Dissimilarity Measures for the Identification of Earthquake Focal Mechanisms

Francesco Benvegna[1,2], Giosué Lo Bosco[1,2,3], and Domenico Tegolo[1,3]

[1] Dipartimento di Matematica e Informatica, Universitá degli Studi di Palermo, Italy
[2] I.E.ME.S.T., Istituto Euro Mediterraneo di Scienza e Tecnologia, Palermo, Italy
[3] C.I.T.C, Centro Interdipartimentale di Teconologie della Conoscenza, Palermo, Italy

Abstract. This work presents a study about dissimilarity measures for seismic signals, and their relation to clustering in the particular problem of the identification of earthquake focal mechanisms, i.e. the physical phenomena which have generated an earthquake. Starting from the assumption that waveform similarity implies similarity in the focal parameters, important details about them can be determined by studying waveforms related to the wave field produced by earthquakes and recorded by a seismic network. Focal mechanisms identification is currently investigated by clustering of seismic events, using mainly *cross-correlation* dissimilarity in conjunction with hierarchical clustering algorithm. By the way, it results that such adoptions have not been sufficiently validated. To shed light on this we have studied the cross correlation dissimilarity on simulated seismic signals in conjunction with hierarchical and partitional clustering algorithms, and compared its performance with a newly one recently introduced for the purpose called *cumulative shape*. In particular, we have properly created synthetic waveforms related to two types of focal mechanisms, showing that the cumulative shape perform better than cross-correlation in the identification of the expected clustering solution.

Keywords: metrics, clustering, seismic signals, waveforms.

1 Introduction

The seismograms are recordings of ground motion which record how the movements have taken place and how have been transmitted through the ground. Seismic waves are spread of low-frequency acoustic energy generated by an earthquake, an explosion or a volcano eruption. They travel through the different level of the Earth's underground where they can be deviated and reflexed by each layer of the earth's crust. During a seismic event, it is possible to identify different groups of waves characterized by several amplitudes and frequencies. These groups may be of different types depending from the waves and their propagation. The main types of waves that are identified are:

– Primary waves *(*P-waves): they are compressional waves orthogonal to ground motion. These are the first waves to arrive and to be detected by the instruments.

A. Petrosino (Ed.): ICIAP 2013, Part II, LNCS 8157, pp. 500–509, 2013.

- Secondary waves (S-waves): they are shear waves parallel to ground motion that arrive after the p-waves.
- Surface Love waves: they are a combination of P-waves and S-waves reflections traveling along the surface which produces entirely horizontal motion.
- Surface Rayleigh waves: combination of P-waves S-waves reflections traveling along the surface which moves the ground up and down, and side-to-side in the same direction of the wave.

P and S-waves are called *body* waves since they travel in the interior of the Earth as opposed to surface waves. The differences between them are the transmission geometry and velocity. For example, in the upper crust the typical p-waves velocity is about 6 km/s while s-waves go at 3.5 km/s. The amplitude of the body waves through an homogeneous elastic medium decreases with the distance and at the same distance the power of the S-waves is greater than P-waves.

Many research activities are related to seismic waves but two of them are very interesting and related to data analysis: *locating hypocenters* and *investigating on focal mechanisms*. The first problem tries to give a suitable location to an unlocated event by the comparison to a well located master event, while the latter is related to the physical phenomena which have generated an earthquake.

In this work we focus our attention on focal mechanisms. In this case, the main assumption is that waveform similarity implies similarity in the focal parameters, so that important details about them can be determined by studying waveforms related to the wave field produced by earthquakes and recorded by a seismic network. For this purpose, focal mechanisms identification is currently investigated by clustering of seismic events, using mainly cross-correlation and/or cross-spectral dissimilarities in conjunction with hierarchical clustering algorithm.

Indeed, in seismology cross-correlation and/or cross-spectral dissimilarities seems to be the most used. Barani et al. [1] use the application of cross-correlation analysis to define groups of dependent events (multiplets) characterized by similar location, fault mechanism and propagation pattern. Badaway et al. [2] did a good analysis on source parameters and fault plane determinations by use of cross-correlation. They use the cross-correlation distance in a classification phase before to develop a focal mechanisms solution.

