 2.2.3),
- Kneip, Liebl PC: Functional completion with a reconstruction operator, with principal component decomposition (Section 2.2.4),
- Kneip, Liebl Align: Functional completion with a reconstruction operator, with alignment (Section 2.2.4).

It is necessary to make a clarification about the last two methods listed above. In Kneip and Liebl (2020), the authors recognize that, additionally to the problem of reconstruction, the estimation of the function underlying the sampled longitudinal points requires some effort. They propose two alternative ways to address

this: *(i)* resorting to the best basis property, use the truncated principal component decomposition to simultaneously provide estimates of the real curve both on the observed and the unobserved domains; *(ii)* operate a preliminary step of non-parametric smoothing. Since in general we are not guaranteed continuity of the non-parametric estimate and the optimal operator at a boundary point, the reconstruction is corrected at the boundary in order to resort continuity (hence the term *align*).

Of all methods listed above and already present in the literature, we were required to manually implement the estimation of the principal component scores. The other scripts are available in the `R package` ReconstPoFD, which can be downloaded and installed from the GitHub account of Dominik Liebl.

The scheme behind the comparison procedure is simple: for each reconstruction methodology, perform the course of action synthesized at the beginning of this section and, downstream of the operational flow, evaluate the corresponding MSE between the original, possibly incomplete, records and the responses predicted by the regression.
In particular, the MSE is evaluated through an event-wise cross-validation, *i.e.* a classical cross-validation that separates observations so that the test set contains records of events that are unobserved in the training set[5]. For each original record y_i observed on a domain O_i, the corresponding squared error err_i is evaluated as

$$\text{err}_i = \frac{1}{|O_i|} \int_{O_i} (y(s) - \hat{y}_i(s))^2 ds.$$

Then, the MSE_j for the j-th training-test partition is given by

$$\text{MSE}_j = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \text{err}_i,$$

and the overall MSE of the method is evaluated by taking the mean of the MSE_j.

The pseudocodes provide an effective scheme of the entire procedure discussed above.

[5]As we work under the ergodic assumption with data that are known to be *event* and *site* dependent, a partition of this kind allows to obtain a more reliable, less underrated estimate of the true error.

Algorithm 1: Analysis

Input: *method*, location, $Y_{\text{train}}, Y_{\text{test}}, X_{\text{train}}, X_{\text{test}}$;

Set $Y \leftarrow Y_{\text{train}}$ for steps (1), (2), (3), (4), (5).

1. Reconstruct using the reconstruction *method*

$Y_{\text{rec}} \leftarrow$ reconstruction(*method*, Y)

2. Create weights

weights $\leftarrow$ create.weights(Y, location = *method*\$location)

3. Weighted smoothing

$Y_{\text{smooth}} \leftarrow$ weighted.smoothing(Y_{rec}, weights)

4. Calibration of the optimal λ

$\lambda_{\text{list}} \leftarrow$ lambda.selection($Y_{\text{smooth}}, X_{\text{train}}$)

5. Weighted functional regression

model $\leftarrow$ weighted.regression($Y_{\text{smooth}}, X_{\text{train}}, \lambda_{\text{list}}$, weights)

6. Evaluation of the MSE on X_{test}

$\hat{Y}_{\text{test}} \leftarrow$ predict(model, X_{test})

$$\text{MSE} \leftarrow \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \frac{1}{O_i} \int_{O_i} (Y^i_{\text{test}}(t) - \hat{Y}^i_{\text{test}}(t))^2 dt$$

Output: MSE

Algorithm 2: 10-fold cross-validation

Set $n_{\text{test}} = \frac{n}{10}, n_{\text{train}} = n - n_{\text{test}}$;

for $k = 1, \ldots, 10$ **do**

1. Split train and test sets

Y^k_{train} and Y^k_{test}

X^k_{train} and X^k_{test}

2. For each method, apply the analysis

for *method* $\in \{methods\}$ **do**

$$\text{MSE}^k_{method} \leftarrow \text{analysis}(method, Y^k_{\text{train}}, Y^k_{\text{test}}, X^k_{\text{train}}, X^k_{\text{test}})$$

for *method* $\in \{methods\}$ **do**

$$\text{MSE}_{method} \leftarrow \frac{1}{10} \sum_{k=1}^{10} \text{MSE}^k_{method}$$

Method	MSE	$\text{MSE}_{\text{ratio}}$
Extrapolation, ξ=0	0.10360	1.00045
Extrapolation, ξ=1	**0.10355**	**1.00000**
Extrapolation, ξ=2	0.10361	1.00063
Extrapolation, ξ=3	0.10360	1.00045
Extrapolation, ξ=4	0.10356	1.00005
Kraus PCS, ξ=1	0.10758	1.03886
Kraus FCHS completion, ξ=1	0.10437	1.00784
Kneip, Liebl PC, ξ=1	0.10491	1.01310
Kneip, Liebl Align, ξ=1	0.10384	1.00280

Table 4.1: Comparison of MSEs between reconstruction methods.

Before discussing the results of the comparison, we need to make an additional clarification concerning the location parameter ξ that defines the weights. As we mentioned in Section 4.3.1, ξ is a hyperparameter that has to be selected through a cross-validation procedure. Indeed, the approach that we adopt consists in evaluating a set of MSEs for the extrapolation method, associated to a sequence of possible candidates for the location parameter. After selecting the best candidate $\xi*$, we proceed to the second part of the comparison, that confronts the extrapolation with the other methods, all being coupled to $\xi*$.

Table reports the results of the comparison.
As we may notice, the extrapolation method associated to weights with $\xi = 1$ is the one providing the smallest MSE. These statistics show whether the differences among the reconstruction methods are absorbed – or not – by the weighted functional analysis. Actually, as the values of $\text{MSE}_{\text{ratio}}$ [6] do not show appreciable divergences, this result reassure us about the robustness of a weighted approach with respect to the missing part of the response variable, and *de facto* justifies our choice to rely on such technique.

[6]We define the $\text{MSE}_{\text{ratio}}$ of a method as the ratio between the MSE evaluated for the method and the minimum among all MSE's.

Chapter 5

Case study: Results and diagnostic

Following all preliminary steps of analysis illustrated in the previous sections, this chapter is devoted to the discussion of the results. Here, we aim to combine model diagnostic with a seismological interpretation of the quantities involved, based on the related literature.
First, we comment on the estimated regression coefficients and on the *goodness-of-fit*. Then, we focus on the ground motion predictions and perform a sensitivity analysis on the hyperparameters of the model, *hinge magnitude*, *reference magnitude* and *pseudodepth*. Finally, the functional model is compared to its scalar counterpart, ITA18.

5.1 Model estimates

This section concentrates on the coefficients estimates appearing in (4.1). Before commenting directly on the form of the coefficients, a multicollinearity analysis allows us to highlight the presence of correlation between the regressors and to get a prior, qualitative idea of the uncertainty associated to the estimates.

5.1.1 Multicollinearity analysis for the regressors

The analysis of multicollinearity among the predictors is a statistical tool that investigates the uncertainty associated to the estimates of the regression coefficients. The easiest way to detect the presence of collinearity is through the inspection of the correlation matrix, that is showed in Figure 5.1. As expected, there is almost perfect positive correlation between the predictors associated to the magnitude-independent

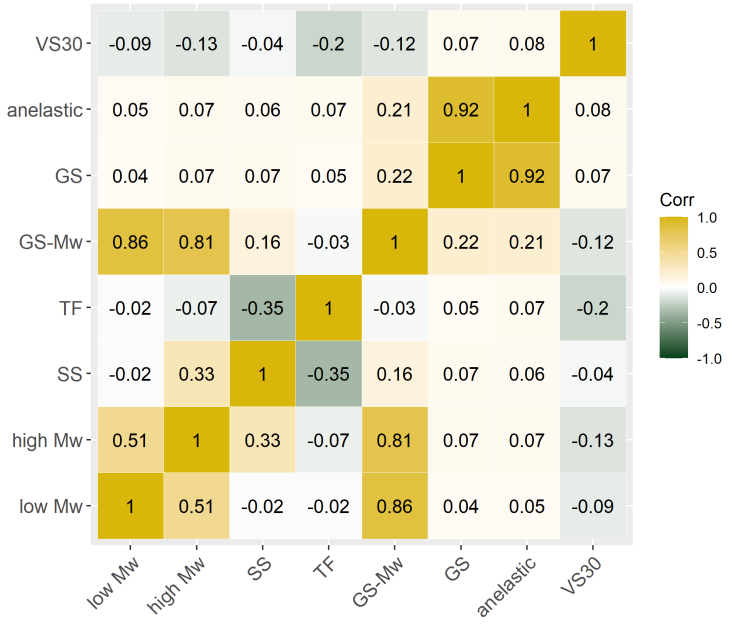


Figure 5.1: Correlation matrix of the predictors.

low M_w	high M_w	SS	TF	GS-M_w	GS	anelastic	V_{S30}
10.68	7.89	1.38	1.24	25.52	7.04	6.36	1.09

Table 5.1: Values of the Variance Inflation Factor for each regressor.

geometrical spreading (GS) and the anelastic attenuation, since both terms depend exclusively on the Joyner-Boore distance. High positive correlation is also present between low and high magnitude regressors and the term accounting for magnitude-dependent geometrical spreading (GS-M_w).

In presence of multicollinearity, *i.e.* when collinearity exists between three or more variables, the Variance Inflation Factor (VIF) quantifies how closely these variables are related with one another, and complements the correlation matrix in capturing associations between more than two predictors. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates problematic levels of multicollinearity, revealing the necessity for model reduction or variable selection. In agreement with the results of the correlation matrix, Table (5.1) reports critical VIF values for the variables mentioned above, and does not uncover correlations for the others (SS, TF, V_{S30}).

Multicollinearity makes it problematic to separate the individual effects of the predictors on the response, leading to an increased variability in the coefficients estimates. Typically, multicollinearity is handled through model reduction – *i.e.* variable selection, combination of collinear variables – and regularization. Although there are extensions of model reduction to functional data analysis, both in the frequentist (Horváth and Kokoszka, 2012, Ramsay and Silverman, 2005) and in the Bayesian (Mehrotra and Maity, 2019) contexts, the functional formulation of our model does not lend itself to the use of such techniques. Indeed, the physical interpretability of the regressors in (4.1) allows one to comment on the results in seismological terms, and to compare them with those of other functional forms, present in the literature, where the same terms appear. For this reason, we do not perform model reduction in this work. Nonetheless, we point out that the introduction of a penalization parameter in the least squares criterion (Section 3.3.1) works as regularization technique, that controls the effects of multicollinearity by reducing the variability associated to the coefficients estimates.

5.1.2 Estimated functional coefficients

In this section, we report the functional coefficients estimated by our model. Figure 5.2 shows the estimates of b_1, b_2, c_1, c_2, c_3 and k, each one associated to the functional boxplot of a bootstrap sample of dimension $B = 1000$ generated from its empirical distribution. The bootstrap sample is obtained following the procedure discussed in Section 3.4. By doing so, the idea is to see the scatter of the sample around the functional estimate as measure of its related uncertainty. The smaller the scatter, the narrower is the distribution of the coefficient around its true value and the lower is the uncertainty associated to the estimate. Notice that, by relying on a bootstrap sample to account for estimates variability, the confidence we get is simultaneous over the whole domain. In the plot, the dashed red line corresponds to zero. We will say that an estimate at T is significantly different from zero if the fences of its functional boxplot at T do not include zero.

Coefficients b_1 and b_2, plotted in the two top left panels, capture the linear dependence of ground motion on low magnitudes and high magnitudes respectively. Both have a positive impact on spectral acceleration that grows in the interval [0, 1 s] and then remains more or less constant until $T = 10$ s.

In the top left panel, coefficient k accounts for the negative scaling of ground motion with the shear-wave velocity. A common issue with this coefficient lies in its instabilities at short periods, where it may get very close to zero or even be positive, conversely to what is observed at all other periods. In our case, the instability is

not pronounced and k remains significantly negative for all T.

The second line of plots displays the coefficients related to the attenuation of ground motion with distance, namely c_1, c_2 and c_3. At all periods, c_2 captures the linear decay of the spectral acceleration with d_{JB}. Coefficient c_1 complements c_2 in capturing the magnitude dependence of geometric spreading due to finiteness of large magnitude ruptures. As expected, c_1 takes positive values to simulate the more gradual decay in near-source distances from large ruptures, and increases towards longer periods (Kotha, Weatherill, et al., 2021). Finally, c_3 accounts for the exponential decay of ground motion with distance, that is the anelastic attenuation. As we may see from the graph, anelastic attenuation affects ground motion at short periods, and its effect vanishes at longer periods. It is crucial to always obtain non-positive values of c_3, since it would indicate an unphysical exponential increase with distance.

