Automatic Seismic Signal Detection via Record Segmentation

Article *in* IEEE Transactions on Geoscience and Remote Sensing · July 2015

DOI: 10.1109/TGRS.2014.2386255

2 authors:

Vasilis Pikoulis
University of Patras
11 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Emmanouil Psarakis
University of Patras
39 PUBLICATIONS **323** CITATIONS
SEE PROFILE

Automatic Seismic Signal Detection via Record Segmentation

Erion-Vasilis Pikoulis and Emmanouil Z. Psarakis

Department of Computer Engineering and Informatics, University of Patras, 26500 Rio-Patras, Greece
email:{pikoulis, psarakis}@ceid.upatras.gr
phone: +30 2610 996969, fax: +30 2610 996971

***Abstract*—The automatic seismic signal detection constitutes a very interesting and challenging task. The main difficulty in solving this problem is attributed to the fact that both the statistical properties of seismic noise and the characteristics of the recorded events, are in general unknown. In this work, by exploiting the particular nature of the signals we are treating, as well as a number of very interesting properties possessed by exchangeable random variables, we treat the problem at hand as a record segmentation one, and propose the use of two functionally linked test statistics for its efficient and robust solution, in a two step procedure. By following this approach, we succeed not only in detecting a seismic wave arrival, but also in identifying the entire interval occupied by the signal, while minimizing the number of the required parameters. The performance of the proposed technique is confirmed by a series of experiments, both in synthetic and real seismic data sets.**

I. Introduction

The successful identification of the event related signals contained in continuously recorded seismic data constitutes the basic ingredient in achieving the ultimate goal of picking, that is the estimation of the arrival times of the seismic waves to the recording stations. As some of the most fundamental problems in Seismology, including event location, event identification, source mechanism analysis, relocation procedures and tomography, rely on travel - time inversion techniques, the reliability of the solutions depends heavily on the accuracy of the estimated arrival times of the waves to a network of seismic stations. Moreover, the aforementioned problems, as well as other ones related to modern applications, such as the monitoring of microseismicity induced by hydraulic fracturing [1], require the detection and analysis of a very large number of small magnitude events implying that the automatic detection technique needs to be both robust and computationally efficient.

The common approach followed in the so called off-line techniques to solve this problem, is to first detect the presence of the existing events and extract segments of the record containing one event each and then apply a picking method to each one in order to estimate the corresponding arrival times. Most well known picking methods that have been proposed, follow this approach [2]–[8]. As it is obvious, the effectiveness of this approach depends strongly on the ability of the segmentation method to obtain a proper splitting of the available record, which in turn is heavily affected by uncontrollable factors such as the magnitude and the duration of the events as well as their separation in time. This has led to a great variety of proposed solutions, dictated by the adopted assumptions on the noise and signal models.

The author in [9] proposes a Maximum Likelihood (ML) detector, by assuming Gaussianity for both signal and noise distributions. In [10], seismic measurements are assumed to consist of a sum of signals corrupted by additive Gaussian white noise, uncorrelated to the signals. Each signal is assumed to consist of a signal pulse, multiplied by a space-dependent amplitude function and with a space-dependent arrival time, estimated by using the ML method. The authors in [11] propose the detection of changes in AR modeled signals, by using the CUSUM test statistic [12]. In [13], both signal and noise are considered zero-mean Gaussian autoregressive (AR) processes. The detection problem is formulated in a hypothesis testing framework, where the used test statistic, given the above mentioned assumptions, follows asymptotically a χ^2 distribution under the null hypothesis. The authors in [14] define a characteristic function obtained by the three-component signal analysis, used for the detection of the events. Approximate confidence intervals are also derived under the Gaussianity assumption. In [15], signal detection is carried out by measuring the correlation between true waveforms, obtained by a network of seismic stations, and expected ones, derived synthetically or empirically. The authors in [16], [17] use a modification of the CUSUM statistic, proposed in [18] for the detection of multiple variance changes in time series and propose an iterative algorithm for the sequential detection of multiple seismic events and the estimation of their arrival times. In [19], ML is used for the detection of seismic events by analysing data recorded by an array of seismic stations. By adopting certain assumptions concerning the selected model parameters, the Short Time Fourier-transformed data are considered as independent, identically, complex normally distributed. Based on this assumption, the authors devise a hypothesis-testing based detection procedure, where they test sequentially the hypotheses of "exactly m events" against "at least $m+1$ events" being present in the time frame under consideration, for $m = 0, 1, \ldots, M_{max}$, where M_{max} is the maximum expected number of events. The authors indicate

The authors would like to thank the Seismological Laboratory of University of Patras, for their support in providing the experimental data set and for offering their expertise on several seismological issues, as well as the reviewers of this work, for their valuable feedback. This work was financed by the University of Patras, "Karatheodori" research program, entitled "The relocation problem of seismic event hypocenter parameters".

that the average power of the m-th signal must be strong enough compared to noise to be detected. In [20] a difference-based test statistic, calculated from the spectrogram of the record, is proposed for the detection of seismic phase arrivals. More recent approaches exploit the underlying P-wave velocity model for the enhancement of the original data and aim at the simultaneous detection and location of events [21]. Finally, a number of techniques that are focused on the detection of particular types of signals, such as the explosion-generated [22], or the low-frequency acoustic ones [23], have also been proposed.

