Functional linear models for the analysis of similarity of waveforms

Modelli lineari funzionali per l'analisi della similaritá di forme d'onda

Francesca Di Salvo, Renata Rotondi and Giovanni Lanzano

Abstract In seismology methods based on waveform similarity analysis are adopted to identify sequences of events characterized by similar fault mechanism and propagation pattern. Seismic waves can be considered as spatially interdependent three dimensional curves depending on time and the waveform similarity analysis can be configured as a functional clustering approach, on the basis of which the membership is assessed by the shape of the temporal patterns. For providing qualitative extraction of the most important information from the recorded signals we propose an integration of the metadata, related to the waves, as explicative variables of a functional linear models. The temporal patterns of this effects, as well as the residual component, are investigated in order to detect a cluster structure. The implemented clustering techniques are based on functional data depth.

Abstract *In sismologia i metodi basati sull'analisi della similaritá tra onde sismiche vengono impiegati per l'identificazione di eventi caratterizzati dallo stesso meccanismo di frattura o di propagazione. Le onde sismiche possono essere considerate come curve tridimensionali, funzioni del tempo e correlate nello spazio, e l'analsi della similaritá delle forme d'onda si puó configurare come un'analisi di clustering funzionale, secondo cui l'appartenenza di un'onda ad un cluster si stabilisce in relazione al pattern temporale. Al fine di estrarre una corretta informazione dai sismogrammi, viene proposto un approccio che integra nell'analisi anche i metadati, riferiti alle onde; questi vengono considerati come esplicative di un modello lineare funzionale. Il pattern temporale degli effetti e della parte residuale, viene analizzato per l'individuazione di strutture di cluster. Le tecniche di clustering implementate sono basate su misure di 'data depth' funzionale.*

Francesca Di Salvo
Department of Agriculture, Food and Forest Sciences, University of Palermo e-mail: francesca.disalvo@unipa.it

Renata Rotondi
CNR IMATI, Milan e-mail: reni@mi.imati.cnr.it

Giovanni Lanzano
INGV, Milan e-mail: giovanni.lanzano@ingv.it

Key words: structured functional principal component, waveforms clustering, functional data depth

1 Introduction

The problem of investigating the seismotectonic structures of an area involves several methods based on waveform similarity analysis. In particular, seismic networks often record signals characterized by similar shapes and methods studying their similarity are adopted to identify sequences of foreshock, main shock and aftershock; in this field the goal is the definition of group of events characterized by similar fault mechanism and propagation pattern, under the hypothesis that a group of dependent events (multiplets) represents a chain led by seismogenetics background of a common earthquake.
The detection of earthquake families or multiplets is finalized to the identification of sources related to the same fault [6] or to obtain instrumental catalogues of independent earthquakes cleaned of dependent ones [3].

Statistical approaches are powerfull tools for detecting dependent events in a seismic data set; waveform similarity analysis is considered to join seismic episodes into a single multiplet [3]; clustering has been demonstrated as a useful method for identifying members of the same group that possess similar waveform. Different techniques for assessing the cluster membership of a earthquake are also reported in literature [8], [1].

Seismic waveform contains information of multiple attributes, that make up the set of integrating metadata concerning the source, the localization of the recording, the time of the event, the dynamics of the registration. A common opinion is that this class of data are also noisy and most techniques, including clustering, can be optimized by using appropriate data preprocessing. Signal filtering as singular value decomposition, as well as short-time Fourier transforms (STFT) are recognized as proper techniques for extracting the key features [9], [5]. In section 2, this complex functional structure is faced by models with explicit functional effect components; this step allow to integrate information from the metadata and to implement cluster analysis on the residual part of the model. Approaches relied on depth measures are considered in order to construct robust tools for the clustering of the curves. In the application, the approach is applied to a set of recordings of the seismic sequence Amatrice - Norcia - Visso, from August 2016 to January 2017, provided by the Engineering Strong Motion database (ESM).

2 The methodology

Seismic waves, that are three dimensional spatially interdependent curves, can be considered as realizations of a multivariate functional random field:

$$f^p(t) = Y^p(t) + \varepsilon^p(t)$$

The couple $(t, f^p(t))$ denotes the time and the oserved value of the function $Y^p(t)$at time t and $t \in [0,T]$. Standardizing the time interval in $[0,1]$:

$Y^p = \{Y^p : [0,1] \to R\}, p = 1,2,3$ is the set of real-valued functions on the closed interval $[0,1]$:

Each $Y^p(t)$ has a set of attributes, among which here we consider the event (metadata concerning the origin identification) and the distance from epicentre of the site where the signal is recorded; we indicate with I the number of events and with J the number of discretized distances. The curves are then indicized by a pair of indexes (i, j) in order to indicate the recordings $f_{ij}^p(t)$ of the i^{th} event at a distance d_j:

$$Y_{11}^p(t), \ldots Y_{ij}^p(t) \ldots Y_{IJ}^p(t) \quad | \quad Y_{ij}^p \in Y^p,\ t \in [0,1]$$

.