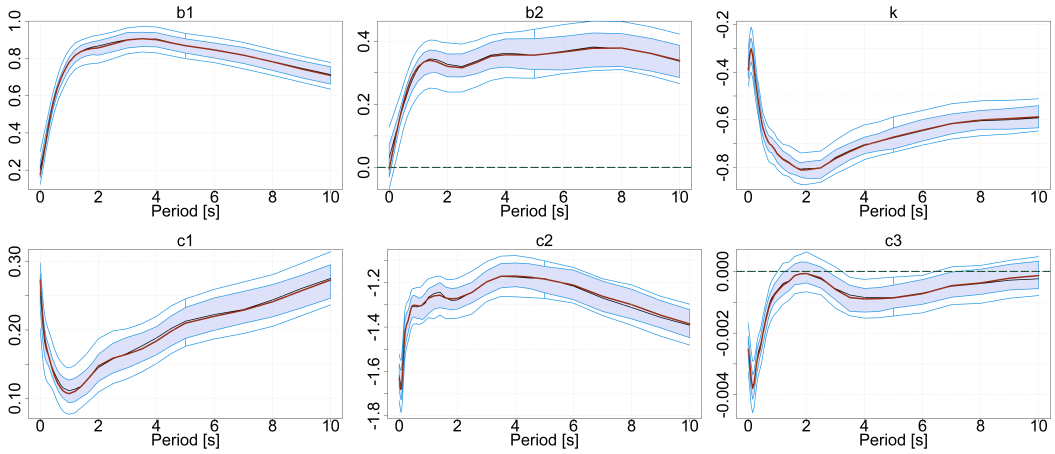


Figure 5.2: Estimated functional coefficients. The red line represents the point estimate of the coefficient. The azure lines correspond delimit the central region and the fences of the functional boxplot. The green dashed line marks zero.

A comment on the coefficients f_1 and f_2 is made separately for two main reasons. Firstly, the faulting mechanism is known to have little impact on the standard deviation of GMPEs (Bommer, Douglas, and Strasser, 2003, Lanzano, Luzi, Pacor, et al., 2019), and to be included in the functional form for purposes of seismic hazard assessment, rather than to get a better performance of the regression model (Section 1.4.1). Secondly, as found in the work of Lanzano, Sgobba, Caramenti, et al. (2021), coefficient f_2 is dependent on the region where the event occurs. Because of this, an ergodic model – as ours is – that neglects both between-event and site-to-site variabilities and does not include any spatial dependence in the coefficients is not expected to capture the effects of the thrust-faulting, to which f_2 is associated. Nev-

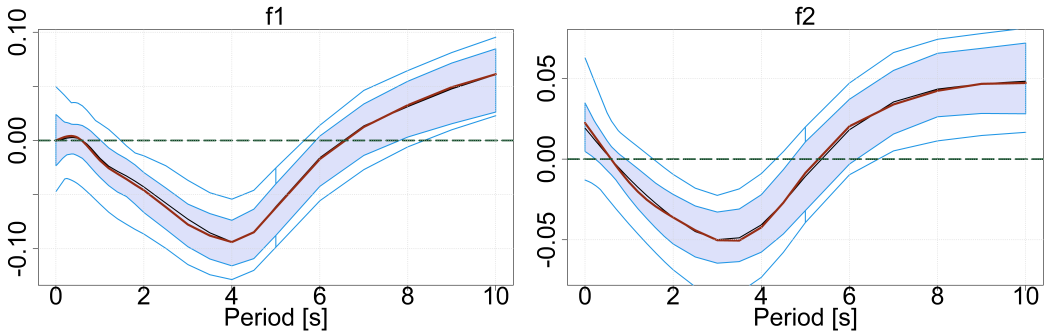


Figure 5.3: Estimated coefficients related to *style-of-faulting*.

ertheless, some indicators help to comment on the estimates. According to Bindi et al. (2011), the main differences in the ground motion due to the faulting mechanism should result in the short-period range – *i.e.* $T < 1$ s – wherein the expected values produced by thrust-faulting are significantly larger than those associated to normal faulting and strike-slip. Here, we partially observe a behaviour of this kind: indeed, the trends of the coefficients in Figure 5.3 are similar in every part of the domain but in the range $[0, 1]s$, where $\hat{f}_2$ is positive and $\hat{f}_1$ is not significant. As matter of fact, we notice that this trend is not fully consistent with that of Bindi et al. (2011), since *(i)* we are not confident that f_2 is different from zero at short periods, *(ii)* there is a gap in the coefficients magnitude at larger periods, implying that a difference among the faulting mechanisms is not only observable for $T \leq 1$ s. Summing up, we do not consider the estimates of f_1 and f_2 to be meaningful, nor to have a trend that is fully consistent with what is observed in the literature. For these reasons, it may be useful to complicate the model with additional terms that reduce the uncertainty associated to these estimates, by allowing for a regionalization of the effects of style-of-faulting.

5.1.3 Goodness-of-fit

In this section, we discuss the behaviour of the residuals to assess whether the functional formulation is able to capture the variability of the dataset. In the scalar case, a good fitting is graphically checked through a scatterplot of the regression residuals against the prediction variables, that shows if some dependence is left unaccounted by the model. If the dots form a cloud around zero with uniform variability along the horizontal axis, then we conclude that the regression succeeds in representing the effects of the independent variables. The functional counterpart

of the scalar scatterplot is displayed in Figure 5.4, in which the integral means [1] of the residuals are plotted against the input variables. As we may notice, the dots are symmetrically distributed around 0 and do not show any particular trend or change in dispersion along the axis. The bottom-left image shows the integral means of the residuals against the integral means of the fitted values. This is intended to check if all amplitudes of ground motion are associated to residuals of the same magnitude and variability. Since also in this case we do not detect any dependence, we conclude that the model does good in capturing the variability of the dataset.

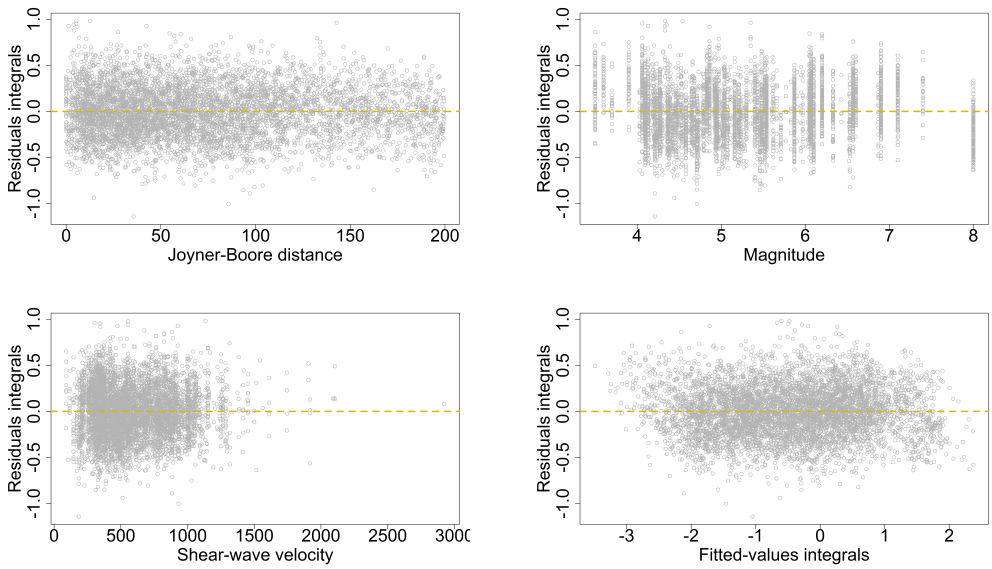


Figure 5.4: Goodness-of-fit. Integral mean of the residuals against the three continuous independent variables: Joyner-Boore distance, magnitude, shear-wave velocity, and against the integral mean of the fitted values. The yellow dashed line marks zero.

5.2 Regression results

In this section, we comment the regression results separately for *source*, *path* and *site* terms, with an eye on how their behaviour is affected by the vibration period. Specifically, the results corresponding to each soil component are displayed at a sequence of increasing vibration periods, that we believe to effectively catch their overall trend with respect to T. Then, a sensitivity analysis is performed for the

[1] Given a function $y : \mathcal{T} \to \mathbb{R}$, its *integral mean* is

$$\int_{\mathcal{T}} y(s)ds.$$

parameters of *hinge magnitude* (M_h), *reference magnitude* (M_{ref}) and *pseudodepth* (h).

5.2.1 *Source*

Figure 5.5 displays the behaviour of the *source* term with respect to the magnitude. The points are the values of spectral acceleration observed for the data used for calibration, regressed by the predicted effects of the *path* and *site* terms, while the green lines represent the values of the soil motion component $F_M(M_w, SoF; T)$ predicted by the regression.

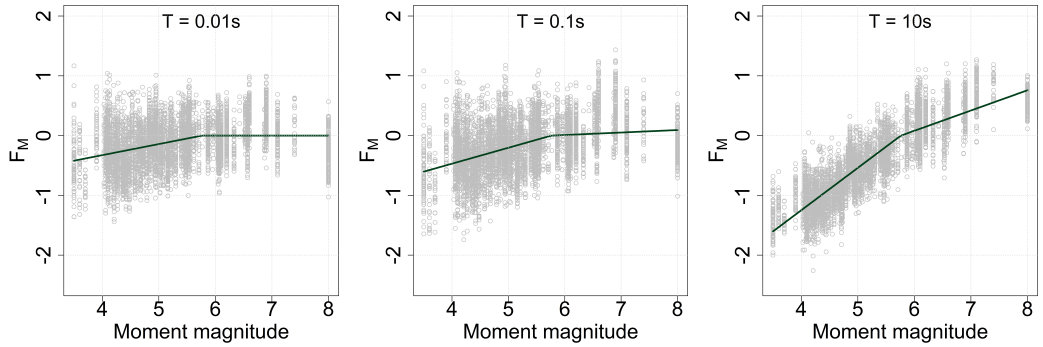


Figure 5.5: Predictions of the *source* term at periods $T = \{0.01s, 0.1s, 10s\}$, versus M_w. The dots are the observed values of spectral acceleration, regressed by the predicted effects of the *path* and *site* terms. The green line represents the predicted effects of the *source* term.

The figure provides a clear representation of the functional form chosen for the source, that is a bilinear function with two terms intersecting at the *hinge* magnitude M_h. As expected, there is a positive dependence of the spectral acceleration with magnitude. This dependency decreases for values of M_w larger than the M_h, showing the presence of a saturation effect. We observe that saturation is accentuated at short periods, and that it manifests itself in the flattening of the line at periods around $T = 0.01\ s$, indicating no dependency of the *source* term on magnitude. Although such behaviour may suggest *oversaturation*[2], the presence of the magnitude as explanatory variable in the *path* term acts as a compensation, recovering the physical soundness within our model. This has been checked by plotting the scaling of the intensity measure with distance, at magnitudes varying in a range of $[4.0, 8.0]$ and for all different faulting mechanisms. Figure 5.6, corresponding to the strike-slip faulting mechanism, shows how the increase of ground motion with magnitude

[2]We refer to *oversaturation* as the anti-physical behaviour of a model that show a negative dependence of ground motion on the moment magnitude M_w, for some values of $M_w > M_h$.

is preserved at all distances and also in scenarios that are poorly sampled, like short distances and high magnitudes.

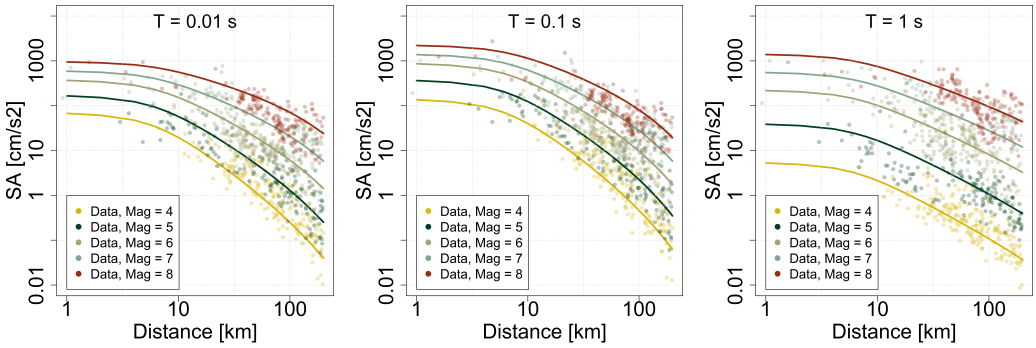


Figure 5.6: Prediction of SA for magnitude-varying scenarios, at periods $T = \{0.01s, 0.1s, 1s\}$. Every coloured line corresponds to a value of magnitude M_w, ranging in $\{4.0, 5.0, 6.0, 7.0, 8.0\}$. The dots represent the observed values of spectral acceleration, regressed by the predicted effects of the *path* and *site* terms. Each dot is coloured according to the magnitude of the recorded event, *e.g.* yellow if $M_w = 4 \pm 0.3$.