In practice, the application of "statistical" detection methods is limited by the assumptions they impose on the characteristics of the background noise, which are usually too strict and rarely met in real noise conditions. It is a widely reported fact that the statistical properties of seismic noise are not easily predictable [23], [24]. Due mainly to this fact, the majority of the techniques used in real applications today are based on the ratio of a Short Term Average (STA) and a Long Term Average (LTA) of some characteristic function of the data [2], [24]–[29]. The simple intuition behind STA/LTA is that regardless of the particular noise model, in areas of noise, the value of the ratio should remain substantially constant, while when a signal occurs, the STA term should be able to capture the change much faster than the LTA, thus resulting in a rise of the ratio values. However, due to the lack of a statistical model in the general case, signal detection is based only on empirical findings, i.e. by comparing the ratio values to an empirically predetermined threshold value (trigger level). Moreover, as a ratio test statistic, STA/LTA is foremost a tool designed to reflect instances rather than intervals of change, leading to certain issues regarding its detection performance (e.g. in events with weak P-waves), as well as its ability to identify the time span of the detected signals [28]. For a solution of the latter problem especially, in practical applications, STA/LTA is equipped with additional parameters, such as the detrigger level, the pre-event time and the post-event time [30]. However, it has to be stressed that the above mentioned parameters are mostly signal-dependent, meaning that their (fixed) values used at the time of application, can only represent average guesses, thus introducing yet another source of uncertainty to the final solution. Some of the aforementioned issues are treated in more detail in Experiment IV, as well as in Section V, where a brief discussion concerning the relation between the STA/LTA-based technique and the proposed one, is held.

In the present work, by taking into account the fact that the ultimate goal of automatic procedures of this nature is to provide the recorded signals for subsequent storage and analysis, not merely indicate their presence in a given record, we treat the problem at hand as a record segmentation rather than a change detection one. In this context, the proposed approach determines the presence of a signal not only based on instances of abrupt change (e.g. the beginning of a seismic wave), but rather by assessing the total disturbance caused by the signal on the assumed features of the noise distribution. This enables it to detect the entire signal interval, rather than just parts of it, as it is the case with most conventional detection techniques. By exploiting the particularities of the signals we are treating, as well as a number of interesting properties satisfied by the employed test statistics under mild assumptions on the background noise process, we are able to present an efficient and robust automatic solution, minimizing at the same time, the number of algorithm parameters.

The remaining of this paper is organized as follows. In Section II the problem formulation is presented. In Section III, the proposed test statistics and their anticipated features in the different parts of a seismogram are considered in detail. Subsequently, by exploiting a number of useful properties of the proposed statistics, an efficient solution for the automatic segmentation of a given seismic record, is proposed. In Section IV the experimental results we obtained from the application of the proposed method on both synthetic and real seismograms are presented. In Section V, we briefly discuss several issues concerning the preprocessing of seismic records and give simple guidelines for the real-time implementation of the proposed segmentation technique. Section VI contains our conclusions. Finally, mathematical proofs and other technical details are summarized in Appendices A-C.

II. Problem Formulation

Let us denote with x_n, $n = 0, 1, \ldots, T-1$, the samples of a seismic record and let K be the unknown number of the recorded seismic events. Then, by denoting with s_n^k, $n = 0, 1, \ldots, T_k - 1$, the signal produced by the k-th event and with n_k the corresponding first arrival time, where $k = 1, \ldots, K$, x_n can be expressed as follows:

$$x_n = w_n + \sum_{k=1}^{K} s_{n-n_k}^k, \tag{1}$$

where w_n is a noise process. The problem at hand is then the identification of the signal intervals, i.e. the estimation of the unknown number of events K, as well as the boundaries of the signal intervals, given by the values of n_k, T_k, $k = 1, \ldots, K$. As we are going to see, the latter set of parameters will be implicitly inferred by the proposed segmentation technique, while K will be obtained as the solution of a well defined optimization problem.

It is important to notice that, as it is widely reported in the literature and already mentioned in the Introduction, the properties of the noise process w_n are not easily predictable, since the factors governing its behaviour remain in general unknown. As a consequence, any meaningful solution to the problem at hand, can only be based on relatively generic (i.e. safe) assumptions for w_n. An example of a real seismic record where, even elementary assumptions regarding the stationarity of w_n do not seem to hold, is shown in Fig. 7 (left, top).

A. Preliminaries

As it is already apparent from the above formulation, in this work rather than detecting instants of change in the given record, we are interested in identifying all the areas of the record, that are occupied by seismic signals. In this context, for every time point n spanned by the record, with $n = 0, 1, \ldots, T-1$, the goal of the proposed technique is

to obtain a classification of x_n into one of the two possible classes, namely "signal" and "noise". As we are going to see, this classification will be based on the values of suitably selected test statistics, that will enable the partition of the input record into areas of pure noise and areas that contain signals. To this end, let $\mathcal{T}$ denote the set of all time points of the record, i.e. let $\mathcal{T} = \{0, 1, \ldots, T-1\}$. Then, all $n \in \mathcal{T}$, for which, the value of the employed test statistic is calculated by using only noise samples, will comprise the subset of $\mathcal{T}$ denoted as $\mathcal{N}$. Similarly, the complement of $\mathcal{N}$ with respect to $\mathcal{T}$, denoted as $\mathcal{E}$, will contain all the time points for which, there is at least one signal sample included in the calculation of the test statistic. Note that, since $\mathcal{E} = \mathcal{T} \backslash \mathcal{N}$, where '\' denotes set subtraction, $\mathcal{N}$ and $\mathcal{E}$ form a partition of $\mathcal{T}$. Thus, in this sense, the goal of the proposed technique can be reformulated as obtaining a partition of set $\mathcal{T}$ into sets $\mathcal{N}$ and $\mathcal{E}$.

III. Motivation and Proposed Solution

The most common approach in solving problems such as the one formulated above, is to obtain a partition of the input data into "signal" and "noise" by comparing the values of some test statistic of the data, with a pre-selected threshold value η (in one-sided test formulations). Thus, instances of the input that yield test statistic values that exceed η are classified as "signal", while the rest, as "noise". As it is readily apparent, the main difficulty with this approach is determining a value for η that strikes the best balance between being "too low", leading to a high false alarms ratio, and "too high" resulting in a low detection ratio. It is well known that a rigorous formulation of the aforementioned problem, as well as its subsequent solution, can only be achieved in cases where the probability distribution of the test statistic under the noise and/or signal hypotheses are assumed as known. For these (and only these) cases, an optimal value of η can be obtained in a hypothesis testing framework, through the optimization of well defined criteria provided in the relative bibliography (e.g. the minimization of the total probability of error).