Furthermore, other dissimilarities have been defined and used for the analysis of seismic signals. In particular, we have recently proposed the so called *cumulative shape dissimilarity* which is based on the difference between the cumulative energies of the signals rather than on their original waveforms [3]. Its reliability has been tested on real seismic data, showing very good result in terms of cluster homogeneity and computational time.

In this paper we deal with focal mechanism identification, by using hierarchical and partitional clustering algorithms in conjunction with *cross-correlation* and *cumulative shape* dissimilarities. To this purpose we have generated several synthetic dataset which simulate seismic events generated by two kinds of

focal mechanisms. Results carried out on such simulated data, show that the cumulative shape is preferable to the cross correlation dissimilarity.

The paper is organized as follows: in the next section we describe the main components involved in waveform analysis of seismic events, section 3 describe the software tool used for generating seismic events related to particular focal mechanism, section 4 describe the experiments and shows the computed results, the last section offers conclusion an future directions.

2 Waveform Analysis

2.1 Data Collection

As described in [4] seismic events can be divided into *artificial* and *natural*. The first are generated by simulation that can be executed in real or in virtually environment, while the other are generated by natural factor such as tectonic earthquakes, volcanic earthquakes and storm microseisms.

The artificial events such as explosions or rock bursts are generated by human activity focusing on scientific aim. These experiments are often a controlled sequence of bursts used to test a detection grid and the transmission medium. Another opportunity is to generate synthetic waveforms by simulating virtually a particular physical model. This allow to test extensively the data analysis methodologies that can be used in order to infer focal mechanism properties.

The natural events are stored when they occur and the researchers study their causes and behavior. Due to different sources an event may be localized from few kilometers of depth up to 700 kilometers. The so called *tectonic earthquakes* are the most dangerous and can be very destructive with a magnitude greater than 6. *Volcanic earthquakes* have a small energy and duration of tremor type, and some instruments have difficult in recordings this type of events. *Miscroseisms* are generated by storms over oceans or large water basins, and are not well localized nor fixed to an origin time.

2.2 Preprocessing

Seismograms are affected by noise and by a not well identified signal portion which corresponds to the event of interest. The noise can make complex the use of a similarity measure between signals because a fine grained measure can look the noise as a significant component of the seismic event. Two seismic events may be generated by the same source, but a different noise could change the waveforms making them quite different. A filtering phase is a mandatory activity in the data preprocessing and must be executed with care. In particular, a light filter could not clean the signal in the right way, otherwise an hard filter could remove a meaningful portion of the signal. Band filters is the common choice to clean seismic signals. Picking and triggering [5] are other important preprocessing techniques in seismograms analysis. Picking is devoted to find the event inside a long seismogram by the identification of P and S-waves, while the main goal of triggering is the automated recognition of the seismic event regardless the noise

background. The two techniques are applied together on seismograms analysis in order to select an event inside a whole signal and to detect the phases on it. Indeed, a trigger algorithm usually found also the P-phase of the event as start so that a picking algorithm can find others. A well known triggering algorithm is the *Z-detector* [6]. This method uses a standardization $Z(i)$ computed on the original discrete signal $x(i)$ defined as:

$$Z(i) = \frac{x(i) - \mu_x}{\sigma_x} \tag{1}$$

where μ_x and σ_x indicates mean and standard deviation of the discrete finite signal $x(i)$. A great advantage of using Z-detector is a good behaviour in background noise's presence. Of course the threshold level required to select the start and the end of the event depends from the background noise.