It is important to acknowledge that the behaviour of the source term at short periods is part of a larger discourse, much debated in literature, that involves the choice of the most appropriate functional form and of a suitable value of hinge magnitude, in relation to the occurrence of the oversaturation effect. The discussion is primarily about the most convenient choice of a functional form that simultaneously allows to: *(i)* capture the non-linear scaling of SA with magnitude, *(ii)* avoid oversaturation. The work of Fukushima (1996) adds a quadratic term for source scaling at low magnitudes, and shows that this term should be associated with a negative coefficient if the magnitude scale considered is M_w, in order to avoid oversaturation at short periods. Accordingly, both Kotha, Bindi, and Cotton (2014) and Kotha, Bindi, and Cotton (2016) include the quadratic term and show that the corresponding coefficient is negative. On the contrary, the model of Boore et al. (2014), that adopts the same functional form as in Lanzano, Luzi, Pacor, et al. (2019) with the addition of the quadratic term, reports a small but positive associated coefficient. In Kotha, Weatherill, et al. (2020), the authors present a ground motion model for Europe that shows a strong positive trend in the coefficient of the quadratic term, resulting in estimates of the short-period spectral acceleration at near-source distances ($d_{JB} \leq 20$ km) that are lower in the range $M_w \geq 6.5$ than in the range $5.7 \leq M_w \leq 6.2$. In order to accommodate this, the model has been revised one year later, with a lowering in the hinge magnitude parameter down to $M_h = 5.7$ that was originally set at $M_h = 6.2$.

To mention other functional forms for the source term, Chiou and Youngs (2014)

adopts hyperbolic magnitude scaling functions, whose effects are, however, difficult to compare with the simpler quadratic and bilinear forms.

In Lanzano, Luzi, Pacor, et al. (2019), the authors argue that the bilinear shape is a convenient choice, since it allows to better control oversaturation without the need to manually lower M_h, nor add other constraints to solve non-physical behaviours. Resorting to a bilinear functional form, the issue is solved within the model through an effective compensation operated by the *path* term.

All things considered, we shall keep in mind that the issue remains open, and that a univocal solution has not been identified yet. Those who adopt a quadratic form argue that the reason behind the oversaturation can be found, among others, in the poor sampling at near-source distances from large magnitude events. Indeed, the strong imbalance in terms of number between these kind of observations and those at low magnitudes would cause the model to seek for a better fit at $M_w \leq M_h$. From this perspective, the calibration of quadratic models on more recent and more complete datasets, such as the NEar Source Strong motion (NESS, Sgobba, Felicetta, et al., 2021), would possibly allow to overcome the oversaturation issue.
Another viable solution, and focus of future work, may reside in the formulation of models which impose a sign constraint on the coefficient estimate, effectively preventing the occurrence of oversaturation.

5.2.2 *Path*

Figure 5.7 displays the attenuation of ground motion with the distance from the source, at two different magnitudes $M_w = 4.0$ and $M_w = 6.0$. Similarly as for Figure 5.5, the dots are obtained by removing *source* and *site* effects from the responses, whereas the curves represent the soil motion component $F_D(M_w, SoF; T)$ fitted by the regression model, at $M_w = 4.0$ and $M_w = 6.0$. As expected, we observe that the attenuation is more rapid at short periods and that it decreases as the period increases (Lanzano, Luzi, Pacor, et al., 2019, Douglas, 2003). Secondly, we observe the compensation operated by this term commented in the previous paragraph; indeed, the attenuation reduces as the magnitude of the earthquake increases, resulting in the overall amplification of ground motion with magnitude that we checked in Figure 5.6. The dependence of the decay rate on magnitude is a peculiarity of the phenomenon of attenuation that was first detected and investigated in Campbell (1981).

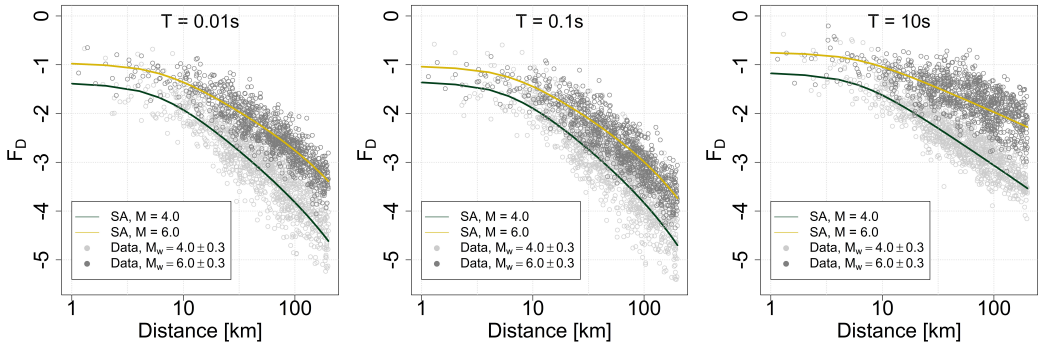


Figure 5.7: Predictions of the path term at periods $T = \{0.01s, 0.1s, 10s\}$, versus d_{JB}. The green and yellow lines is the predicted effect of the *path* term for $M_w = 4.0$ and $M_w = 6.0$, respectively. The dots are the observed values of spectral acceleration, regressed by the predicted effects of the *source* and *site* terms, coloured light if $M_w = 4.0 \pm 0.3$, dark if $M_w = 6.0 \pm 0.3$.

5.2.3 *Site*

Figure 5.8 shows the scaling of ground motion with the shear-wave velocity representing the local soil conditions of the site, at periods $T = 0.01s, 1s, 10s$. Recalling that higher values of shear-wave velocity are measured at rock and stiff sites and that lower values are observed at soil sites, then the plots illustrate a well known feature of ground motion models. Low values of shear-wave velocity correspond to larger amplitudes of the spectral acceleration, and higher values of V_{S30} to smaller amplitudes, as expected (Douglas, 2003). In this case, we do not identify a peculiar trend in the slope with the period.

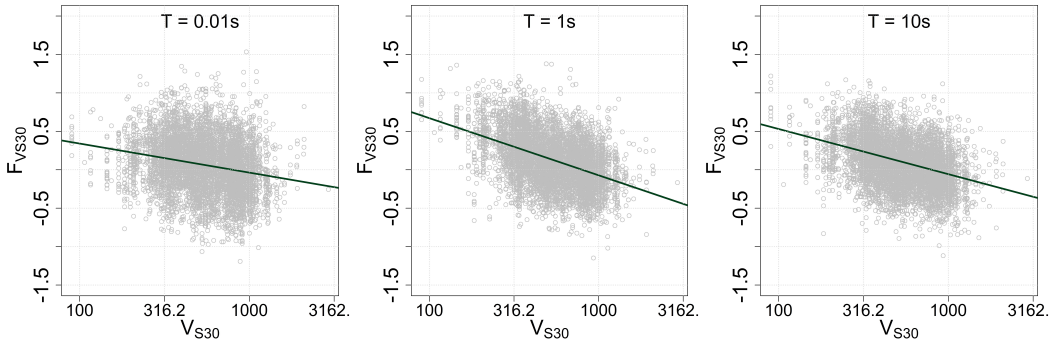


Figure 5.8: Predictions of the site term at periods $T = \{0.01s, 0.1s, 10s\}$, versus V_{S30}. The dots are the observed values of spectral acceleration, regressed by the predicted effects of the *source* and *path* terms. The green line represents the predicted effects of the *site* term.

As mentioned in the comment for coefficient k, its instabilities at short periods could make it become zero or even positive. This would result in an unexpected

increase of the soil motion at rock sites. Since the irregularity is sufficiently limited in our case, we do not observe an undesirable behaviour of the site term. Instead, what we observe is a confirmation of something that is already known in the literature, namely that soil sites are expected to show significantly larger response spectral amplitudes than rock sites at almost all periods of engineering interest, with the maximum ratio occurring around $T = 1\ s$ (Douglas, 2003). In the specific of our model, the minimum occurs at $T = 2\ s$.

5.2.4 Sensitivity analysis on M_h, $M_{\mathbf{ref}}$ and h

This section is devoted to the assessment of the impact that parameters M_h, M_{ref} and h have on the outcome of the regression. Our model considers them to be fixed and equal to the average of their period-dependent counterpart used for the calibration of ITA18.

To the best of our knowledge, a general approach for the identification of the hinge magnitude does not exist in the literature. After a preliminary step of non-linear regression, Lanzano, Luzi, Pacor, et al. (2019) define M_h to be a period-dependent step-wise parameter, taking values in the 5.5-6.5 range. This choice causes the presence of jumps in the prediction of the spectrum for scenarios close to hinge magnitude. To smooth the discontinuities in the estimates, the work of Sabetta et al. (2021) corrects M_h to have a smoother variation in the range of periods $[0.25s, 0.7s]$. In Kotha, Weatherill, et al. (2021), the lowering of the hinge magnitude from 6.2 to 5.7 solves the issue of the previous model that associated a negative dependence of ground motion on large magnitudes.

In our case, a sensitivity analysis that varies M_h in $\{5.5, 5.7, 6.2, 6.5\}$ does not clearly indicate which value is best. The left and the right panels of Figure 5.9 show that both the point-wise mean squared error (PMSE)[3] and the estimated standard deviation $\hat{\sigma}$ [4] of the model are not affected by M_h at short periods, while they reduce for $M_h = (6.2, 6.5)$ at periods larger than 1. More precisely, the lowest value of MSE and $\hat{\sigma}$ correspond to $M_h = 6.5$. We observe that the greatest reductions in $\hat{\sigma}$ and in MSE with respect to their value for $M_h = 5.7$ are 3% and 10% respectively, both occurring at periods around $T = 7\ s$ for $M_h = 6.5$.

[3]The pointwise mean squared error is evaluated as the mean of the mean squared errors produced by a cross-validation procedure, that separates training and test sets event-wise and for each partition j evaluates

$$MSE_j(t) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (\hat{y}_i(t) - y_i(t))^2.$$

[4]The evaluation of $\hat{\sigma}$ is done following the argument of Section 3.3.3.

In light of the little changes in MSE and $\hat{\sigma}$, we are interested in checking whether one of the values of M_h leads to oversaturation effect, as the occurrence of this phenomenon would make us discard the value. Figure 5.10 shows that $M_h = \{6.2, 6.5\}$ are associated to a negative slope for $M_w > M_h$, suggesting oversaturation at short periods. Actually, by repeating the magnitude scaling analysis made at the end of Section 5.2.1, we observe that neither $M_h = 6.2$ nor $M_h = 6.5$ are associated to such non-physical behaviour (see Appendix B.1.1).

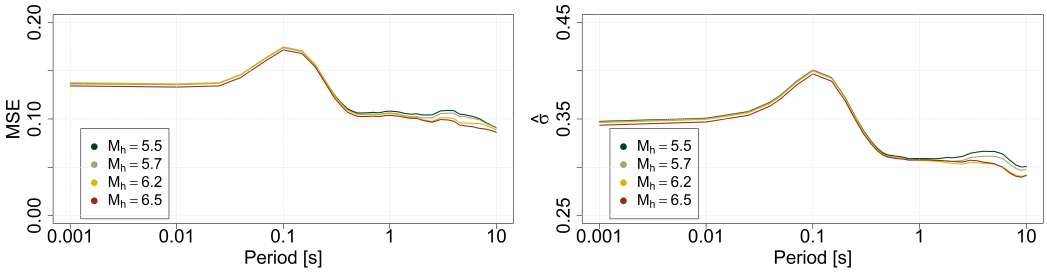


Figure 5.9: MSE (left) and $\hat{\sigma}$ (right) for $M_h = \{5.5, 5.7, 6.2, 6.5\}$.

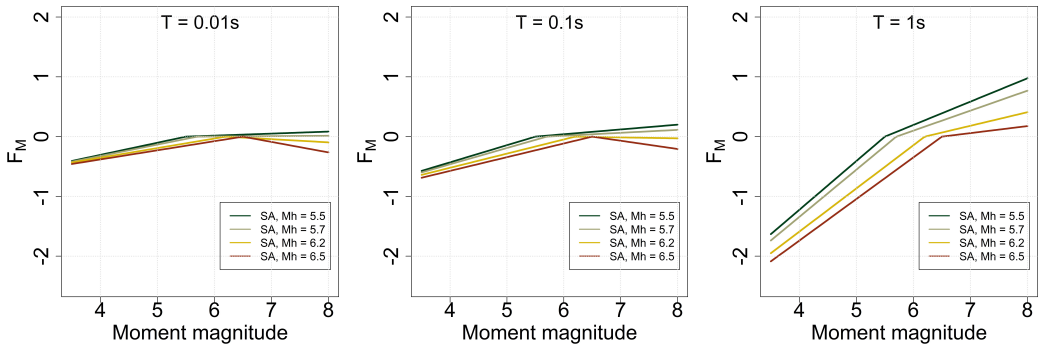


Figure 5.10: Variation of the source term with M_h in $\{5.5, 5.7, 6.2, 6.5\}$, at $T = \{0.01s, 0.1s, 1s\}$.