Unfortunately, as already mentioned, since the adoption of a specific probability distribution for the seismic noise process w_n in Equ. (1) seems practically infeasible, in the particular case of the problem at hand, there exists no clear way of expressing threshold selection as a function of the parameters of the problem. Consequently, in this case, η can only be selected in an empirical fashion, meaning that the actual impact of each selection to the overall performance of the technique remains basically unknown. It has to be mentioned though, that this is the path followed by the majority of the existing techniques for the problem at hand.

In this work, instead of basing the solution on the proper selection of a threshold value, we propose a two-step approach, that results ultimately in an efficient, automatic and robust detection technique. The main motivation for the proposed approach lies on the fact that the aforementioned issues related to the one-step thresholding approach, originate mainly from the requirement of a threshold that controls at the same time, the detection, as well as the false alarm probabilities. Instead, the proposed approach addresses the two detection related subproblems, in a sequential fashion, rather than a simultaneous one, by using the following two-step decision procedure:

S_1 : Detect signal intervals, without taking into account the false alarms.

S_2 : Separate "true" signal intervals from "false" noise intervals, from the ones detected in S_1.

The best way to envision the first of the above steps is to consider that we are trying to solve the detection problem for the ideal case of a noiseless record, i.e. without taking into account the noise process. The solution to this problem will be based on a simple thresholding procedure, by combining a suitable averaging test statistic with a (relatively) low value for η. Note that by ignoring (for the time being) the noise related false alarms, we are able to relax to a significant extent the requirements on the value of η. This turns threshold selection into a relatively straightforward task, that is nonetheless performed in an automatic as well as a dynamic fashion.

On the other hand, the main effort of the proposed procedure is concentrated in addressing the second of the above steps, namely removing the falsely detected intervals from the ones identified in S_1. As it is going to be explained in detail, the solution to this problem will be based on the properties of a difference-based test statistic, that is functionally linked to the one employed in step S_1, as well as the notion of exchangeable random variables.

As we are going to see, by following the proposed approach, we succeed in avoiding the aforementioned issues related to the strict hypothesis testing framework (i.e. the demand of exact knowledge over the distribution of the used statistic under the null and/or alternative hypotheses), while still proposing a well-defined, automatic detection technique.

A. S_1: Solving the detection subproblem

As already mentioned, the goal of this step is to address separately the detection problem, discarding for the time being the false alarms one. The solution to this problem will be based on the following test statistic, denoted as $L_n(M)$:

$$L_n(M) = \frac{1}{M} \sum_{m=n}^{n+M-1} x_m^2, \; n = 0, 1, \ldots, T-M, \quad (2)$$

where M is the selected window length[1] and y_n denotes some positive transformation of x_n. Although we do not place any assumption on the applied positive transformation, in this work we are implicitly referring to amplitude manipulations such as the absolute or the squared value of the initial signal. In any case, we consider y_n as an enhanced, positive version of x_n. Moreover, regarding the employed window length, values in the range of $0.5 - 4$ secs ($M = 50 - 400$ for the common sampling frequency of 100 Hz), are considered sufficient for the task at hand. Note that M should be smaller than the duration of an event and that in any case, $M \ll T$ should hold. Finally, in the subsequent analysis, it is assumed that the recorded signals are separated by an interval of at least M

[1]For the sake of simplicity, from now on, the dependence of the above defined quantity on the parameter M will be omitted.

noise samples. We are going to elaborate more on these topics in Section V.A.

As it will become shortly clear, the fact that L_n is defined as a forward operator (i.e. it uses only present and future samples), is necessary for its consistency with the definition of the test statistic proposed in the next subsection. This definition introduces a left shift in the output of L_n, which is however global and can be trivially compensated in a post-processing step (by introducing the proper delay). Finally, we should mention that this compensation has been performed in all the subsequent examples involving L_n.

Let us now discuss briefly the anticipated features of L_n under the signal and noise scenarios, namely for $n \in \mathcal{E}$, and $n \in \mathcal{N}$, respectively. In the former case, due to the positivity of y_n, as well as the smoothing effect of the averaging operator, L_n can be considered as an approximation of the signal envelope, giving roughly a positive outline of its amplitude. For $n \in \mathcal{N}$ on the other hand, assuming first order ergodicity for w_n, every value of L_n represents a different estimation of the same quantity, namely the stochastic mean of random variable (RV) y_n in noise. Thus, it is very reasonable to assume that in the noise parts of the record, L_n will vary around a constant level, which we will denote as $m_{L|\mathcal{N}}$. Note that this is a value that lies somewhere in the middle of the noise distribution of L_n, not towards its tail, a fact that fortifies to a great extent the following assumption: throughout the duration of each seismic signal, L_n will exceed $m_{L|\mathcal{N}}$ with an exceedingly high probability.

Let us now consider a thresholding scheme, where we identify all the intervals of the L_n sequence where L_n exceeds $m_{L|\mathcal{N}}$. By "identification" we mean that we discover all the intervals of the form $[n_1, n_2]$, with $n_2 > n_1$, such that $L_{n_1-1} \leq m_{L|\mathcal{N}}$, $L_{n_2+1} \leq m_{L|\mathcal{N}}$, and $L_n > m_{L|\mathcal{N}}$, for $n_1 \leq n \leq n_2$. Under this scenario, based on the previous analysis, we anticipate that K of the identified intervals will correspond to the actual intervals occupied by the recorded signals, while the rest will correspond to noise parts of the record, due to the random oscillations of L_n around $m_{L|\mathcal{N}}$, in noise. From now on, we will refer to the former group of intervals as the "true" or "signal" ones, and to the second group as the "false" or "noise" ones. In any case, by implementing the aforementioned thresholding scheme and by setting the threshold value η equal to $m_{L|\mathcal{N}}$, the task of step S_1 of the proposed technique, can be considered as completed.