The presence of these effects on their dynamics is modeled in a functional two-way crossed design [10] [11]:

$$Y_{ij}^p(t) = \mu^p(t) + X_i(t) + Z_j(t) \tag{1}$$

Each curve is decomposed into an event-effect $X_i(t)$ and epicentral distance effect $Z_j(t)$, both affecting its shape. Using the Karhunen-Loéve expansion for the effects $X_i(t),\ Z_j(t)$ the model (1) becomes:

$$Y_{ij}^p(t) = \mu^p(t) + \sum_{k=1}^{\infty} \phi_k^X(t)\xi_{ik}^X + \sum_{l=1}^{\infty} \phi_l^Z(t)\xi_{jl}^Z \tag{2}$$

The proposed approach is described for $p = 1$ but its generalization for $p > 1$ can follow from the framework of functional principal components for multidimensional curves. The functional processes $X_i(t), Z_j(t)$ characterize the observed variability and their covariance operator can be estimated using method of moments estimators from the observed curves, [10]. Principal component decomposition of covariance operators results in structured principal components aims to derive the effects of the two attributes on the temporal pattern of the waves: other structures and correlation on the data are detected on the basis of the shape of the temporal patterns using a functional data depht measure. Based on the concept of data depth, this robust nonparametric tool is applied for clustering purposes in the functional data setting, providing an order within a sample of curves. Data depth notion measures the centrality of an observation within a sample and allows the definition of a natural ordering from center outwards; several depth notions generalize unidimensional concept, here we focus on Modified Band Depth [7]. Given the set of n continuous functions $f(t)$ in $[0,1]$ (the double indicization i, j is not necessary here) and λ, a Lebesgue measure in $[0,1]$, for any f of the sample the *Modified Band Depth* is the portion of time that $f(t)$ is inside the regions, made up of $2,3,\ldots,k,\ldots,K$ of the n curves:

$$MBD_n(f) = \sum_{k=2}^{K} n_{(k)}^{-1} \sum_{1<\ldots r_1, r_2 \ldots <n} \frac{\lambda(A(f; f_{r1}, f_{r2}))}{\lambda(T)} \tag{3}$$

where $A(f; f_{r1}, f_{r2})$ the region delimited by f_{r1} and f_{r2}, is as follows:

$$A(f; f_{r1}, f_{r2}) = \{f(t) : min_{r=r_1,r_2} f_r(t) \leq f(t) \leq max_{r=r_1,r_2} f_r(t)\} \tag{4}$$

Two previous paper [2], [4] describe the algorithm based on the notion of Modified Band Depth and adopted here for clustering the curves.

3 The application

A set of recordings of the seismic sequence Amatrice - Norcia - Visso, from August 2016 to January 2017 is considered; data are preprocessed through alignment techniques [12] dealing with different lengths of the sequences, temporal aspects, and signal filtering. The resulting sequences are aligned signals of the same length, sampled at 10 Hz; the curves are represented in figure 1 (a) .After having estimated the model (2), the estimated effects $\hat{X}_i(t)$ and $\hat{Z}_i(t)$, are represented in figure 1, respectively in (b) and in (c). The figure 2 (left) shows the functional residuals from

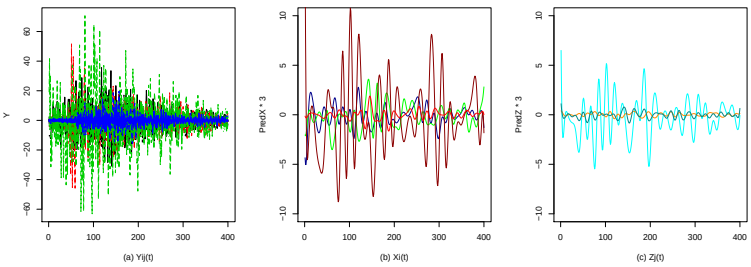

Fig. 1 Waves (a) and estimated components (b) and (c)

the functional model (2), that are the input of the Modified Band depth algorithm. The algorithm finds an intrinsic order within the set of the curves, and the similarity between consecutive curves can be analyzed. In the figure 2 (right) the kernel of 50% inner curves is represented (on the top) from the 50% of the the most external curves (bottom). It is evident that the two groups differ not for shape but for the amplitude variability. As intermediate result we report in figure 3 a structure of three groups, the (25%) of the inner curves (a), the (25%) of intermediate curves (b) and the (50%) of the external curves (c): the separation is obtained on the basis of *MBD* and the clusters are well separated. The deepest curves of the three clusters are compared in (d).

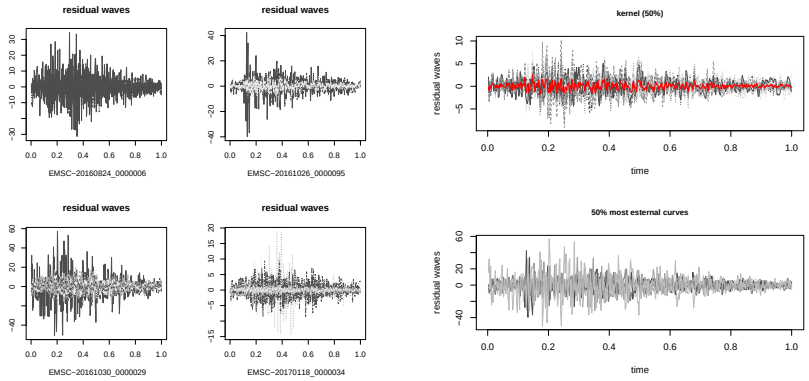

Fig. 2 left: Residuals for the main 4 events; right: the kernel (top) and the external curves (down) of the whole set of residuals from the functional linear model.