All things considered, this analysis does not identify an objective criterion for setting M_h. In our case, the low reduction in MSE and $\hat{\sigma}$ observed for $M_h = \{6.2, 6.5\}$ convince us to keep $M_h = 5.7$ as fixed value, as it is the mean of the values assumed by M_h in Lanzano, Luzi, Pacor, et al. (2019).
Nonetheless, to better stick to the formulation of ITA18, it may be preferable to introduce in the functional form a dependence of M_h on the period. Notice that such a change could be handled in two ways, and either way would be non-trivial in a functional framework. If M_h is assumed to be a known function, *e.g* resorting to the parameter proposed in Sabetta et al. (2021), then the overall model should be reformulated as a *function-on-function* regression. If M_h is considered unknown, then the regression would be non-linear in the coefficients. Both these options open interesting perspectives for further research.

A similar analysis conducted on M_{ref} and h does not show the same criticalities. Indeed, a variation of the parameters in the ranges identified by Lanzano, Luzi, Pacor, et al. (2019) does not correspond to any significant variation in the MSE error ($< 1\%$), in $\hat{\sigma}$ ($< 1\%$), nor in the *path* term in which these parameters appear (see Appendix B.1.2, B.1.3). The effect of the alteration of M_{ref} and h is found in negligible modifications in the estimates of c_2 and c_3, that are effectively compensated by the intercept term.
This suggests that our fixing $M_h = 5.7$, $M_{\text{ref}} = 4.5$ and $h = 5.9$ km may be the most convenient choice, as their variation has poor effect on the outcomes of the regression.

5.3 Comparison with Scalar ITA18

This section is devoted to the comparison between the results of ITA18 and its functional extension. Below, we refer to the first as *Scalar ITA18* and to the latter as *Functional ITA18*.

5.3.1 Comparison of the functional coefficients

Figure 5.11 shows the comparison between Scalar ITA18 and Functional ITA18 estimates of the coefficients. Each functional estimate is associated to 37 simultaneous confidence intervals[5] [6]. We observe that the general trend of Scalar ITA18 coefficients is always captured by the functional model, but that the roughness of the first is smoothed out at periods that correspond to the highest variability of data. We refer in particular to the peaks that ITA18 estimates present in correspondence of $T = 0.1$ s. Such smooth behaviour is the result of the regularization that we operate on the functional coefficients, through the introduction of the penalization term in the least squares criterion. Recall that Scalar ITA18 does not employ any form of regularization for its point-wise estimates.

[5]Following the procedure discussed in Section 3.4, the SCIs for a coefficient are constructed using the bootstrap sample generated from its empirical distribution. More precisely, the SCI at a sampling period T is given by the amplitude at T of the fence of the functional boxplots built for the bootstrap sample.

[6]It is worth noticing that, by introducing the assumption of Gaussianity of the point-wise residuals and by using the value of the point-wise variance of the coefficient estimate (Section 3.3.2), we obtain the one-at-a-time confidence intervals. Nonetheless, we are not interested in the information they bring, for two reasons: *(i)* the bootstrap approach prevents us from introducing any far-fetched hypothesis on the distribution of data, *(ii)* a simultaneous confidence is more informative for our scopes, since it allows to draw conclusions for the entire estimated curve and not only point-wise.

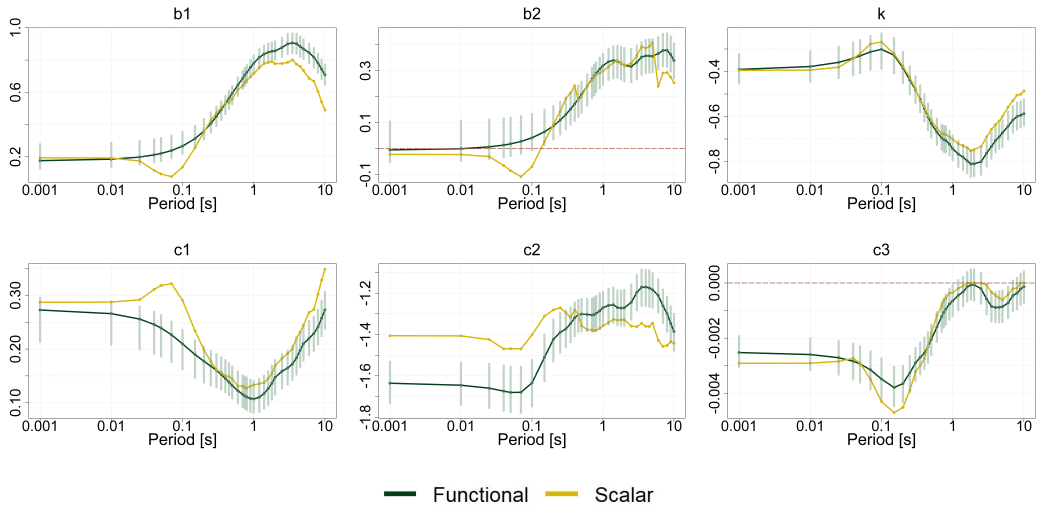


Figure 5.11: Comparison of the estimated coefficients between Scalar ITA18 and Functional ITA18. The green vertical segments indicate the simultaneous confidence intervals at the sampling points. The red dashed line marks zero.

5.3.2 Comparison of ground motion predictive performances

Figure 5.12 shows a comparison of the point-wise mean squared errors. From the comparison of the coefficient estimates, where we observe a clear-cut smoothing of the predictions at $T = 0.1\ s$, we would have expected to observe a strong reduction of the PMSE in the model functional at that time. In fact, the reduction is very small and the trend of the two statistics is similar along the period axis.

Figure 5.13 shows the comparison between the estimated point-wise standard deviations of Functional and Scalar ITA18. Notice that $\hat{\sigma}$ for the Functional model is obtained following the argument in Section 3.3.3. While Scalar ITA18 is associated to a lower $\hat{\sigma}$ at very short periods, Functional ITA18 provides a lower standard deviation at almost all periods $T > 0.5\ s$. Eventually, the two estimates align for $T > 8\ s$. However, we emphasize that the result of this comparison is not very informative, since the estimate we make of the *degrees-of-freedom* (dof) of the model has no sound theoretical basis.

Figure 5.14 shows the observations against the predictions of Functional and Scalar ITA18, for magnitudes $M_w = 4.0$ and $M_w = 6.8$. The two rows of plots are associated to normal faulting and strike-slip scenarios respectively, whereas the columns correspond to periods $T = 0.01\ s$ and $T = 1\ s$. The behaviour of Functional ITA18 closely follows that of its scalar counterpart at all periods; still, we notice one interesting deviation. Indeed, the functional model provides higher predictions of

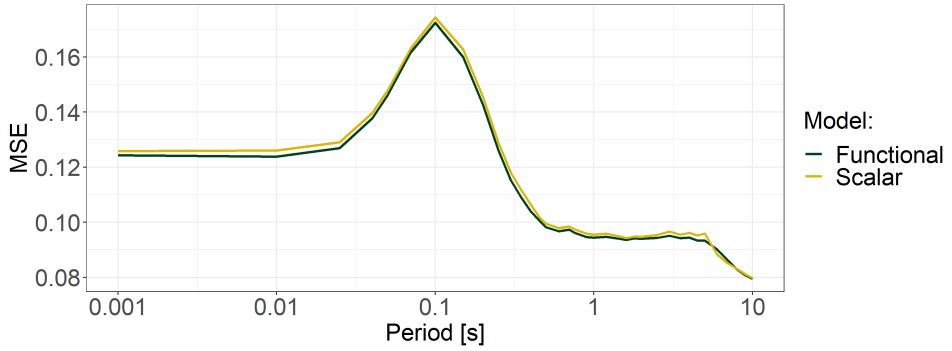


Figure 5.12: Point-wise Mean Squared Error for Functional ITA18 and Scalar ITA18.

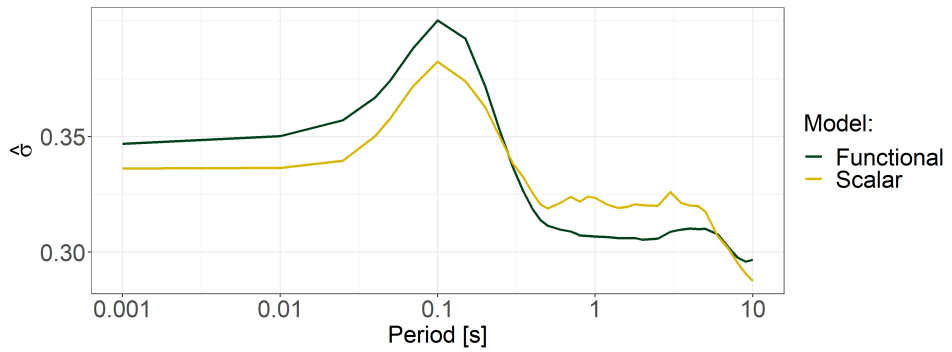


Figure 5.13: Estimated residual standard deviation $\hat{\sigma}$ for Functional ITA18 and Scalar ITA18. Notice that $\hat{\sigma}$ estimate of Functional ITA18 does not use the correct *dof* of the model, hence the comparison is not very reliable. We show the plot for the sake of completeness.

near-source ground motions for normal faulting and thrust faulting (see Appendix B.2) scenarios at short periods. Note that this behaviour is not present in strike-slip scenarios.

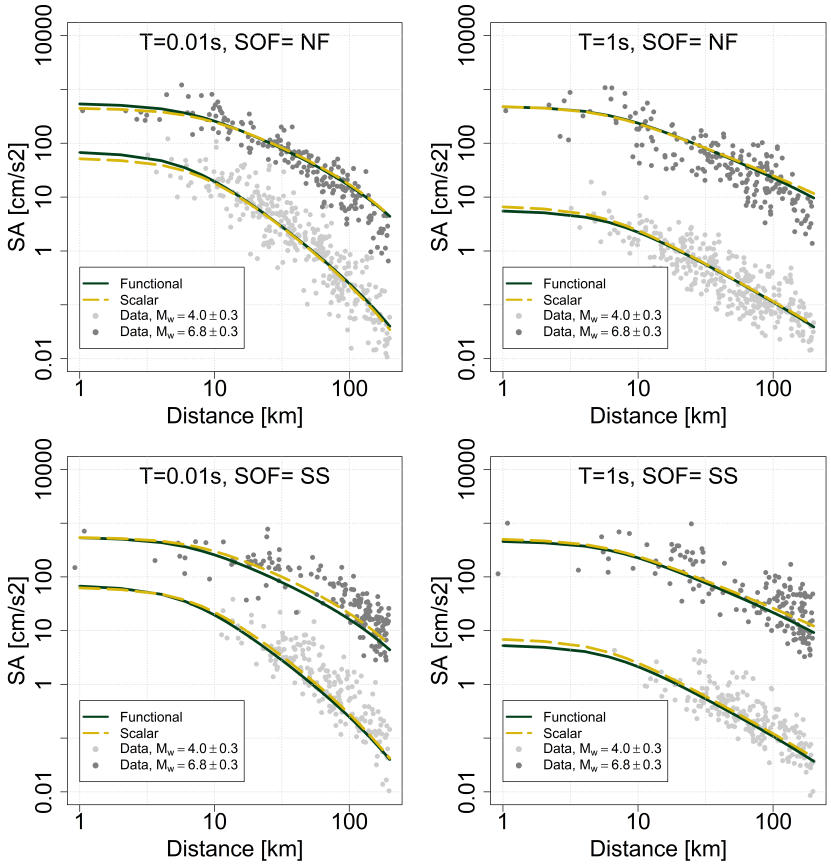


Figure 5.14: Comparison of the ground motion predicted by Functional ITA18 and Scalar ITA18 for normal faulting and strike-slip scenarios, at $T = \{0.01s, 1s\}$. Each plot shows the predictions for both $M_w = 4.0$ and $M_w = 6.0$. The dots are coloured light if $M_w = 4.0 \pm 0.3$, dark if $M_w = 6.0 \pm 0.3$.