The issue of course is that $m_{L|\mathcal{N}}$ is unknown and cannot be directly estimated by sample averages, since set $\mathcal{N}$ is unknown. However, for this task, the inherent sparsity of seismic records, as well as the robustness of the median as a statistic prove themselves very helpful. More specifically, let m_L denote the median of the obtained L_n sequence (over the whole record). Then, based on the two aforementioned features and despite the presence of signals in the record, we maintain that for our purpose, m_L constitutes a good estimator of the noise level of L_n, $m_{L|\mathcal{N}}$. This assumption is solidified by the fact that we are considering records of arbitrary length (the longer the record, the safer the sparsity assumption becomes). This bring us at the first of the two basic assumptions we made in this work in order to achieve our goal:

$$\text{Assumption } \mathcal{A}_1: \quad P\left(L_n <= m_L | n \in \mathcal{E}\right) \approx 0. \tag{3}$$

After the above presentation, we are now in position to reformulate the first step of the proposed technique in more specific terms, as follows:

S_1 : Given y_n, $n = 0, \ldots, T-1$, calculate its running average L_n, as defined in Equ. (2). Then use the median of L_n, m_L, as the threshold value η and identify the record intervals where L_n exceeds $\eta = m_L$.

Concluding this subsection, in Fig. 1 (top), we present an example illustrating the use of L_n, as well as that of m_L in the first step of the proposed technique, in a synthetic record with $K = 3$ signals. Note the signal intervals that are identified by the application of S_1, i.e. by thresholding L_n, at $\eta = m_L$.

B. Step S_2: Separating true from false

The outcome of the thresholding procedure described in S_1, is the identification of L intervals, K of which correspond to signals (unknown at this stage), while the rest correspond to noise, with $L \geq K$. The goal of the second step of the proposed technique is the separation of the former intervals from the latter, thus obtaining the overall solution to the problem at hand.

1) Outline: Let us denote as $\widehat{\mathcal{E}}(\eta)$ the set of all time points identified by thresholding L_n at level $\eta = m_L$, i.e. let $\widehat{\mathcal{E}}(\eta) = \{n \in \mathcal{T} | L_n > \eta\}$. Based on our previous analysis, $\widehat{\mathcal{E}}(\eta)$ can be described as the union of two sets of time points, namely the ones that belong to the K desired intervals and the ones that belong to the $L-K$ unwanted ones. By using assumption $\mathcal{A}_1$, we can consider that for $\eta = m_L$, the first of the above sets constitutes a good approximation of the true set time points spanned by signals, i.e. of the desired set $\mathcal{E}$ defined in Section II.A. The second set on the other hand, which we will denote as $\widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$, is formed by the union of the $L-K$ false intervals identified in S_1 and consists of all n such that $n \in \mathcal{N}$ and $L_n > \eta$, i.e.:

$$\widehat{\mathcal{N}}_{\mathcal{E}}(\eta) = \{n \in \mathcal{N} | L_n > \eta\}. \tag{4}$$

This set constitutes basically the unwanted contribution of noise to the outcome of S_1, that has to be discarded, in order to obtain $\mathcal{E}$ from $\widehat{\mathcal{E}}(\eta)$ (see also Fig. 1 (top)) .

To this end, we are going to work as follows. First, we will sort the L identified intervals with respect to an attribute that will be shortly introduced, ensuring that the K desired intervals are grouped in the beginning of the obtained sequence, followed by the noise ones. This yields the following sequence of intervals:

$$\widehat{\mathcal{E}}_1(\eta), \widehat{\mathcal{E}}_2(\eta), \ldots, \widehat{\mathcal{E}}_L(\eta), \tag{5}$$

with $\bigcup_{l=1}^{L} \widehat{\mathcal{E}}_l(\eta) = \widehat{\mathcal{E}}(\eta)$, such that:

$$\bigcup_{l=1}^{K} \widehat{\mathcal{E}}_l(\eta) = \mathcal{E}, \qquad \bigcup_{l=K+1}^{L} \widehat{\mathcal{E}}_l(\eta) = \widehat{\mathcal{N}}_{\mathcal{E}}(\eta). \tag{6}$$

Since, as it is obvious, the desired solution to the problem at hand is given by the first K members of the sequence in Equ.

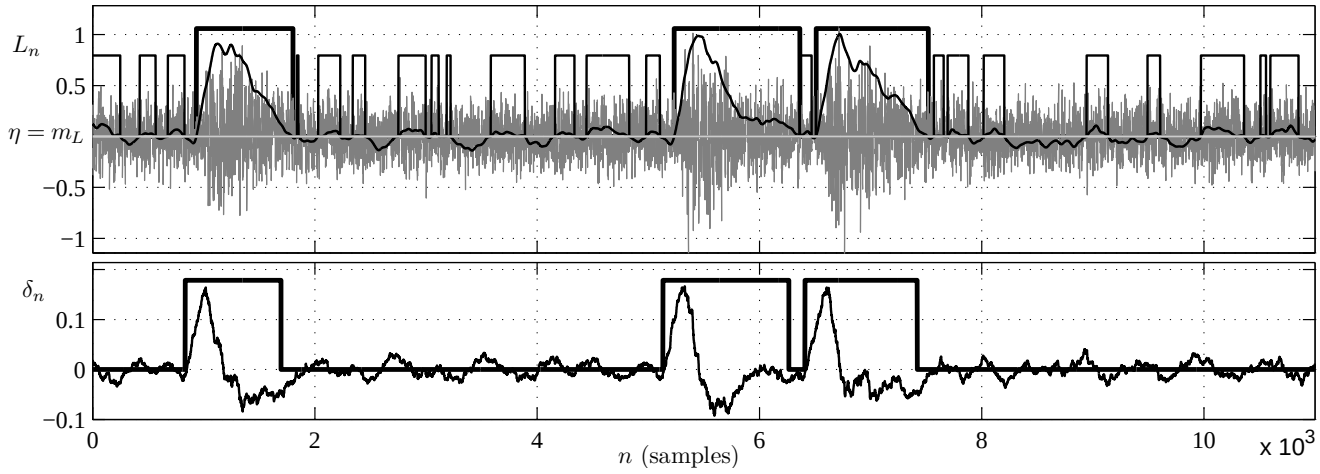


Fig. 1. Application of the proposed test statistics to a synthetic record (grey waveform, top). The obtained L_n (Equ. (2)) and δ_n (Equ. (7)) sequences are depicted by the black curves on the top and bottom plots, respectively. The rectangular lines indicate the intervals identified by thresholding L_n at level $\eta = m_L$ (light-grey horizontal line, top). The higher amplitude "pulses" indicate the signal (true) intervals (i.e., set $\mathcal{E}$), while the lower amplitude ones indicate the noise (false) intervals (i.e., $\widehat{\mathcal{E}}(\eta)$). Note that in this example, a window length of $M = 200$ samples was selected and that, the L_n sequence has been scaled, translated and delayed by $M/2 = 100$ samples, for the sake of better visibility.