2.3 Dissimilarity Measures for Seismic Signals

The *cross-correlation* is one of the most used proximity measure for clustering and classification of seismic events. Giving two discrete signals $x_1(i)$ and $x_2(i)$ both of finite length n, the cross correlation dissimilarity is defined as:

$$\delta_R(x_1, x_2) = 1 - \frac{1}{\sigma_{x_1}\sigma_{x_2}} \underset{k=1,..,2n-1}{max} R_{x_1,x_2}(k-n). \tag{2}$$

Where R_{x_1,x_2} denotes the *cross correlation* between x_1 and x_2, and is defined as follows

$$R_{x_1,x_2}(k) = \begin{cases} \sum_{i=0}^{n-k-1} (x_1(i+k) - \mu_{x_1}) \times (x_2(i) - \mu_{x_2}) & \text{if } k \geq 0 \\ R_{x_1,x_2}(-k) & \text{otherwise} \end{cases} \tag{3}$$

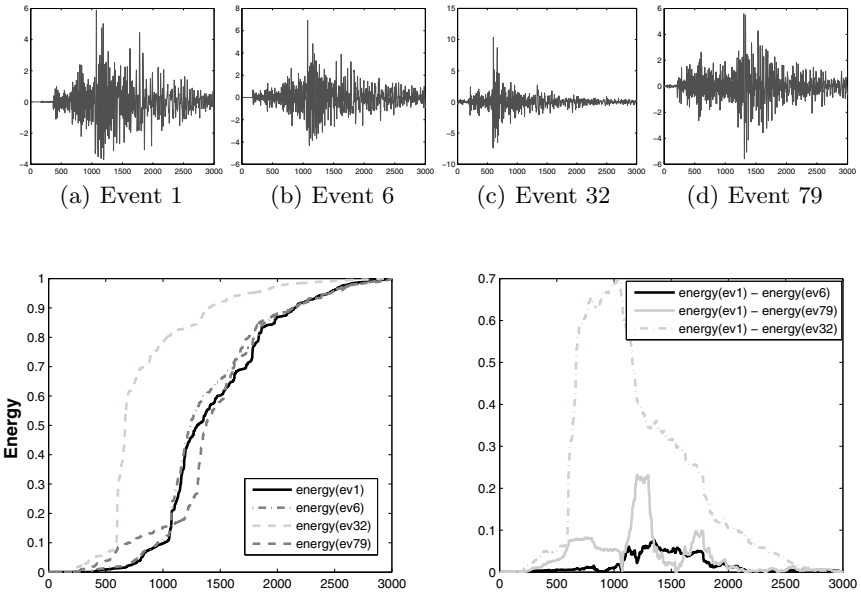
for $k = 1-n, .., n-1$. Such dissimilarity is largely used to catch difference in shape between seismic signals, but in this context it has also shown some drawbacks. One of the most important, is that for a signal of length n its computational time is $O(n^2)$.

Recently, we have proposed the so called *cumulative shape dissimilarity* δ_s [3]. It is based on on the difference between the cumulative energies of the signals rather than on their original waveforms. This assures that it fully satisfies three important properties: (a) it gives high weight to the difference among the initial part of the signals, (b) it is very low sensitive to background and impulsive noise and (c) it is capable of detecting where two wave shapes are similar regardless of magnitude. For completeness, we recall its definition:

$$\delta_s(x_1, x_2) = \sum_k |sd_{12}(k+1) - sd_{12}(k)| \tag{4}$$

where sd_{12} denote the cumulative sums between x_1 and x_2, that is:

$$sd_{12}(k) = \left| \frac{\sum_{r=1}^{k} x_1^2(r)}{\sum_{r=1}^{n} x_1^2(r)} - \frac{\sum_{r=1}^{k} x_2^2(r)}{\sum_{r=1}^{n} x_2^2(r)} \right| \tag{5}$$



(a) Event 1 (b) Event 6 (c) Event 32 (d) Event 79

(e) Cumulative energies of the events (f) Difference of cumulative energies

Fig. 1.

Equation 5 defines a normalized non-decreasing curve with values between 0 and 1. Finally, δ_s in 4 represents the sum of the derivative of the difference between the cumulative sums of x_1 and x_2. In some way the cumulative shape checks that the two cumulative energy curves rise in the same way at the same time. The value of the measure is higher when the energy curves have the same shape. Figures 1a-d show four example of seismic events, figure 1e their cumulative sums while figure 1f the differences between them. We can observe that the cumulative energies plot (1e) allows to identify the P and S-waves arrivals as the two concavities of the curves. Moreover, is less sensible in time to amplitude values of the tail: the curve rises quickly at first, but less when the value of the initial energy is added to that remaining. The background noise is less evident because its value is constantly added to the energy curve. Two curve with similar shape have a similar cumulative energy curve so is more simple to detect close events (see event 1 and event 6 in figure 1).