5.3.3 Comparison at near-source scenarios

In probabilistic seismic hazard, large-magnitude events occurring near a site are considered critical scenarios that require pre-eminent investigation. Ground motion models are expected to perform well for such rare scenarios, *i.e.* to predict realistic values of soil motion with low prediction uncertainty (Kotha, Weatherill, et al., 2021). In light of this, it is somehow problematic that the ESM flatfile (Lanzano, Sgobba, Luzi, et al., 2018), which is the pan-European dataset model ITA18 was calibrated on, is composed by less than 9% of near-source events of magnitude $M_w \geq 5.5$, implying that the model is poorly constrained by data of this kind.

Only recently, Sgobba, Felicetta, et al. (2021) took advantage of the availability of a richer dataset of near-source records to evaluate corrective terms for the estimates of ITA18 in correspondence of such scenarios. The dataset is NESS2, which is an updated version of the worldwide NEar Source Strong motion flat file (Pacor, Felicetta, Lanzano, Sgobba, Puglia, D'Amico, Russo, Baltzopoulos, et al., 2018), containing an increased number of near-source recordings of moderate-to-strong earthquakes. The corrected model is obtained in Sgobba, Felicetta, et al. (2021) by logarithmically combining the ITA18 median predictions (SA_{ITA18}) with a correction factor δ_c, according to the equation

$$\log_{10}(SA_{\text{ITA18, corrected}}) = SA_{\text{ITA18}} + \delta_c.$$

In particular, given the little dependence of near-source predictions on V_{S30}, the corrective term δ_c is estimated conditionally on magnitude, distance and style-of-faulting. The authors observe that the corrected terms are associated with higher predictions of SA for distances less than 30 km. In this context, what we noted in the previous section is of particular interest and motivates the analysis reported here, that checks the near-source behaviour of Functional ITA18 with respect to Scalar ITA18 and its correction, hereafter referred to as *Corrected ITA18*.

First, we compare the PMSEs of Functional and Scalar ITA18, evaluated for data partitioned in three classes of distances, which are

- *class 1*: $d_{JB} \leq 10$km,
- *class 2*: $10\text{km} < d_{JB} \leq 30$km,
- *class 3*: $d_{JB} > 30$km.

The sequence in Figure 5.15 shows that for *classes 1* and *2* Functional ITA18 corresponds to a lower MSE, while little to no difference is observed for *class 3*. This suggests that the functional model is better able to fit the observations for near-source scenarios.

Figure 5.16 provides a zoomed in picture of Figure 5.14, showing the scaling of ground motion with distance in the range [0, 30] km as predicted by Functional and Scalar ITA18 respectively. The images show the scaling at three different magnitudes, *i.e.* $M_w = \{4.8, 5.8, 7\}$, and each plot is repeated for the sequence of increasing vibration periods $T = \{0.01s, 0.1s, 1s\}$. The plots correspond to a normal-faulting scenario [7]. Interestingly, distinctions in the prediction curves are observed at periods

[7] The analysis has been repeated for thrust-faulting scenarios and for strike-slip scenarios (see Appendix B.3). The first shows results very similar to the NF case. The latter confirms what

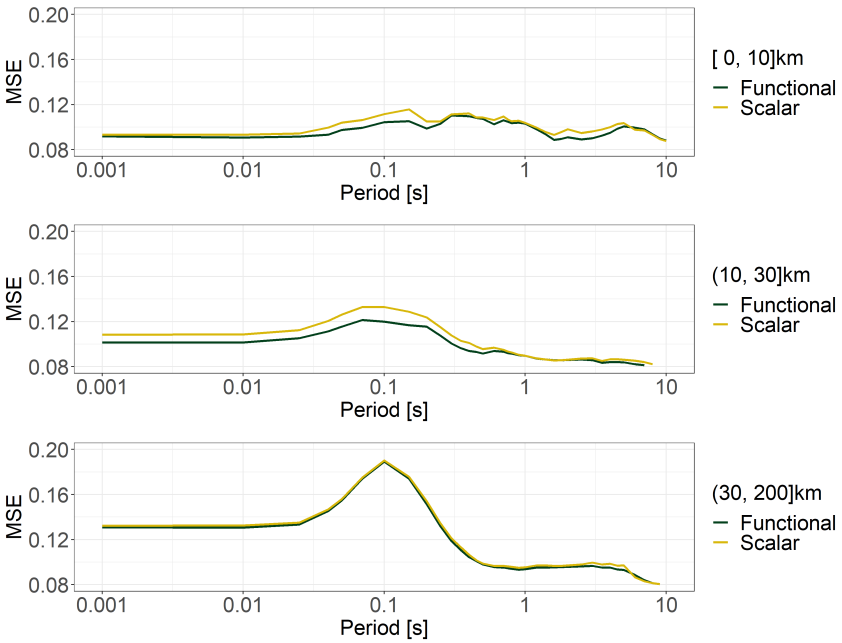


Figure 5.15: Comparison of MSE of Functional ITA18 and Scalar ITA18 for three classes of distance. Top to bottom: $d_{JB} \leq 10$ km, 10 km $< d_{JB} \leq 30$ km and $d_{JB} \geq 30$ km.

$T < 1\ s$ for critical hazard scenarios, *i.e.* for estimates associated to $M_w = \{5.8, 7\}$, whereas the two models present similar trends for $M_w = 4.8$. Generally, the functional model overestimates ground motion with respect to Scalar ITA18 in these scenarios.

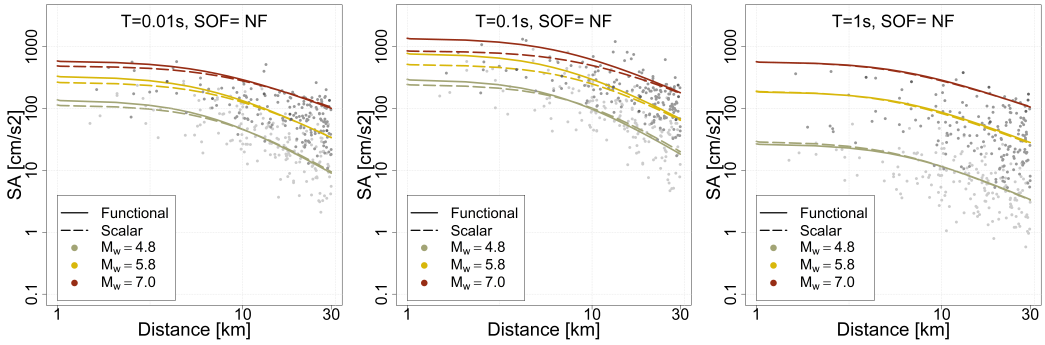


Figure 5.16: Comparison of the ground motion predicted by Functional ITA18 and Scalar ITA18 for normal faulting near-source scenarios, at $T = \{0.01s, 0.1s, 1s\}$. The continuous line is SA predicted by Functional ITA18. The dashed line is SA predicted by Scalar ITA18. In each plot, the lines are coloured according to the magnitude of the corresponding scenario. The dots are the observed values of SA, coloured darker as the magnitude of the corresponding event increases.

previously detected in Figure 5.14, namely that no difference is observed between the near-source trends of Scalar and Functional ITA18.

The fact that Functional ITA18 is associated to higher ground motion predictions than its scalar counterpart in normal and thrust faulting scenarios motivates its comparison with the near-source corrected predictions of the scalar model, at these faulting regimes. Similarly to the previous figure, Figure 5.17 shows the near-source scaling of spectral acceleration with distance as predicted by Functional and Corrected ITA18 [8]. For periods around $T = 0.01\ s$, the functional model aligns to the corrected scalar estimates, while at $T = 0.1\ s$ it predicts higher soil motions at very short distances, and then approaches the corrected predictions at $d_{JB} \simeq 30$ km. We notice that the differences among predictions tend to increase with the magnitude, and eventually to disappear for all magnitudes at periods $T \geq 1\ s$.

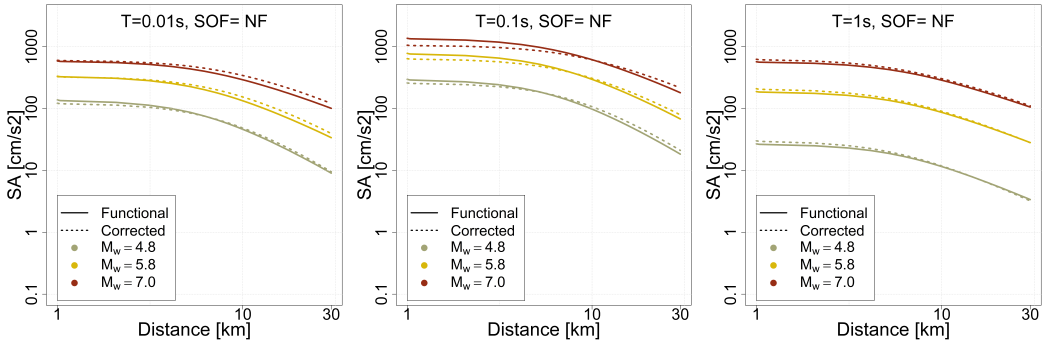


Figure 5.17: Comparison of the ground motion predicted by Functional ITA18 and Corrected ITA18 for normal faulting near-source scenarios, at $T = \{0.01s, 0.1s, 1s\}$. The continuous line is SA predicted by Functional ITA18. The dashed line is SA predicted by Corrected ITA18. In each plot, the lines are coloured according to the magnitude of the corresponding scenario.

Although the functional model is fitted on the same dataset as Scalar ITA18, the functional estimates seem to better explain soil motion for critical hazard scenarios, especially at very short periods, by predicting higher values of spectral acceleration. Given their relevance in the context of probabilistic seismic hazard, it may be of interest to deepen this qualitative investigation on critical hazard scenarios, by calibrating the functional model on the NESS2 dataset and checking how the coefficients estimates are affected by new observations.

[8] Here we report the analysis for the normal faulting scenario. The analysis is repeated for thrust faulting scenario, showing very similar results to what observed for normal faulting.

Conclusions

The present work proposes a new approach to the analysis of partially observed functional data, maintaining a thorough focus on the application aspect that motivated the analysis itself, aimed at the formulation of Ground Motion Prediction Equations in the field of engineering seismology. Based on the proposed methodology, the reconstruction of an incomplete observation is associated with the definition of a weight, quantifying the confidence associated with the values that the reconstructed functional datum assumes in the different parts of the domain. Accordingly, the classical *penalized smoothing* and *penalized functional regression* are extended to account for the introduction of weights. The soundness of the proposed techniques with respect to the adopted reconstruction method is tested via a cross-validation procedure, that confirms the validity of the approach and its adaptability to other application contexts, characterized by incomplete functional data that are extended to the unobserved part of the domain. The benefits of such novel approach are twofold. First, a functional embedding enables us to capture the smoothness underlying the longitudinal discrete observations. Second, the reconstruction of incomplete data avoids the loss of information that the removal of missing values would involve, by extending the definition of the functional data over a domain that is wider than the one common to all data pre-reconstruction.

Concerning the application aspects of this analysis, the functional extension of the ground motion model ITA18 (Lanzano, Luzi, Pacor, et al., 2019) for the prediction of the median spectral amplitudes is calibrated on the ESM dataset (Lanzano, Sgobba, Luzi, et al., 2018) and shows results that are satisfactory. Indeed, the diagnostic conducted on the model does not report any anomaly, whereas the coefficient estimates and prediction results show trends that are seismologically interpretable, standing as useful tool in the understanding of the phenomenon under study. The simplifying assumption made in the definition of the functional form, *i.e.* that parameters M_h, M_{ref} and h are *a priori* fixed and independent of the period, is checked through a sensitivity analysis and does not entail any abnormal trend in estimation and predictive performances of the model. A comparison with ITA18 highlights an

interesting behaviour of the functional model for critical hazard scenarios, where the latter does better in predicting high values of spectral acceleration.

In conclusion, the proposed approach applied to ground motion prediction equations allows one to predict the whole spectrum of ground motion, simultaneously on a continuous range of periods, with a sustainable computational burden. Through the construction of bootstrap simultaneous confidence bands, conclusions can be drawn simultaneously on the functional estimate, rather than only point-wise. Finally, the methodology operates a natural smoothing of the predictions, solving the shortcomings of univariate and multivariate approaches that show non-physical discontinuous patterns in the estimates and predictions.