(5), our initial problem is now reduced to the estimation of the unknown number of signals K. As we are going to see, this estimation will result via an iterative scheme where we remove sequentially all the intervals of the sorted sequence from the record, and estimate the point at which all that remains, can be characterised as pure noise.

2) A difference based test statistic: Having presented an outline of the main ideas employed in step S_2, we will now describe in detail each of the problems involved in this process, as well as the approach taken towards their solution. To this end, we begin by defining the following test statistic, denoted as δ_n, obtained as a sequence of differences between suitably spaced L_n values, i.e:

$$\delta_n = L_n - L_{n-M},\ n = M, M+1, \ldots, T-M, \qquad (7)$$

where L_n the averaging test statistic used in step S_1 (see Equ. (2)).

As it is clear from its definition, for every n, δ_n is the difference of the average of the input y_n, over two adjacent, non-overlapping windows of length M. The left window involves the M samples immediately preceding time point n (interval $[n-M, n-1]$), while the right one, the M samples beginning at this time point (interval $[n, n+M-1]$). Thus, in the sense that it is obtained as a manipulation of "pre-" and "post-" quantities, δ_n is very similar in nature with a great number of ratio-based test statistics used in change detection or hypothesis testing problems. The reason we use subtraction instead of division, is attributed to the fact that in this work, δ_n will be used for the identification of intervals rather than points of change. More specifically, we exploit the fact that the value of the difference is sensitive to the magnitude of the operands, while the ratio is not.

As it will become shortly apparent, the totally different characteristics of δ_n in the time point sets $\mathcal{E}$, $\mathcal{N}$, and $\widehat{\mathcal{E}}_{\mathcal{N}}(\eta)$, namely the signal, noise, and false alarms ones respectively, will become our main tools in achieving the goal set in step S_2 of the proposed technique. Let us now describe these interesting attributes of δ_n, beginning by the desired set $\mathcal{E}$.

a) $n \in \mathcal{E}$: In signal intervals, the values of the above mentioned statistic depend on uncontrollable factors such as the magnitude and the shape of the events and as such, cannot be assessed in a rigorous, mathematical framework. However, by taking into account the general characteristics of seismic signals, several generic features are to be anticipated. More specifically, the presence of the $k-$th signal in the record, is reflected in the values of δ_n by the presence of a narrow peak (of approximately $2M$ samples), at the onset of each signal, followed by a large interval (comparable to the duration of the signal), of values that are biased towards a negative level, due to the fading of the signal amplitude. Note also that depending on the magnitude of the change at n_k, as well as the fading rate of the signal, δ_n is expected to take extreme values (positive and negative, respectively) during the entire signal interval. This is helped by the fact that, as a difference based statistic, δ_n is sensitive to signal amplitude, which is a highly desirable feature for the detection problem at hand.

b) $n \in \mathcal{N}$: As mentioned previously, in noise intervals, L_n and L_{n-M} constitute two identical estimators of the same quantity (the mean of y_n), using noise samples from disjoint but adjacent windows. Consequently, it seems intuitively safe to assume that in pure noise, the statistical properties of their difference, namely δ_n, should be independent of their order in the subtraction operation. This can be equivalently expressed as demanding the distributions of δ_n and $-\delta_n$, for $n \in \mathcal{N}$, to coincide, leading to the following statement, which constitutes the second basic assumption made in this work:

Assumption $\mathcal{A}_2$: The pdf of δ_n in noise, $f_{\delta_n}(z|\mathcal{N})$, is even symmetric, that is $f_{\delta_n}(z|\mathcal{N}) = f_{\delta_n}(-z|\mathcal{N})$.

Note that, while under our perception of normal noise conditions, the above assumption seems very reasonable, from a mathematical correctness standpoint, it can hold only if a set

of conditions are imposed to the ingredients of δ_n, namely L_n and L_{n-M}, and consequently to the noise process w_n. A relatively mild condition ensuring the validity of $\mathcal{A}_2$ states that for $n \in \mathcal{N}$, RVs L_n and L_{n-M} must be exchangeable [31], that is, their joint pdf, $f_{L_n,L_{n-M}}(x,y)$ is bivariate symmetric, i.e.:

$$f_{L_n,L_{n-M}}(x,y) = f_{L_n,L_{n-M}}(y,x). \tag{8}$$

If the above condition holds, then the following proposition ensures the validity of $\mathcal{A}_2$:

Proposition 1: *The pdf of the difference of two positive, exchangeable RVs is even symmetric.*
Proof. For a proof of Proposition 1, see Appendix A.