The application of the cumulative shape requires a good cut of the signal at P-waves arrival. Although a fast alignment can be applied between two signal on the first part of them, is preferable to have each signal that starts with P-waves. Note that the cumulative shape is more sensible to signal triggering than cross-correlation, anyway a good cut is a necessary condition for both. Cross-correlation dissimilarity is indeed affected by the cut because its value is

computed on different subpartition of the signal so that an unrequited or a portion miss can affect the computed value in different ways. Finally, it is important to remark that cumulative shape measure is faster than cross-correlation ($O(n)$ vs $O(n^2)$).

2.4 Clustering and Validation

Clustering is an unsupervised learning technique used on pattern recognition for data partitioning [7] that can be seen as a three step process [8,9]: (1) choice of a distance function; (2) choice of a clustering algorithm and (3) choice of a methodology to assess the statistical significance of clustering solutions. Regarding point (2), we recall that the main distinction between clustering algorithm is among *hierarchical* and *partitional* methods. Hierarchical algorithms are the most known, they are simple to implement and to understand also for naive users. A great advantage of these algorithms is that the cluster solution is composed by a tree structure called *dendrogram*. Among the hierarchical methods for clustering, the most used are known as *single-linkage*, *complete linkage* and *average linkage*. The partitional clustering techniques create a flat configuration, a partitioning, of the data with a desired number of clusters K. The most well known partitional algorithm is the k-means algorithm. An extension of the k-means is the k-medoids algorithm [10], where the *medoid* is a prototype of a cluster that can be different from the simple centroid. Generally, the performance of a clustering algorithm can be established by means of *clustering internal and external indices*: the former gives a reliable indication of how well a partitioning solution captures the inherent separation of the data into clusters [11], the latter measures how well a clustering solution agrees with the *ground truth* for a given data set. Due to the nature of our supervised experiments, we dispose of a ground truth so the right choice is the adoption of the external indices, due also to their superior accuracy compared with internal ones. In this work we have used the `Adjusted Rand index` [12] and [13],

3 Simulation Model

The possibility to generate synthetic waveforms by simulating virtually a particular physical model allows to test extensively the data analysis methodologies that can be used in order to infer focal mechanism properties.

The E3D simulation tool developed by the Lawrence Livermore National Laboratory of the University of California, is a software tool which allows to generate seismic signals following a particular simulation model, based on models defined by [14], [15], [16] and [17]. The software simulates wave propagation by solving the elastodynamic formulation of the full wave equation on a staggered grid. The solution scheme is 4th-order accurate in space, 2nd-order accurate in time.

The computation of a simulated dataset requires a long computational time and lot of resources. Theses simulations are usually executed on high performance clusters (HPC) because they use a massive parallelism.

The first step for data simulation is the simulation of the earth structure. To this purpose, a velocity model composed by five block is a good assumption. Each block is an horizontal layer over the distance between the detection station and projection of the source in the earth surface. The hypothesis of an horizontal layer is used to simplify the model without losing some real characteristics of the Earth's crust. Each block is defined by six parameters:

Start depth: starting z position of block element
End depth: ending z position of block element
Gradient: vertical gradient (units per km)
P: P-wave velocity in km/sec
S: S-wave velocity km/sec
r: density g/cm^3

The previous parameters are used to describe the physics of the propagation path in a more realistic way. The values are fixed to simulate a real mean of several rock types. A fault source is defined by three main parameters: strike, dip and rake. As described in [18] we report the definition of the previous parameters:

Strike: it is the direction of a line created by the intersection of a fault plane and a horizontal surface, 0° to 360°, relative to North. Strike is always defined such that a fault dips to the right side of the trace when moving along the trace in the strike direction. The hanging-wall block of a fault is therefore always to the right, and the footwall block on the left. This is important because rake (which gives the slip direction) is defined as the movement of the hanging wall relative to the footwall block.
Dip: it is the angle between the fault and a horizontal plane, 0° to 90°.
Rake: it is the direction a hanging wall block moves during rupture, as measured on the plane of the fault. It is measured relative to fault strike, ±180°. For an observer standing on a fault and looking in the strike direction, a rake of 0° means the hanging wall, or the right side of a vertical fault, moved away from the observer in the strike direction (left lateral motion). A rake of ±180° means the hanging wall moved toward the observer (right lateral motion). For any rake>0, the hanging wall moved up, indicating thrust or reverse motion on the fault; for any rake $< 0°$ the hanging wall moved down, indicating normal motion on the fault.