New horizons of research open up in multiple directions. One consists in the recalibration of the model on the NEar Source Strong motion (NESS, Sgobba, Felicetta, et al., 2021) dataset, that would allow for further investigation on the behaviour of the functional model for critical hazard scenarios. An interesting extension of the model could introduce in the functional form a dependence of M_h on the period. Such a change could be handled in two ways, and either way would be non-trivial in a functional framework. If M_h is assumed to be a known function, then the overall model should be reformulated as a *function-on-function* regression. If M_h is considered unknown, then the regression would be non-linear in the coefficients. Another direction of research may go towards a functional extension of the model of Lanzano, Sgobba, Caramenti, et al. (2021), that, by introducing a site and event dependence in the regression coefficients, allows for a regionalization of the estimates, and reduces the residual standard deviation of the model. Finally, the most consistent extension of this work combines the functional estimation for the median with the geostatistical model for the residuals proposed in Menafoglio et al. (2020), providing a fully functional, effective tool for the construction of period-continuous seismic shaking maps.

Bibliography

Anderson, J. G. and J. N. Brune (1999). "Probabilistic Seismic Hazard Analysis without the Ergodic Assumption". In: *Seismological Research Letters* 70.1, pp. 19–28.

Anderson, J. G. and Y. Uchiyama (2011). "A Methodology to Improve Ground-Motion Prediction Equations by Including Path Corrections". In: *Bulletin of the Seismological Society of America* 101.4, pp. 1822–1846.

Al-Atik, L. et al. (2010). "The Variability of Ground-Motion Prediction Models and Its Components". In: *Seismological Research Letters* 81.5, pp. 794–801.

Bindi, D. et al. (2011). "Ground motion prediction equations derived from the Italian strong motion database". In: *Bulletin of Earthquake Engineering* 9.6, pp. 1899–1920.

Bommer, J. J., J. Douglas, and F. O. Strasser (2003). "Style-of-Faulting in Ground-Motion Prediction Equations". In: *Bulletin of Earthquake Engineering* 108.1, pp. 171–203.

Boor, C. de (2001). "A Practical Guide to Splines, Revised Edition". In: Springer Nature.

Boore, D. M. (2010). "Orientation-independent, nongeometric-mean measures of seismic intensity from two horizontal components of ground motion". In: *Bulletin of the Seismological Society of America* 100.4, pp. 1830–1835.

Boore, D. M. et al. (2014). "NGA-West2 Equations for Predicting PGA, PGV, and 5% Damped PSA for Shallow Crustal Eartquakes". In: *Earthquake Spectra* 30.3, pp. 1057–1085.

Bosq, D. (1998). "Linear Processes in Function Spaces". In: Springer. Chap. 1.

Campbell, K. W. (1981). "Near-source attenuation of peak horizontal acceleration". In: *Bulletin of the Seismological Society of America* 71.6, pp. 2039–2070.

Cao, G., L. Yang, and D. Todem (2012). "Simultaneous Inference For The Mean Function Based on Dense Functional Data". In: *Journal of Nonparametric Statistics* 24.2, pp. 359–377.

Chang, C., X. Lin, and R. T. Ogden (2017). "Simultaneous confidence bands for functional regression models". In: *Journal of Statistical Planning and Inference* 188, pp. 67–81.

Chiou, B. S.-J. and R. R. Youngs (2014). "Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra". In: *Earthquake Spectra* 30, pp. 1117–1153.

Cuevas, A., M. Febrero, and R. Fraiman (2004). "An anova test for functional data". In: *Computational Statistics and Data Analysis*.

Cuevas, A. and R. Fraiman (2004). "On the Bootstrap Methodology for Functional Data". In: *Compstat – Proceedings in Computational Statistics*.

Degras, D. A. (2011). "Simultaneous confidence bands for functional regression models". In: *Statistica Sinica* 21.4, pp. 1735–1765.

— (2017). "Simultaneous confidence bands for the mean of functional data". In: *WIREs Computational Statistics* 9.3, e1397.

Douglas, J. (2003). "Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates". In: *Earth-Science Reviews* 61.1, pp. 43–104.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26.

Eurocode8 (2003). "Eurocode 8: Design Provisions for Earthquake Resistance of Structures, Part 1.1: General rules, seismic actions and rules for buildings." In: *CEN, European Committee for Standardisation TC250/SC8, PrEN1998-1.*

Ferraty, F. and P. Vieu (2006). "Nonparametric Functional Data Analysis". In: Springer-Verlag New York.

Fukushima, Y. (1996). "Scaling relations for strong ground motion prediction models with M^2 terms". In: *Bulletin of the Seismological Society of America* 86.2, pp. 329–336.

Goldsmith, J., S. Greven, and C. Crainiceanu (2013). "Corrected Confidence Bands for Functional Data Using Principal Components". In: *Biometrics* 69.1, pp. 41–51.

Horváth, L. and P. Kokoszka (2012). "Inference for Functional Data with Applications". In: Springer. Chap. 2.

Huang, C. and C. Galasso (2019). "Ground-motion intensity measure correlations observed in italian strong-motion records". In: *Earthquake Engineering Structural Dynamics* 48.15, pp. 1634–1660.

James, G. et al. (2013). "An Introduction to Statistical Learning : with Applications in R". In: Springer.

Johnson, R. A. and D. W. Wichern (2018). "Applied Multivariate Statistical Analysis". In: Pearson Education.

Kneip, A. and D. Liebl (2020). "On the optimal reconstruction of partially observed functional data". In: *The Annals of Statistics* 48.3, pp. 1692–1717.

Kotha, S. R., D. Bindi, and F. Cotton (2014). "Pan-European ground-motion prediction eqautions for the average horizontal component of PGA, PGV and 5 %-damped PSA at spectral periods up to 3.0s using the RESORCE dataset". In: *Bulletin of Earthquake Engineering* 12, pp. 391–430.

— (2016). "Partially non-ergodic region specific GMPE for Europe and Middle-East". In: *Bulletin of Earthquake Engineering* 14, pp. 1245–1263.

Kotha, S. R., G. Weatherill, et al. (2020). "A regionally-adaptable ground-motion model for shallow crustal earthquakes in Europe". In: *Bulletin of Earthquake Engineering* 18, pp. 4091–4125.

— (2021). "Near-Source Magnitude Scaling of Spectral Accelerations: Analysis and Update of Kotha et al. (2020) Model". In: *Bulletin of Earthquake Engineering*.

Kraus, D. (2015). "Components and completion of partially observed functional data". In: *Journal of the Royal Statistical Society* 77.4, pp. 777–801.

Lanzano, G., L. Luzi, F. Pacor, et al. (2019). "A Revised Ground-Motion Prediction Model for Shallow Crustal Earthquakes in Italy". In: *Bulletin of the Seismological Society of America* 109.2, pp. 525–540.

Lanzano, G., L. Luzi, E. Russo, et al. (2018). *Engineering Strong Motion Database (ESM) flatfile*. en. DOI: 10.13127/ESM/FLATFILE.1.0.

Lanzano, G., S. Sgobba, L. Caramenti, et al. (2021). "Ground-Motion Model for Crustal Events in Italy by Applying the Multisource Geographically Weighted Regression (MS-GWR) Method". In: *Bulletin of the Seismological Society of America*. ISSN: 0037-1106.

Lanzano, G., S. Sgobba, L. Luzi, et al. (2018). "The pan-European Engineering Strong Motion (ESM) flatfile: compilation criteria and data statistics". In: *Bulletin of Earthquake Engineering* 17, pp. 561–582.

Liebl, D. and M. Reimherr (2019). "Fast and Fair Simultaneous Confidence Bands for Functional Parameters". In:

Lin, P.-S. et al. (2011). "Repeatable source, site, and path effects on the standard deviation for empirical ground-motion prediction models". In: *Bulletin of the Seismological Society of America* 101.5, pp. 2281–2295.

Luzi, L., F. Pacor, and R. Puglia (2017). *Italian Accelerometric Archive v 2.3*. en. DOI: 10.13127/ITACA.2.3.

Mehrotra, S. and A. Maity (2019). "Simulataneous Variable Selection, Clustering, and Smoothing in Function on Scalar Regression". In:

Menafoglio, A. et al. (2020). "Simulation of seismic ground motion fields via object-oriented spatial statistics with an application in Northern Italy". In: *Stochastic Environmental Research and Risk Assessment* 34, pp. 1607–1627.

Pacor, F., C. Felicetta, G. Lanzano, S. Sgobba, R. Puglia, M. D'Amico, E. Russo, G. Baltzopoulos, et al. (2018). "NESS1: A Worldwide Collection of Strong-Motion Data to Investigate Near-Source Effects". In: *Seismological Research Letters* 89, pp. 2299–2313.

Pacor, F., C. Felicetta, G. Lanzano, S. Sgobba, R. Puglia, M. D'Amico, E. Russo, and L. Luzi (2018). *NEar-Source Strong-motion flatfile (NESS), version 1.0*. en. DOI: 10.13127/NESS.1.0. URL: http://ness.mi.ingv.it/.

Palano, M. (Dec. 2014). "On the present-day crustal stress, strain-rate fields and mantle anisotropy pattern of Italy". In: *Geophysical Journal International* 200.

Politis, D. N. and J. P. Romano (1994). "Limit theorems for weakly dependent Hilbert space valued random variables with application to the stationary bootstrap". In: *Statistica Sinica* 4, pp. 461–476.

Puglia, R. et al. (2018). "Strong-motion processing service: a tool to access and analyse earthquakes strong-motion waveforms". In: *Bulletin of Earthquake Engineering* 16, pp. 2641–2651.

Ramsay, J. O. and B. W. Silverman (2005). "Functional Data Analysis". In: Springer.

Římalová, V. et al. (2020). "A permutation approach to the analysis of spatiotemporal geochemical data in the presence of heteroscedasticity". In: *Environmetrics* 31.4, e2611.

Sabetta, F. et al. (2021). "Simulation of non-stationary stochastic ground motions based on recent Italian earthquakes". In: *Bulletin of Earthquake Engineering* 19.9, pp. 3287–3315.

Sgobba, S., C. Felicetta, et al. (2021). "NESS2.0: An Updated Version of the Worldwide Dataset for Calibrating and Adjusting Ground-Motion Models in Near Source". In: *Bulletin of the Seismological Society of America*. ISSN: 0037-1106.

Sgobba, S., G. Lanzano, et al. (2019). "Spatial Correlation Model of Systematic Site and Path Effects for Ground-Motion Fields in Northern Italy". In: *Bulletin of the Seismological Society of America* 109.4, pp. 1419–1434.

Stafford, P. J. (2014). "Crossed and Nested Mixed-Effects Approaches for Enhanced Model Development and Removal of the Ergodic Assumption in Empirical Ground-Motion Models". In: *Bulletin of the Seismological Society of America* 104.2, pp. 702–719.

Worden, C. B. et al. (2018). "Spatial and Spectral Interpolation of Ground-Motion Intensity Measure Observations". In: *Bulletin of the Seismological Society of America* 108.2, pp. 866–875.

Yao, F., H.-G. Müller, and J.-L Wang (2005). "Functional Data Analysis for Sparse Longitudinal Data". In: *Journal of the American Statistical Association* 100.470, pp. 577–590.

Appendix A

Theoretical results

A.1 Weighted Functional Penalized Least Squares

In the following, t as integration parameter is omitted for clarity of notation.
Let $W(t) = \mathrm{diag}(\boldsymbol{w}(t))$.

$$
\begin{aligned}
WFLS &= \int (\mathrm{C}\boldsymbol{\phi} - X\mathrm{B}\boldsymbol{\theta})^\intercal W(\mathrm{C}\boldsymbol{\phi} - X\mathrm{B}\boldsymbol{\theta}) \\
&= \int (\mathrm{C}\boldsymbol{\phi})^\intercal W(\mathrm{C}\boldsymbol{\phi}) + \int (X\mathrm{B}\boldsymbol{\theta})^\intercal W(X\mathrm{B}\boldsymbol{\theta}) - \int (X\mathrm{B}\boldsymbol{\theta})^\intercal W(\mathrm{C}\boldsymbol{\phi}) - \int (\mathrm{C}\boldsymbol{\phi})^\intercal W(X\mathrm{B}\boldsymbol{\theta}) \\
&= \int \mathrm{tr}\left[(W\mathrm{C}\boldsymbol{\phi})(\mathrm{C}\boldsymbol{\phi})^\intercal\right] + \int \mathrm{tr}\left[(WX\mathrm{B}\boldsymbol{\theta})(X\mathrm{B}\boldsymbol{\theta})^\intercal\right] - \int \mathrm{tr}\left[(W\mathrm{C}\boldsymbol{\phi})(X\mathrm{B}\boldsymbol{\theta})^\intercal\right] - \int \mathrm{tr}\left[(X\mathrm{B}\boldsymbol{\theta})(W\mathrm{C}\boldsymbol{\phi})^\intercal\right] \\
&= \int \mathrm{tr}\left[W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\mathrm{C}^\intercal\right] + \int \mathrm{tr}\left[WX\mathrm{B}\boldsymbol{\theta}\boldsymbol{\theta}^\intercal\mathrm{B}^\intercal X^\intercal\right] - \int \mathrm{tr}\left[(X\mathrm{B}\boldsymbol{\theta})^\intercal(W\mathrm{C}\boldsymbol{\phi})\right] - \int \mathrm{tr}\left[(X\mathrm{B}\boldsymbol{\theta})(W\mathrm{C}\boldsymbol{\phi})^\intercal\right] \\
&= \int \mathrm{tr}\left[\mathrm{C}^\intercal W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \int \mathrm{tr}\left[X^\intercal WX\mathrm{B}\boldsymbol{\theta}\boldsymbol{\theta}^\intercal\mathrm{B}^\intercal\right] - 2\int \mathrm{tr}\left[\mathrm{B}\boldsymbol{\theta}\boldsymbol{\phi}^\intercal\mathrm{C}^\intercal WX\right] \\
&= \int \mathrm{tr}\left[\mathrm{C}^\intercal W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \int \mathrm{tr}\left[\mathrm{B}^\intercal X^\intercal WX\mathrm{B}\boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right] - 2\int \mathrm{tr}\left[\mathrm{B}^\intercal(\boldsymbol{\theta}\boldsymbol{\phi}^\intercal\mathrm{C}^\intercal WX)^\intercal\right].
\end{aligned}
$$

Here, the operations of integration and summation, implied by the trace, may be interchanged, and hence the previous can be reformulated as

$$
\begin{aligned}
WFLS &= \int \mathrm{tr}\left[\mathrm{C}^\intercal W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \int \mathrm{tr}\left[\mathrm{B}^\intercal X^\intercal WX\mathrm{B}\boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right] - 2\int \mathrm{tr}\left[\mathrm{B}^\intercal(\boldsymbol{\theta}\boldsymbol{\phi}^\intercal\mathrm{C}^\intercal WX)^\intercal\right] \\
&= \mathrm{tr}\left[\int \mathrm{C}^\intercal W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \mathrm{tr}\left[\int \mathrm{B}^\intercal X^\intercal WX\mathrm{B}\boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right] - 2\mathrm{tr}\left[\int \mathrm{B}^\intercal X^\intercal W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\theta}^\intercal\right] \\
&= \mathrm{tr}\left[\int \mathrm{C}^\intercal W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\phi}^\intercal\right] + \mathrm{tr}\left[\int \mathrm{B}^\intercal X^\intercal WX\mathrm{B}\boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right] - 2\mathrm{tr}\left[\mathrm{B}^\intercal \int X^\intercal W\mathrm{C}\boldsymbol{\phi}\boldsymbol{\theta}^\intercal\right].
\end{aligned}
$$

In order to minimize this quantity, we have to take its derivative with respect to

B. The first term does not depend on B and hence it disappears. The derivative of the third is equal to

$$-2 \int X^\intercal W \mathrm{C} \boldsymbol{\phi} \boldsymbol{\theta}^\intercal$$

and is easily obtained by recalling that the derivative of $\mathrm{tr}(\mathrm{B}^\intercal \mathrm{A})$ with respect to B is A. The derivation of the term in the middle requires a little more work.
First, recall that

$$\nabla_A \mathrm{tr}(ABA^\intercal C) = CAB + C^\intercal AB^\intercal$$

Then, the derivative of the middle term is obtained through the following calculations

$$\begin{aligned}
\nabla_B \mathrm{tr}\left[\int \mathrm{B}^\intercal X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right] &= \int \nabla_B \mathrm{tr}\left[\mathrm{B}^\intercal X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right] \\
&= \int \nabla_B \mathrm{tr}\left[\mathrm{B}\boldsymbol{\theta}\boldsymbol{\theta}^\intercal \mathrm{B}^\intercal X^\intercal W X\right] \\
&= \int \left(X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal + X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right) \\
&= 2 \int X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal
\end{aligned}$$

Similarly for the penalization term, we observe that

$$\begin{aligned}
&\int \left[L\mathrm{B}\boldsymbol{\theta}\right]^\intercal \left[L\mathrm{B}\boldsymbol{\theta}\right] \\
&= \int \mathrm{tr}\left[(\mathrm{B}L\boldsymbol{\theta})(\mathrm{B}L\boldsymbol{\theta})^\intercal\right] \\
&= \int \mathrm{tr}\left[\mathrm{B}(L\boldsymbol{\theta})(L\boldsymbol{\theta})^\intercal \mathrm{B}^\intercal\right] \\
&= \mathrm{tr}\left[\mathrm{B}R\mathrm{B}^\intercal\right],
\end{aligned}$$

where we define $[R]_{ij} = \langle L\theta_i, L\theta_j \rangle_{L^2(\mathcal{T})}$.

Taking advantage again of the properties of the derivative of the trace, we observe that

$$\begin{aligned}
\nabla_B \mathrm{tr}\left[\mathrm{B}R\mathrm{B}^\intercal\right] &= \nabla_B \mathrm{tr}\left[\mathrm{B}R\mathrm{B}^\intercal I\right] \\
&= I\mathrm{B}R + I\mathrm{B}R^\intercal \\
&= 2\mathrm{B}R.
\end{aligned}$$

Therefore, we find that B satisfies the following system

$$\int X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal + \lambda \mathrm{B} R = \int X^\intercal W \mathrm{C} \boldsymbol{\phi}\boldsymbol{\theta}^\intercal.$$

Taking the vec() on both sides of the equation and exploiting the linearity of the vec() operator, we get

$$\begin{aligned}
\mathrm{vec}\left(\int X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right) + \lambda \mathrm{vec}(\mathrm{B}R) &= \mathrm{vec}\left(\int X^\intercal W \mathrm{C} \boldsymbol{\phi}\boldsymbol{\theta}^\intercal\right) \\
\int \mathrm{vec}\left(X^\intercal W X \mathrm{B} \boldsymbol{\theta}\boldsymbol{\theta}^\intercal\right) + \lambda \mathrm{vec}(I\mathrm{B}R) &= \mathrm{vec}\left(\int X^\intercal W \mathrm{C} \boldsymbol{\phi}\boldsymbol{\theta}^\intercal\right) \\
\int \left(\boldsymbol{\theta}\boldsymbol{\theta}^\intercal \otimes X^\intercal W X\right) \mathrm{vec}(\mathrm{B}) + (\lambda R \otimes I)\mathrm{vec}(\mathrm{B}) &= \mathrm{vec}\left(\int X^\intercal W \mathrm{C} \boldsymbol{\phi}\boldsymbol{\theta}^\intercal\right) \\
\left[\int \left(\boldsymbol{\theta}\boldsymbol{\theta}^\intercal \otimes X^\intercal W X\right) + (\lambda R \otimes I)\right] \mathrm{vec}(\mathrm{B}) &= \mathrm{vec}\left(\int X^\intercal W \mathrm{C} \boldsymbol{\phi}\boldsymbol{\theta}^\intercal\right).
\end{aligned}$$

A.2 Estimation of the *degrees-of-freedom*

Without loss of generality, fix t_1 as time point. The *residual sum of squares* in t_1 is

$$\mathrm{RSS}(t_1) = \sum_{i=1}^{n} (\hat{\epsilon}_i(t_1))^2 = \mathrm{tr}\left(\hat{\boldsymbol{\epsilon}}(t_1)\hat{\boldsymbol{\epsilon}}(t_1)^\intercal\right).$$

Define A to be a $(n \times nT)$-dimensional block matrix in the form

$$\left[\; I_n \;\middle|\; O \;\middle|\; O \;\middle|\; \ldots \;\middle|\; O \;\right].$$

Then we may write $\hat{\boldsymbol{\epsilon}}(t_1) = A\mathrm{vec}(\hat{\mathcal{E}})$, so that

$$\mathrm{RSS}(t_1) = \mathrm{tr}\left((A\mathrm{vec}(\hat{\mathcal{E}}))(A\mathrm{vec}(\hat{\mathcal{E}}))^\intercal\right).$$

Taking the expectation of $\mathrm{RSS}(t_1)$, we get

$$\begin{aligned}
\mathbb{E}\left[RSS(t_1)\right] &= \mathbb{E}\left[\mathrm{tr}\left(A\mathrm{vec}(\hat{\mathcal{E}})\mathrm{vec}(\hat{\mathcal{E}})^\intercal A^\intercal\right)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(A(I-H)\mathrm{vec}(Y)\mathrm{vec}(Y)^\intercal(I-H)^\intercal A^\intercal\right)\right] \\
&= \mathbb{E}\left[\mathrm{tr}\left(A(I-H)\mathrm{vec}(\mathcal{E})\mathrm{vec}(\mathcal{E})^\intercal(I-H)^\intercal A^\intercal\right)\right] \\
&= \mathrm{tr}\left[\mathrm{Cov}(\mathrm{vec}(\mathcal{E}))(I-H)^\intercal A^\intercal A(I-H)\right].
\end{aligned}$$

Notice that we cannot isolate the covariance term from the corrective term, unless we assume that $\text{Cov}(\text{vec}(\mathcal{E})) = \sigma I_{nT}$. Such hypothesis is highly improbable: by implying that the time points are uncorrelated and that the variance of the error is equal at all time points, the hypothesis nullifies the reasons behind the use of a functional approach.
Nonetheless, we observe that if $\text{Cov}(\text{vec}(\mathcal{E})) = \sigma^2 I_{nT}$, we are able to identify a corrective term for $\text{RSS}(t_1)$. Indeed, we may write

$$\mathbb{E}\left[RSS(t_1)\right] = \text{tr}\left(\text{Cov}(\text{vec}(\mathcal{E}))(I-H)^\intercal A^\intercal A(I-H)\right) = \sigma \text{tr}\left((I-H)^\intercal A^\intercal A(I-H)\right),$$

and identify

$$\delta_1 = \text{tr}\left[(I-H)^\intercal A^\intercal A(I-H)\right] = n - 2\text{tr}\left[H_{1:n,1:n}\right] + \text{tr}\left[H_{1:n,\cdot}(H_{1:n,\cdot})^\intercal\right]$$

as an estimate of the *effective degrees-of-freedom* of the penalized regression at t_1.

Appendix B

Additional figures

B.1 Sensitivity analysis

B.1.1 Oversaturation check for M_h parameter

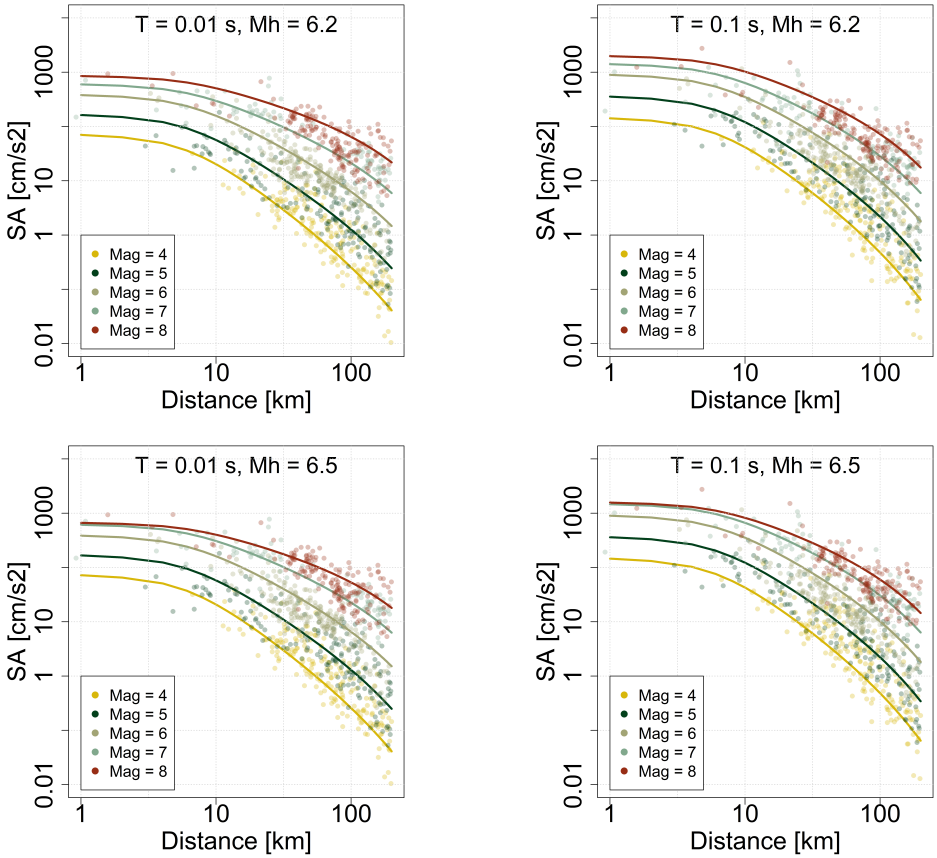


Figure B.1: Check of oversaturation for models corresponding to $M_h = \{6.2, 6.7\}$, at periods $T = \{0.01s, 0.1s\}$. The lines correspond to the predicted values of spectral acceleration, coloured according to the magnitude of the scenario which they refer to. The dots represent the observed values of spectral acceleration, and are coloured according to the magnitude of the recorded event, *e.g.* yellow if $M_w = 4 \pm 0.3$.

B.1.2 Results of sensitivity analysis for $M_{\mathbf{ref}}$

Notice that the plots below show lines that are almost perfectly overlapped and that cannot be distinguished.

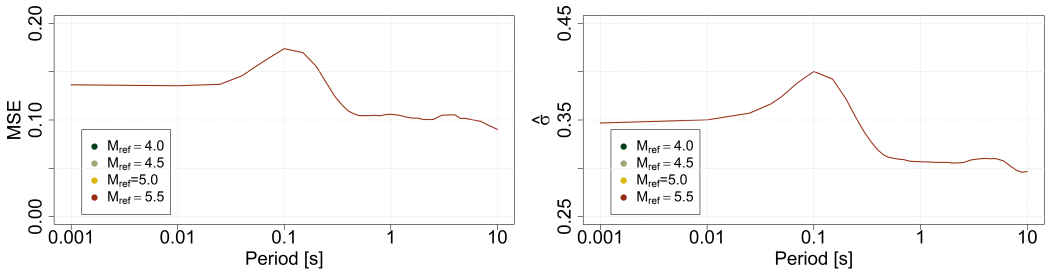


Figure B.2: MSE (left) and $\hat{\sigma}$ (right) for $M_{\text{ref}} = \{4.0, 4.5, 5.0, 5.5\}$. We notice perfect overlapping of the curves.

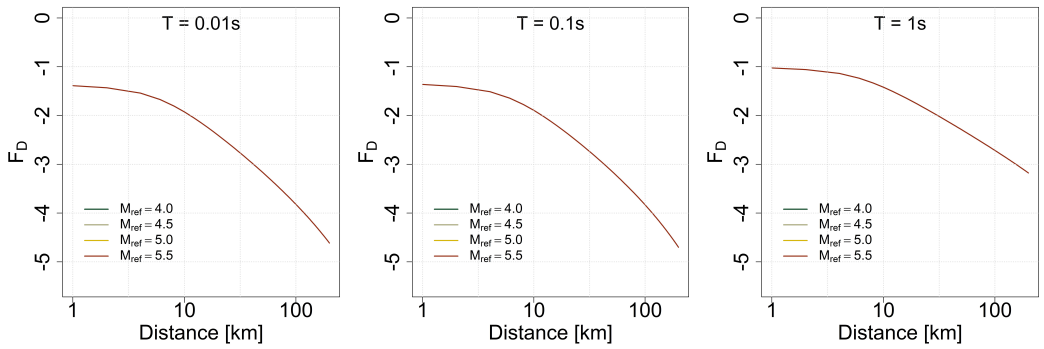


Figure B.3: Variation of the source term with M_{ref}, at $T = \{0.01s, 0.1s, 1s\}$. We notice perfect overlapping of the curves.

B.1.3 Results of sensitivity analysis for h

Notice that the plots below show lines that are almost perfectly overlapped and that cannot be distinguished.

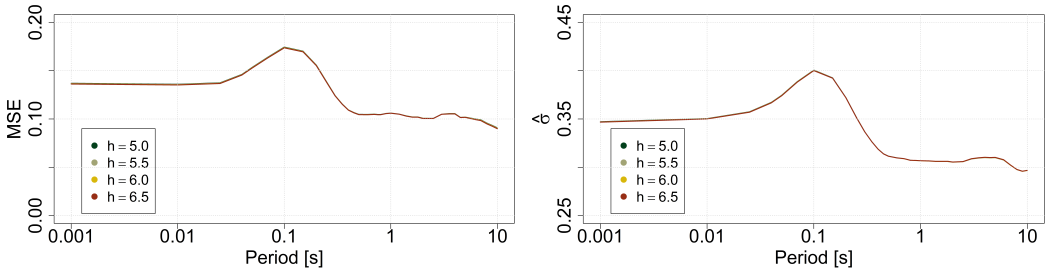


Figure B.4: MSE (left) and $\hat{\sigma}$ (right) for $h = \{5.0, 5.5, 6.0, 6.5\}$ km.

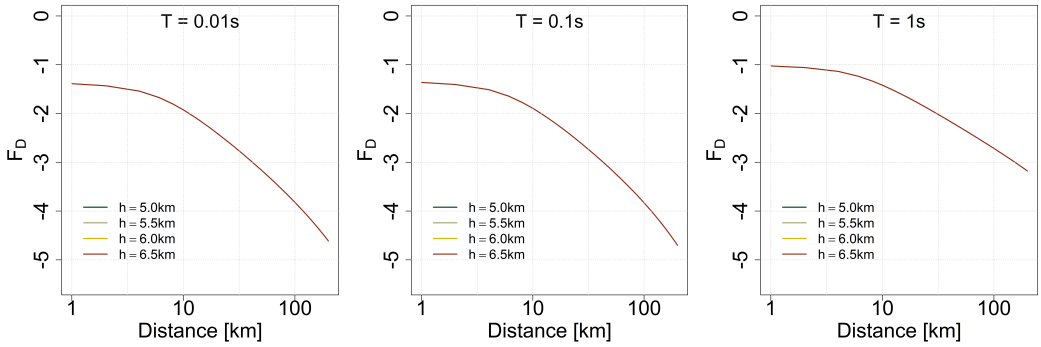


Figure B.5: Variation of the source term with h, at $T = \{0.01s, 0.1s, 1s\}$.

B.2 Comparison of ground motion predictive performances

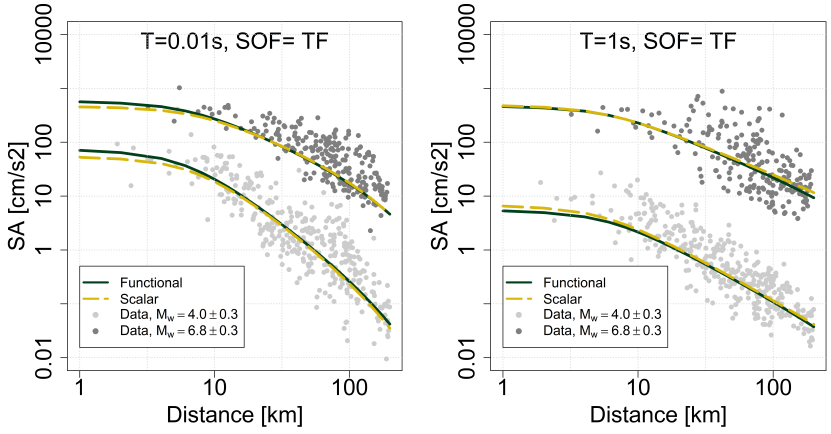


Figure B.6: Comparison of the ground motion predicted by Functional ITA18 and Scalar ITA18 for thrust faulting scenarios, at $T = \{0.01s, 1s\}$. Each plot shows the predictions for both $M_w = 4.0$ and $M_w = 6.0$. The dots are coloured light if $M_w = 4.0 \pm 0.3$, dark if $M_w = 6.0 \pm 0.3$.

B.3 Comparison at near-source SS and TF scenarios

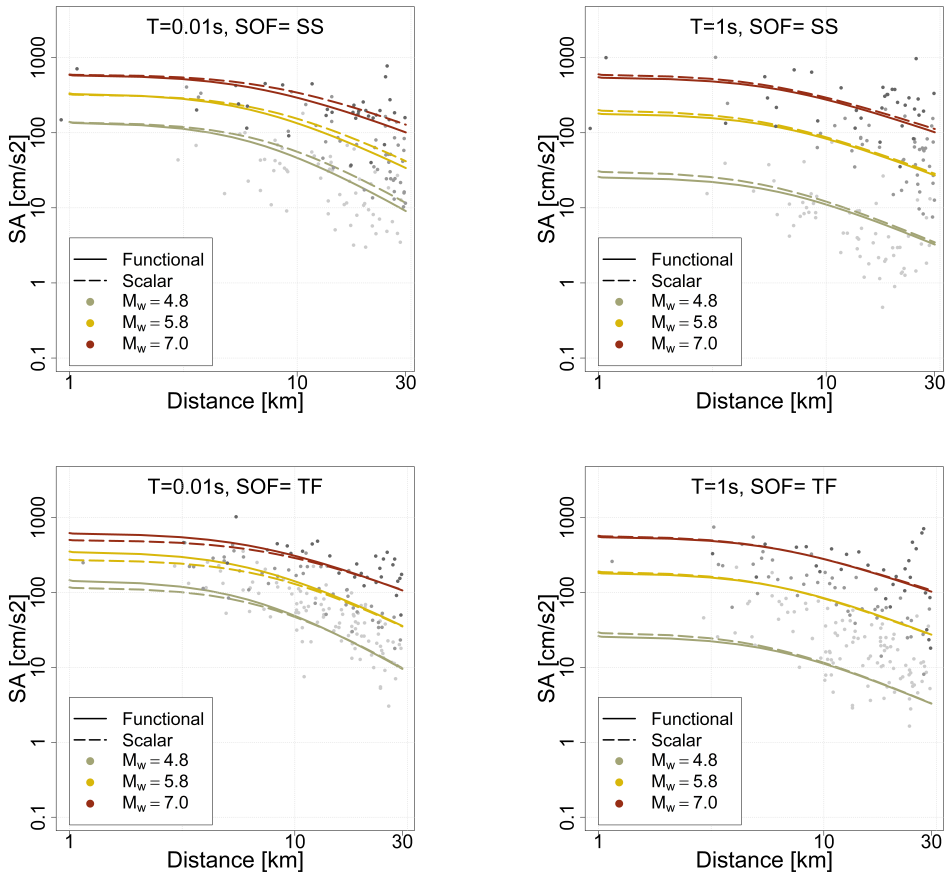


Figure B.8: Comparison of the ground motion predicted by Functional ITA18 and Scalar ITA18 for strike-slip and thrust faulting near-source scenarios, at $T = \{0.01s, 1s\}$. The continuous line is SA predicted by Functional ITA18. The dashed line is SA predicted by Scalar ITA18. In each plot, the lines are coloured according to the magnitude of the corresponding scenario. The dots are the observed values of SA, coloured darker as the magnitude of the corresponding event increases.

Appendix C

Codes

The application of the methodologies proposed in this work has been done through the implementation of algorithms in the `R` language.
The main `R` scripts used for the analysis can be downloaded from the repository of the author's GitHub account:
`https://github.com/tbortolotti/Weighted-functional-analysis-for-seismic-gmm.git`

In particular, the structure of the repository is organized as follows:

- Function `main`, that sequentially calls all the weighted functional methodologies leading from raw data to the diagnostic over the regression estimates.,
- Folder **method_comparison**,
- Folder **methods**,
- Folder **Regression**.

Below, the composition of the folders is illustrated briefly.

Folder **method_comparison**:
The folder contains the `R scripts` that allows one to test the soundness of the weighted functional analysis with respect to the reconstruction method.In particular,

- The main script `cv_event` performs the cross-validation. For every proposed method, it calls function $\texttt{method}_c omparison_e vent, givingasinputthe\texttt{reconstruction method}a$
- -- Function `my_reconstructKenipLiebl` and the corresponding folder **KL** are convenient local versions of the method proposed in the work of Kneip

and Liebl (2020), but no modifications are present with respect to the original scripts of the authors.

- Function `my_reconstructKraus1` implements the reconstruction method of Kraus (2015) for the estimation principal component scores of partially observed functional data. Function `finegrid_evaluations` simply evaluates the curves to be reconstructed on a finest grid, for the method to work better.

Folders **methods** and **Regression**:
These folder contain the implementations of all the methods that compose the analysis workflow that leads raw data to the estimation of the functional coefficients. Below, we comment on the main functions.

- Function `build_Sphi` operates the construction of the smoothing map, as illustrated in Section 3.2.2.
- Function `create_weights` associates each observation to a weight, constructed following the rationale illustrated in Section 4.3.1.
- Function `pwMSE_event` performs an event-wise cross-validation for the evaluation of the MSE.
- Function `stderrors` evaluates the variability associated to the coefficients estimates, as illustrated in Section 3.3.2.
- Function `wt_bsplinessmoothing` operates the weighted penalized smoothing of the reconstructed observations.
- Function `f_Regress` operates the weighted penalized regression of the smoothed observations, conditionally on the predictor variables given in input. A crucial input of the function is the parameter `blist`, which is the output of the function `lambda_select`, previously called, that operates a cross-validation to choose the penalization parameters for the coefficient estimates.