A direct consequence of Proposition 1 is that in noise intervals, the median value of δ_n coincides with its mean value, or equivalently,

$$P\left(\delta_n > 0 | n \in \mathcal{N}\right) = 1/2. \tag{9}$$

We must stress at this point that the symmetry expressed by Equ. (8), is satisfied always under the assumption of an i.i.d noise process, while in the i.d. case, where this symmetry reduces to

$$f_{L_n,L_{n-M}}(x|y) = f_{L_n,L_{n-M}}(y|x), \tag{10}$$

there are well known families of bivariate pdfs [32], [33], [34] satisfying this property, which are used for the approximation of pdfs of RVs whose analytical form is unknown. Bivariate Normal and Gamma distributions with identical marginals (regardless of the amount of correlation), constitute principal examples of the aforementioned class of pdfs [32]. Although we are not able to prove, without imposing more assumptions on the noise process, that the joint pdf of the RVs defined in Equ. (7) possesses such a symmetry, we consider that in most cases it is true.

c) $n \in \widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$: By its definition in Equ. (4), $\widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$ is a specific subset of $\mathcal{N}$, formed by all $n \in \mathcal{N}$ that satisfy the additional condition $L_n > \eta$. We will now examine the way this condition affects the symmetry of the pdf of δ_n in noise, as expressed in assumption $\mathcal{A}_2$. More specifically, we will concern ourselves with the value of the probability of δ_n being positive, in the unconditional and conditional noise cases, i.e. for $n \in \mathcal{N}$ and $n \in \widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$, respectively. As we are going to see, the difference between these probabilities can be exploited for the discrimination between the "pure noise" set $\mathcal{N}$ and its subset $\widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$. This in turn will enable the identification of the desired set $\mathcal{E}$ from the identified one, $\widehat{\mathcal{E}}(\eta)$, which is the ultimate goal of the second step. In this sense, the greater this difference is, the easier the solution of the identification problem becomes.

As we have already seen in Equ. (9), the first of the two aforementioned probabilities, equals $1/2$. We will now calculate the second one, which we will simply denote as $p(\eta)$, i.e.:

$$\begin{aligned} p(\eta) &= P(\delta_n > 0 | n \in \widehat{\mathcal{N}}_{\mathcal{E}}(\eta)) \\ &= P(\delta_n > 0 | n \in \mathcal{N}, L_n > \eta). \end{aligned} \tag{11}$$

To this end, the following proposition, stating an interesting statistical property of the difference of two positive exchangeable RVs, is of crucial importance.

Proposition 2 : Let $\mathcal{X}$, $\mathcal{Y}$ be two positive exchangeable RVs, with $f(x)$ denoting their common pdf and let $\mathcal{Z} = \mathcal{X} - \mathcal{Y}$. Let also $d_{\mathcal{Z}}(\eta)$ denote the following conditional probability:

$$d_{\mathcal{Z}}(\eta) = P\left(\mathcal{Z} > 0 | \mathcal{X} > \eta\right), \tag{12}$$

where η denotes a threshold value. Then,

$$d_{\mathcal{Z}}(\eta) = 1 - \frac{1}{2} P(\mathcal{Y} > \eta | \mathcal{X} > \eta). \tag{13}$$

Proof. For a proof of Proposition 2, see Appendix B.

By combining Equs. (11)-(13), $p(\eta)$ can be written as:

$$\begin{aligned} p(\eta) &\equiv d_{\delta_n}(\eta | \mathcal{N}) \\ &= 1 - \frac{1}{2} P(L_{n-M} > \eta | L_n > \eta, n \in \mathcal{N}). \end{aligned} \tag{14}$$

Subsequently, as can be seen from Equs. (9) and (14), while the probability of having a positive value for δ_n, given $n \in \mathcal{N}$, equals $1/2$, the same probability conditioned on $n \in \widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$, is always greater than $1/2$, by an amount that depends on the value of η, as well as the joint pdf of L_n and L_{n-M} in noise.

Although further conclusions are not possible without additional assumptions on this joint pdf, it is seems very useful to investigate further the case where w_n is an i.i.d. noise process. In this particular case, RVs L_n and L_{n-M} are also i.i.d. given $n \in \mathcal{N}$, and the probability on the right hand side of Equ. (14) can be further simplified, as follows:

$$\begin{aligned} &P(L_{n-M} > \eta | L_n > \eta, n \in \mathcal{N}) \\ &= \frac{P(L_{n-M} > \eta, L_n > \eta, n \in \mathcal{N})}{P(L_n > \eta, n \in \mathcal{N})} \\ &= \frac{P(L_{n-M} > \eta, L_n > \eta | n \in \mathcal{N}) P(n \in \mathcal{N})}{P(L_n > \eta | n \in \mathcal{N}) P(n \in \mathcal{N})} \\ &= \frac{P(L_{n-M} > \eta | n \in \mathcal{N}) P(L_n > \eta | n \in \mathcal{N})}{P(L_n > \eta | n \in \mathcal{N})} \\ &= P(L_{n-M} > \eta | n \in \mathcal{N}) \\ &= P(L_n > \eta | n \in \mathcal{N}). \end{aligned} \tag{15}$$

Consequently, using Equs. (14), (15), in the i.i.d. noise case $p(\eta)$ simply reduces to

$$p(\eta) = 1 - \frac{1}{2} P(L_n > \eta | n \in \mathcal{N}), \tag{16}$$

meaning that under this particular noise model, $p(\eta)$ is an increasing function of the selected threshold η. Although this statement cannot be proven in the general (exchangeable) noise case, it serves as a guideline for the impact of the selected threshold on the value of $p(\eta)$ and its expected distance from $1/2$. Moreover, in Appendix C, we present a simple procedure for estimating $p(\eta)$ from record sample averages (see Equ. (49)), for noise cases beyond the i.i.d scenario. By using this estimation procedure, we have concluded that $p(\eta)$ constitutes a very robust feature of δ_n, with limited sensitivity to the selected noise model, as discussed in the next paragraph and shown in our experiments section.

d) The selection of $\eta = m_L$: In the paragraph following assumption $\mathcal{A}_1$, we argued that the selection of $\eta = m_L$ already meets the requirement set by the first step of the proposed technique, namely the correct detection of the signal intervals. In this paragraph we maintain that this particular selection leads to values of $p(\eta)$ that depart significantly form $1/2$, thus enabling the solution of the false alarms problem as well.

To this end, let it be reminded that m_L equals the median of the L_n sequence, which, under the sparsity assumption for the record, provides a good estimation of the theoretical median of L_n in noise. For the moment, let us assume that m_L is exactly equal to this theoretical value. Then, under the i.i.d. noise scenario, the theoretical value of $p(\eta)$, is exactly $3/4$, as can be easily seen from Equ. (16) (in this case, the probability on the r.h.s. of (16) equals $1/2$). Notice now that, due to the one-sided contribution of "higher than noise" L_n values in the signal intervals of the record, in reality, m_L is expected to be an overestimation of its theoretical counterpart. Then, since in the i.i.d. noise case, $p(\eta)$ is a increasing function of η, $p(m_L)$ is expected to be even higher than $3/4$, i.e. sufficiently higher that $1/2$ for the task at hand.