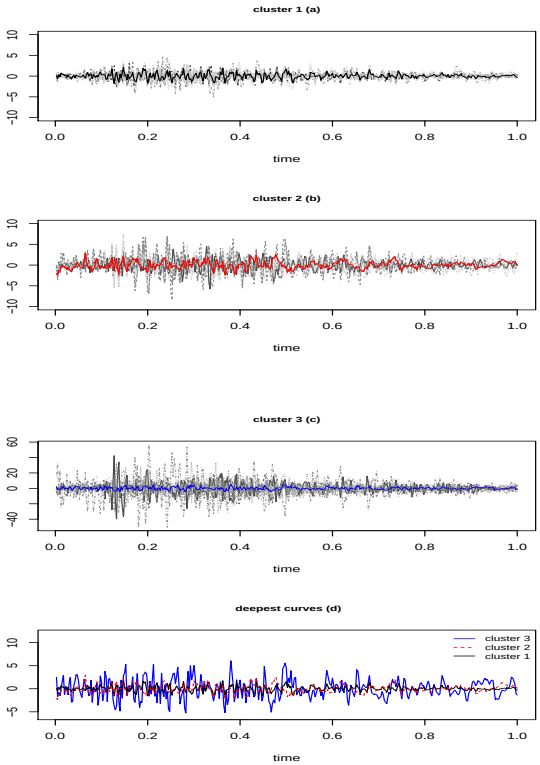

Fig. 3 Three final clusters (a) , (b), (c) obtained from ordered residuals waves. The deepest curves of the three clusters (d).

4 Results and discussion

The proposed approach is an adaptive data-driven method based on the integration of information from metadata in the analysis of the temporal pattern of the waves. Seismic waveform and attributes are used as inputs in a clustering process. Since the seismic waveform contains also some unnecessary information such as noise, they are preprocessed through alignment technique. The functional linear model derives the variability due to the events and the epicentral distance: this results improves the final identification of the clusters of curves with similar amplitude and high correlations. The interest is motivated by further analysis finalized to the study of other seismic features or geological features of the sites. The methodology can be easily extended from two-way to a m-way model [11] and to the case of three-dimensional waves, as the functional principal component analysis is well known technique for $p-$dimensional curves, $p > 1$.

References

1. Adelfio, G., Chiodi, M., D'Alessandro, A. and Luzio, D., D'Anna, G., Mangano, G. Simultaneous seismic wave clustering and registration. Computers Geosciences, 8(44), 6069 (2012)
2. Adelfio G., Di Salvo F., Sottile G. Depth-based methods for clustering of functional data TIES 2017 Conference, Bergamo, Italy, July 24th 26th, (2017).
3. Barani, S. Ferretti, G., Massa, M., Spallarossa D. : The waveform similarity approach to identify dependent events in instrumental seismic catalogues , Geophys. J. Int. doi: 10.1111/j.1365-246X.2006.03207.x (2006)
4. Di Salvo, F., Rotondi, R., Lanzano, G.: Detecting clusters in spatially correlated waveforms, GNGTS conference, Trieste, November 13th - 16th (2017)
5. Hao-kun, D., Jun-xing, C.,Ya-juan, X., Xing-jian, W., Seismic facies analysis based on self-organizing map and empirical mode decomposition, Journal of Applied Geophysics, 112, 5261 (2015)
6. Jagla, E. A., Kolton A. B.: A mechanism for spatial and temporal earthquake clustering, J. Geophys. Res. doi:10.1029/2009JB006974 (2010)
7. Lopez-Pintado, S., Romo, J., : Depth-based inference for functional data, Computational Statistics and Data Analysis 51 (10), 4957-4968, (2007).
8. Reasenberg, P. : Second-order moment of Central California seismicity, 1969 - 1982, J. geophys. Res., **90**, 54785495 (1985)
9. Silvestrov, I., Tcheverda V.:SVD analysis in application to full waveform inversion of multicomponent seismic data,Journal of Physics: Conference Series 290 doi:10.1088/1742-6596/290/1/012014 (2011)
10. Shou, H., Zipunnikov, V., Crainiceanu, C. M., Greven, S. : Structured Functional Principal Component Analysis, Biometrics, 71(1), 247257 doi.org/10.1111/biom.12236, (2015)
11. Suk, H. W., Hwang, H. : Functional Generalized Structured Component Analysis. Psychometrika, 81(4), 940968. doi.org/10.1007/s11336-016-9521-1, (2016)
12. Tucker, D.J., Wu, W. , Srivastava, A., Generative models for functional data using phase and amplitude separation, Computational Statistics and Data Analysis, 61, 5066, (2013)