Once defined the model for the focal mechanisms, the simulation can occur after having properly located the source event and the detection station across the model.

4 Experimental Results

To perform simulation and test we used the E3D simulation tool on a computing infrastructure based on an high performance cluster. On our simulation we have chosen a 2D model on a grid of size $30Km$ of length and $60km$ of depth. The source event are located at length $7.5km$ from origin while the detection station

Table 1. Layer of the simulated model

Start depth	End depth	Gradient	P	S	r
0	5	0.25	3.1	1.5	2.3
5	17	0.05	5.8	3.3	2.67
17	28	0.02	6.7	4.5	2.8
28	31	0.45	6.92	4.7	2.9
31	60	0.01	8.25	5.5	3.2

is at length $22.5km$. The blocks which describes the crust structure are reported on table 1, and have been defined by expert seismologists.

A full 3D model must include the strike, dip and rake parameters but in a 2D model we can use only the first two. In details we have created two different model of focal mechanisms:

Strike slip: have walls that move sideways, not up or down. That is, the slip occurs along the strike, not up or down the dip. In these faults, the fault plane is usually vertical, so there is no hanging wall or footwall. The forces creating these faults are lateral or horizontal, carrying the sides past each other.

Reverse: form when the hanging wall moves up. The forces creating reverse faults are compressional, pushing the sides together.

We have defined two sources located at increasing depths in the range [2040] km by step of 5 km. We recall that the sources are located at length $7,5$ km from the origin. For each one of the 5 source location, we have generated 20 events of length 20 seconds in al cloud large about $2-3$ km around the basic depth. The simulation defines a ground truth of $K = 2$ clusters, each one composed by the 20 events generated by the two simulated focal mechanisms. In all the experiments, we have used the average linkage algorithm as hierarchical clustering method, while the k-medoid algorithm as partitional. In this latter case, the prototype for each cluster is represented by the median event of each cluster.

The evaluation index computed on the results is the Adjusted Rand Index. We recall that the higher its value is, the better the clustering result. In particular, values closer to 1 shows a clustering solution closer to the ground truth.

Figure 2 summarize results of Adjusted rand for the two considered distance computed by hierarchical and partitional clustering, for each one of the 5 considered source locations.

Results show that the Average linkage seems to be the right algorithm for studying the focal mechanism. This is due to the average values of adjusted rand computed for both distances by the hierarchical algorithm, that is better than what obtained for the partitional one (0.88 vs 0.62). Moreover, the cumulative shape has shown to be the right choice in both the hierarchical and partitional cases (0.84 vs 0.65) . Anyway, the significative difference between the two distances is in terms of computational time, since cumulative shape is linear while cross correlation is quadratic.

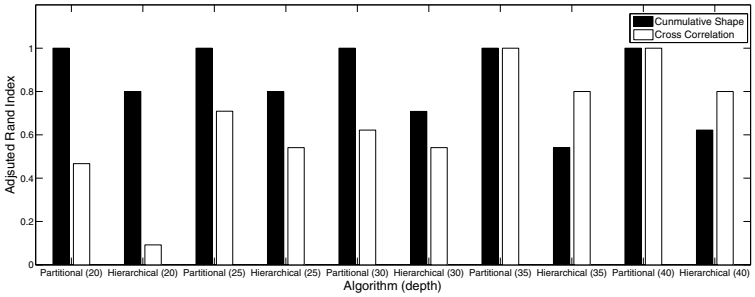


Fig. 2. Results computed by hierarchical and partitional method for 5 source location at depths 20,25,30,35,40. Hierarchical and Partitionals' Adjusted rand values are shown in order for all the 5 considered depths.

5 Conclusion

In this paper we have tested two dissimilarity measures for seismic signals in the specific problem of focal mechanisms identification. Our interest was to understand if the common adoption in the seismic data analysis signals framework to use *cross-correlation* dissimilarity in conjunction with hierarchical clustering algorithm can be also used for identifying focal mechanisms of an earthquake. To this purpose, we have compared the cross correlation dissimilarity on simulated seismic signals in conjunction with hierarchical and partitional clustering algorithms with a newly one recently introduced for the purpose called *cumulative shape*. Results confirm that the hierarchical clustering algorithm (in particular average linkage) seems to be also the right adoption for focal mechanisms identification. Moreover, the cumulative shape has shown to be preferable to the classical cross correlation especially because of the reduced computation time. Future developments will be devoted to extend this simulation to more than two sources adopting a full 3D model.

Acknowledgement. The authors would like to thank Prof. Dario Luzio (DISTeM, University of Palermo) and Dr. Antonino D'Alessandro (Istituto Nazionale di Geofisica e Vulcanologia, Palermo) for the definition of the 2D model parameters, and Dr. Francesco Grigoli (Institute of Earth and Environmental Science, University of Potsdam) for the suggestion about the use of the simulation software.

References

1. Barani, S., Ferretti, G., Massa, M., Spallarossa, D.: The waveform similarity approach to identify dependent events in instrumental seismic catalogues. Geophysical Journal International 168(1), 100–108 (2007)

2. Badawy, A., Fattah, A.K.A.: Source parameters and fault plane determinations of the 28 december 1999 northeastern cairo earthquakes. Tectonophysics 343, 63–77 (2001)
3. Benvegna, F., D'Alessando, A., Bosco, G.L., Luzio, D., Pinello, L., Tegolo, D.: A new dissimilarity measure for clustering seismic signals. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part II. LNCS, vol. 6979, pp. 434–443. Springer, Heidelberg (2011)
4. Borman, P.: New Manual of Seismological Observatory Practice. IASPEI, GFZ German Research Centre for Geosciences, Potsdam (2012)
5. Mitchell, W., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S., Trujillo, J.: A comparison of select trigger algorithms for automated global seismic phase and event detection. Bullettin of the Seismological Society of America 88(1), 95–106 (1998)
6. Stewart, S.W.: Real-time detection and location of local seismic events in central. Bulletin of the Seismological Society of America 67 (1977)
7. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: a Review. ACM Computing Surveys 31(3), 264–323 (1999)
8. Giancarlo, R., Bosco, G.L., Pinello, L., Utro, F.: The Three Steps of Clustering in the Post-Genomic Era: A Synopsis. In: Rizzo, R., Lisboa, P.J.G. (eds.) CIBB 2010. LNCS, vol. 6685, pp. 13–30. Springer, Heidelberg (2011)
9. Giancarlo, R., Lo Bosco, G., Pinello, L., Utro, F.: A methodology to assess the intrinsic discriminative ability of a distance function and its interplay with clustering algorithms for microarray data analysis. BMC Bioinformatics 14(suppl. 1) (2013)
10. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley (1990)
11. Shamir, R., Sharan, R.: Algorithmic approaches to clustering gene expression data. In: Current Topics in Computational Biology, pp. 269–300 (2001)
12. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336), 846–850 (1971)
13. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification 2, 193–218 (1985)
14. Madariaga, R.: Dynamics of an expanding circular fault. Bulletin of the Seismological Society of America 66(3), 639–666 (1976)
15. Virieux, J.: P-SV wave propagation in heterogeneous media: Velocity-stress finite-difference method. Geophysics 51(4), 889–901 (1986)
16. Levander, A.R.: Fourth-order finite-difference p-sv seismograms. Geophysics 53(11), 1425–1436 (1988)
17. Larsen, S., Harris, D.: Seismic wave propagation through a low-velocity nuclear rubble zone. Technical report, Lawrence Livermore National Lab., CA (1993)
18. Aki, K., Richards, P.G.: Quantitative Seismology: Theory and Methods. Univ. Science Books (2002)