Although these statements hold only for the i.i.d. noise model and can not be proven in the general case, based on a number of experiments on a great variety of noise models, we can safely state that $p(\eta)$ constitutes a very robust feature of the used test statistic. This fact is clearly demonstrated by Experiment I in Section IV, where it is shown that, even in noise cases that depart greatly form the ideal i.i.d. process, $p(m_L)$ remains constantly above the 0.75 level, i.e. close to its ideal value. Finally, we should mention that, even in critical noise conditions where the value of $p(m_L)$ could prove to be lower than expected, the estimation procedure presented in Appendix C (see Equ. (49)), provides a very useful tool for the selection of a suitable value for η. More specifically, we recommend a search over an interval of values of η that lie within the noise level of L_n, and select the one that yields the highest possible $p(\eta)$, estimated by the aforementioned procedure. It has to be stressed though, that such an extreme noise case was not encountered in our experiments on real seismic records, where we always used the $\eta = m_L$ as the selected threshold in step S_1 of the proposed technique.

All the above mentioned features of the proposed test statistic, are apparently depicted in the example of Fig. 1 (the δ_n sequence is shown in the bottom plot). By exploiting these attributes of δ_n, we are now in place to address the two subproblems involved in the second step of the proposed technique. This will be the topic of the next two subsections.

S_{2a}: Sorting the identified intervals: As already mentioned, the first subtask of step S_2 is obtaining a suitable ordering of the L intervals identified in step S_1, so that the resulting sorted sequence of Equ. (5) meets the condition set by Equ. (6). Based on the aforementioned attributes of δ_n, the solution to this problem is fairly straightforward. More specifically, we have already seen that, being a detection tool, δ_n attains its most extreme values in signal intervals. Thus it seems reasonable to assume that if we sort the L obtained intervals based on the variance (equivalently, energy) of δ_n in each of them, in descending order, then the K signal intervals will be placed first in the ordered sequence. In other words, our goal will be met. Note that, while there are many other criteria that can be used for this task (e.g., the duration of the intervals, the energy of L_n or the signal itself, to name a few), we found that the use of δ_n robustifies this procedure to a great extent.

S_{2b}: Estimation of K: After the completion of the previous task, we have obtained a sorted sequence of L intervals, $\widehat{\mathcal{E}}_l(\eta),\ l = 1, \ldots, L$, ensuring that the K desired signal intervals are positioned in the beginning of this sequence. The only problem that remains unsolved at this stage, is that K is still unknown. Thus, the estimation of this number will be our next (and final) task.

In order to achieve this goal, we are going to exploit the properties of δ_n in the different parts of the record, as they are defined by the above mentioned sequence of intervals (see also Equ. (6)). Specifically, the time points falling into the first K intervals of the sequence, define the set of true (or correct) detections $\mathcal{E}$. This is the part of the record where the values of δ_n are both extreme and irregular. The complement of this set with respect to set $\mathcal{T}$ (all of the time points of the record), is set $\mathcal{N}$, where the δ_n values are calculated using only noise samples. The pdf of δ_n in this set is assumed even symmetric. Finally, the time points falling into the last $L-K$ intervals of the sorted sequence, comprise the subset $\widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$ of $\mathcal{N}$, that contains all the unwanted false alarms. The values of δ_n in this set are highly non-symmetrical, since for $n \in \widehat{\mathcal{N}}_{\mathcal{E}}(\eta)$, the probability of δ_n being positive is much greater than the probability of it being negative.

Our approach for solving the task at hand is based on an iterative scheme where we discard the values of δ_n that correspond to the intervals of the sorted sequence, one interval at a time, and assess the statistical properties of the remaining set. This procedure defines the following sets of δ_n values:

$$\Delta_l = \{\delta_n | n \in \widetilde{\mathcal{N}}_l\},\ l = 0, \ldots, L, \tag{17}$$

where $\widetilde{\mathcal{N}}_0 = \mathcal{T}$ and $\widetilde{\mathcal{N}}_l = \widetilde{\mathcal{N}}_{l-1} \backslash \widehat{\mathcal{E}}_l,\ l = 1, \ldots, L$. Note that in order to improve readability, we have omitted the dependence of the above quantities on the threshold value η from their notation.

Based on the previous analysis, we know that sets $\Delta_l,\ l = 0, \ldots, K-1$, will contain extreme (as well as highly irregular) δ_n values, caused by the presence of the K signals in the record. These signal induced extremities (or "outliers") will have been removed by the next iteration of the algorithm, yielding set Δ_K. After this point, in the next $L-K$ iterations, the discarded values of δ_n will correspond to the false alarms intervals, meaning that they will be positive in their vast majority. As a result, sets $\Delta_l,\ l = K+1, \ldots, L$, will increasingly exhibit a negative bias in their values. Consequently, of all the δ_n sets defined in Equ. (17), Δ_K should be the closest representative of a random population with pdf $f_{\delta_n}(z|\mathcal{N})$. Thus, this will become our strategy for estimating K: find the member of the sequence $\Delta_l,\ l = 0, \ldots, L$ of sets, that best reflects the properties of δ_n in noise, namely the absence of extreme values and the even symmetry of the corresponding pdf.

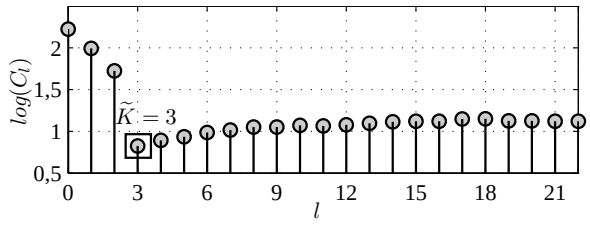


Fig. 2. The C_l sequence which is obtained from the application of S_2 to the synthetic record of Fig. 1.

In order to assess these properties, we define the following Cost Function (CF):

$$C_l = v_l \mathcal{D}_l, \tag{18}$$

where

$$v_l = \mathrm{E}[\delta_n^2 | \widetilde{\mathcal{N}}_l], \tag{19}$$

is the variance of δ_n conditioned on $n \in \widetilde{\mathcal{N}}_l$, and

$$\mathcal{D}_l = \sup_x \left| \int_0^x \left[f_{\delta_n}(z|\widetilde{\mathcal{N}}_l) - f_{\delta_n}(-z|\widetilde{\mathcal{N}}_l) \right] dz \right|, \tag{20}$$

quantifies the desired even symmetry of the underlying pdf.

Let us now give a more detailed description of the proposed procedure. In the l-th iteration of the algorithm, for $l = 0, \ldots, L$, we consider the values of set Δ_l as a sample drawn from a random population with pdf $f_{\delta_n}(z|n \in \widetilde{\mathcal{N}}_l)$, and obtain an estimation of the latter, by a detailed histogram of the sample. By using this estimation, we evaluate the above defined CF in order to assess the degree by which the estimated pdf exhibits the aforementioned characteristics, namely the lack of extreme values, measured by v_l, and the even symmetry imposed by Proposition 1, measured by $\mathcal{D}_l$.

Since, as it was mentioned, sets $\Delta_l,\ l = 0, \ldots, K-1$, will gradually contain fewer extreme values and will increasingly exhibit the aforementioned even symmetry, we expect a rapidly decreasing sequence of C_l values (dominated mostly by v_l), up to the K-th iteration. From this point on, sets $\Delta_l,\ l = K+1, \ldots, L$ will result to increasingly non-symmetric histograms, thus leading to an increasing sequence of C_l values (dominated mostly by $\mathcal{D}_l$), for $l = K+1 \ldots, L$. As an example, the C_l sequence that is obtained by the application of S_2 to the synthetic record of Fig. 1, clearly exhibiting the aforementioned behavior, is depicted in Fig. 2.

Thus, we are now in place to define the desired estimation of K as follows:

$$\widetilde{K} = \arg\min_{0 \le l \le L} C_l. \tag{21}$$

Note that in the example shown in Figs. 1, 2, C_l attains its minimum for $l = 3$, which is the correct number of events in this case. The result of the segmentation is indicated in Fig. 1 (top), by the thick rectangular line.

Following the above presentation, the second step of the proposed technique can be now reformulated as follows:

S_{2a} : Sort the L intervals identified in S_1, in order to obtain the ordered sequence $\widehat{\mathcal{E}}_l(\eta),\ l = 1, \ldots, L$ defined in Equ. (5).

S_{2b} : By using the ordered sequence of intervals, form sets $\Delta_l,\ l = 0, \ldots, L$ of δ_n values, as defined in Equ. (17). For every member of the latter sequence, evaluate the CF defined in Equ. (18) and obtain an estimation of K, as given by Equ. (21). The solution to the problem at hand is then given by the first $\widetilde{K}$ intervals of the sequence obtained in S_{2a}.

Having completed the presentation of the proposed technique, in the next section we are going to apply it in a number of experiments.

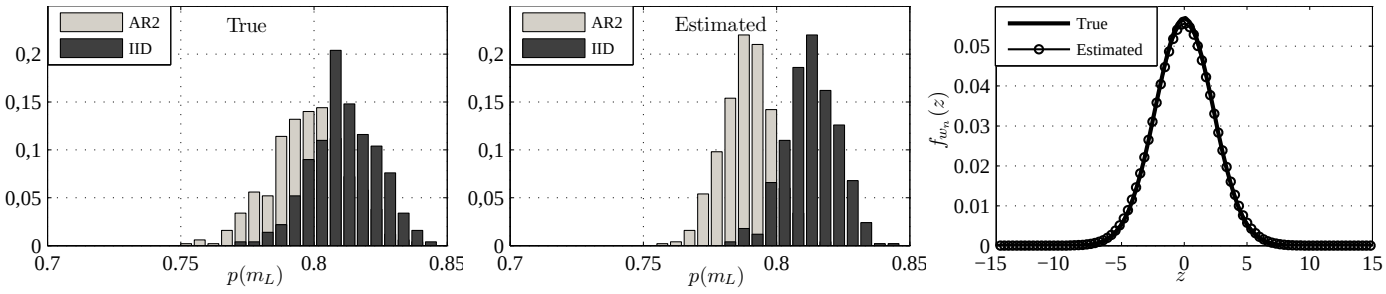
IV. Experimental Results

In this section we evaluate the performance of the proposed method by applying it in both synthetic as well as real data sets. In the simulations with synthetic data, four (4) noise processes were used. A white Gaussian process (denoted as IID in the following), an ARMA(2,2) one, as well as two (2) first order AR processes (denoted AR1 and AR2, respectively). In fact, the first and last noise models represent two extremes cases: the IID process plays the role of the "control sample", i.e. where all assumptions made are valid, while the AR2 one, due to the extreme magnitude of its pole (0.9), is used to test the performance of the method in cases where all assumptions fail. The details for the other noise processes, as well as the values concerning the other parameters of the technique, are summarized in Table I. We must stress at this point that the same experiments were also performed under a great number of other noise process scenarios and that the one presented here are selected as the most representative ones.

TABLE I
Experimental Set-up

w_n	**AR1**	**AR2**	**ARMA**	**IID**
Poles (mag.)	0.7	0.9	0.4472, 0.4472	–
Zeros (mag.)	–	–	0.5477, 0.5477	–
M	100			
T	30000			
SNR	$\mathrm{SNR}_{min} = -1$, $\mathrm{SNR}_{max} = 10$			
K	$K_{min} = 5$, $K_{min} = 15$			

In order to construct a data record, first the noise sequence w_n was created and then a value for the number of events K was randomly selected in a range $[K_{min}, K_{max}]$. