Leveraging weighted functional data analysis to estimate earthquake-induced ground motion

T. Bortolotti, R. Peli, G. Lanzano, S. Sgobba and A. Menafoglio

Abstract Ground motion models are fundamental tools for seismic hazard assessment, providing estimates of earthquake-induced ground motion based on seismic variables. A novel approach grounding on weighted functional data analysis is employed to extend a scalar ground motion model for Italy to the functional context. By incorporating observation-specific functional weights in the estimation routine, we aim to improve the accuracy and stability of model calibration in the presence of incomplete functional data. Through a simulation study, we show the effectiveness of the proposed methodology in enhancing the estimation accuracy and reducing variability compared to the traditional approach. Our findings highlight the potential of weighted functional data analysis to enhance the ground motion estimates and seismic hazard assessment, offering valuable insights for civil protection planning.

Key words: Functional data analysis, functional weights, partially observed data, ground motion model

Teresa Bortolotti
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, e-mail: teresa.bortolotti@polimi.it

Riccardo Peli
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, e-mail: riccardo.peli@polimi.it

Giovanni Lanzano
Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Milan, Italy, e-mail: giovanni.lanzano@ingv.it

Sara Sgobba
Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Milano, Milan, Italy, e-mail: sara.sgobba@ingv.it

Alessandra Menafoglio
MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, e-mail: alessandra.menafoglio@polimi.it

1 Introduction

Ground motion models play a critical role in seismic hazard assessment by estimating earthquake-induced ground motion based on seismic variables. ITA18 is a scalar ground motion model for Italy proposed in [3] which separately estimates the mean of a longitudinal sequence of ground motion intensity measures defined over a set of vibration periods. By leveraging a novel weighted approach in the field of functional data analysis, in [1] we extend the ITA18 model to the functional framework. The inherent incompleteness of seismic response profiles across oscillation periods poses a significant challenge, prompting the need for advanced methodologies that can enhance the accuracy and stability of the estimated functional ground motion model. We employ state-of-the-art methods to reconstruct the incomplete response profiles and incorporate observation-specific functional weights in the fitting of the model. In doing so, we aim to address the uncertainty associated with partially observed seismic data, seeking a solution that is somewhere between reducing the domain of analysis to the part common to all curves and considering the full domain by analysing solely the fully observed curves. Through a comprehensive simulation study, we evaluate the performance of our proposed weighted functional methodology under various scenarios. By assessing the impact of different reconstruction methods and weight definitions on estimation outcomes, we show the effectiveness of our approach in reducing the estimation uncertainty and enhancing predictive accuracy. Eventually, the proposed methodology proves effective in improving the estimation precision in the calibration of the ground motion model, offering a more reliable framework for seismic risk assessment and civil protection planning in regions susceptible to seismic activity.

2 Method

Let $y_1, \dots, y_n$ be reconstructed curves belonging to $L^2(T)$, where T is an open subset of $\mathbb{R}$. The weighted approach couples each y_i to a functional weight $w_i : T \to [0,1]$, taking value 1 where the curve is observed and decreasing to zero the more the reconstruction becomes uncertain. Analytically, this translates into formulating a new optimization criteria for concurrent regression, including the curve-specific functional weights. In doing so, full weight is given to errors made on the observed parts of the curves, and less weight to errors made on the reconstructed parts.

2.1 Weighted concurrent regression

The reconstructed functional observations are assumed to follow a concurrent regression model, namely

$$y_i = \sum_{j=1}^{q} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $x_{i1}, \ldots, x_{iq}$ are independent functional covariates, and $\beta_1, \ldots, \beta_q$ are the functional coefficients defined on T. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be realizations of a zero-mean stochastic process. We define a penalized fitting criterion which minimizes

$$\sum_{i=1}^{n} \int_T w_i \left(y_i - \beta_j x_{ij}\right)^2 + \sum_{j=1}^{q} \int_T \lambda_j \left(D^2 \beta_j\right)^2. \tag{2}$$

The second sum in (2) is a roughness penalty regularizing the estimates, and $\lambda_1, \ldots, \lambda_q$ are penalization parameters tuned via generalized cross-validation. Each coefficient is associated to a specific penalization parameter, meaning that the estimates of the coefficients are allowed to have diverse levels of smoothness.

The dimensionality of problem (2) is reduced by assuming that each β_j belongs to a finite dimensional space spanned by suitable basis functions $\theta_{j1}, \ldots, \theta_{jL_j}$. Under this assumption, the vector $\beta := (\beta_1, \ldots, \beta_q)$ can be expressed as $\beta = \Theta b$ and is uniquely identified by $b \in \mathbb{R}^L$, with $L = \sum_{j=1}^{q} L_j$. We show in [1] that the solution to problem (2) can be found in closed form for b, and reads

$$[J + R]\, b = \int_T \Theta(s)^T X(s)^T W(s) y(s) ds, \tag{3}$$

where $J := \int_T \Theta(s)^T X(s)^T W(s) X(s) \Theta(s) ds$, and R accounts for the ridge regularization.

2.2 Definition of the weights

We consider two alternative systems of weights, namely logistic weights and reconstruction driven weights.

Logistic weights Let the reconstructed curve y_i be originally observed up to t_i. Then the logistic weight is defined as

$$w_i(t) = \begin{cases} 1\,, & t \le t_i \\ \frac{1}{1+e^{(t-\mu_i)\alpha_i}} + d_i\,, & t > t_i \end{cases}, \tag{4}$$

where $\mu_i = (t_i + t_N)/2$ and $\alpha_i = a\hat{\sigma}_{t_i}$, $\hat{\sigma}_{t_i}$ being the empirical standard deviation of the observed values at t_i. Although continuity is not required, d_i is a corrective term guaranteeing continuity of the weight at t_i. Parameter $a > 0$ controls the rate of decay of the logistic functions. The advantage of using logistic weights lies in their interpretability and in the possibility of controlling the downweighting of the missing trajectory.

Reconstruction driven weights An alternative definition of the weights allows for fragmented patterns of missing information and relies on the covariance operator of the reconstruction error. Let y_i be a curve observed on O_i and reconstructed on M_i, $O_i \cup M_i = T$, $O_i \cap M_i = \emptyset$. Let $\mathscr{C}$ denote the covariance operator of y_i, and $\mathscr{V}_i$ the covariance operator of the reconstruction error. Additionally, let $c(t)$ be the diagonal of the kernel of $\mathscr{C}$ and $v_i(t)$ the diagonal of the kernel of $\mathscr{V}_i$. We define the reconstruction driven functional weight associated to y_i as

$$w_i(t) = 1 - \sqrt{\frac{\hat{v}_i(t)}{\hat{c}(t)}}, \ \forall t \in M_i, \tag{5}$$

where $\hat{c}(t)$ and $\hat{v}_i(t)$ are the sample versions of $c(t)$ and $v_i(t)$, respectively. Notice that $v_i(t)$ quantifies the amount of uncertainty associated to the predicted trajectory in t, while $c(t)$ quantifies the uncertainty that there would be about $y_i(t)$ if we ignored the observed part. Then, $w_i(t)$ quantifies the reduction in uncertainty on $y_i(t)$ achieved through the reconstruction.

3 Simulation study

A simulation study is conducted to validate the proposed methodology and assess its robustness under different scenarios. We investigate the effectiveness of the weights in reducing the impact of the adopted reconstruction methods on the estimates, evaluate the accuracy of estimates with varying weight definitions, and analyze the performance of the weighted methodology with increasing fractions of partially observed data. Through simulations, we demonstrate the effectiveness of weighted functional analysis in reducing the variability of the estimates compared to the unweighted functional approach. The predictive performance of the methodology is further analyzed through leave-one-out cross-validation (LOO CV) on synthetic data, evaluating the functional prediction error under different weight definitions. By examining the empirical distributions of prediction errors, we show in [1] the existence of a trade-off between fully relying on observed data and neglecting information from missing trajectories, highlighting the importance of finely-tuned weighting systems in achieving optimal estimation outcomes. Such finding is in agreement with the work of [5], where a classification problem on partially observed functional data is tackled by considering an intermediate domain extension between the common and the full domain.

4 Case study

The analysed dataset includes 5607 seismological records, relative to 146 earthquakes and 1657 stations [4]. The spectral acceleration profiles of each record are

considered as response of the ground motion model. Each profile is identified by 37 longitudinal ground motion intensity measures, namely the peak ground acceleration and the spectral acceleration recorded at 36 vibration periods in the interval $[0.04\ s, 10\ s]$. In the dataset under analysis, less than 75 % of the records are observed up to the largest period of 10 s, while the remaining profiles are only partially observed up to a certain period $T_i < 10\ s$.

The functional ground motion model that we propose is a functional extension of the model proposed in [3], which separately fits a scalar linear regression to the peak ground acceleration and the 36 longitudinal values of spectral acceleration. The embedding of the scalar model into a functional framework reads

$$\begin{aligned}\log_{10} SA = a + b_1(M_w - M_h)1_{(M_w \le M_h)} + b_2(M_w - M_h)1_{(M_w \ge M_h)} + f_1 SoF_1 + f_2 SoF_2 \\ + c_1(M_w - M_{\text{ref}})\log_{10} R + c_2 \log_{10} R + c_3 R + k \log_{10} \frac{V_0}{800} + e. \end{aligned} \tag{6}$$

In (6), SA is a random variable with values in the space of square integrable functions, a, M_h, M_{ref} and R are known functions with domain T and e is assumed to be generated by a zero mean stochastic process.

Before fitting (6), we carry out the three critical steps of calibrating the ridge penalization parameters entering (2), selecting the optimal system of weights, and selecting the optimal reconstruction method. In particular, the calibration of the ridge penalization, implying the search of an optimum in a nine-dimensional space, requires the use of an evolutionary algorithm for parameter selection (Centofanti et al. 2023). In all three steps, optimality is by means of a global measure of error in predicting the observed parts of the spectral acceleration profiles. To highlight the impact of the weighted functional methodology on the model calibration, we here comment on the estimates of coefficient c_1, obtained with the scalar and with the weighted functional approaches (see Figure 1). The results relating to the other model coefficients are in [1]. The functional boxplot associated to the point estimate is obtained via bootstrap sampling. The functional estimate follows the trend of the scalar estimate while displaying a smoother behavior. However, we notice a significant difference in the right half of the range due to how the data is weighted. When we extend the scalar model to longer periods, we overlook some information from partially observed SA profiles, relying only on 75 % of the data used for functional estimation. Conversely, by leveraging the correlation among SA ordinates for profile reconstruction and using a weighted functional approach, our method prevents this loss of information. This allows us to obtain dependable estimates throughout the relevant oscillation period range.

Acknowledgements Teresa Bortolotti and Alessandra Menafoglio acknowledge the support by MUR, grant Dipartimento di Eccellenza 2023–2027.

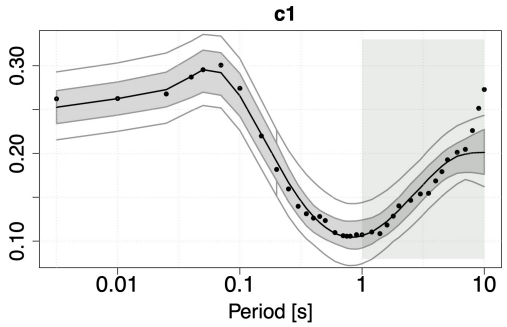

Fig. 1 Functional boxplot of the bootstrap sample of the estimated functional regression coefficient c_1 and comparison with its ITA18 point estimates (black dots). The black line represents the functional point estimate of the coefficient. The bootstrap sample has size 1000.

References

1. Bortolotti T., Peli R., Lanzano G., Sgobba S., Menafoglio A.: Weighted Functional Data Analysis for the Calibration of a Ground Motion Model in Italy. Journal of the American Statistical Association. 1-19 (2024)
2. Centofanti F. et al.: Adaptive smoothing spline estimator for the function-on-function linear regression model. Computational Statistics. **38**, 191-216 (2023)
3. Lanzano G. et al.: A Revised Ground Motion Prediction Model for Shallow Crustal Earthquakes in Italy. Bulletin of the Seismological Society of America. **109(2)**, 525-540 (2019)
4. Lanzano G. et al.: Parametric table of the ITA18 GMM for PGA, PGV and Spectral Acceleration ordinates. Istituto Nazionale di Geofisica e Vulcanologia (INGV) (2022) https://doi.org/10.13127/ita18/sa_flatfile/
5. Stefanucci M., Sangalli L. Brutti P.: PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set. Statistica Neerlandica. **72(3)**, 246–264 (2018)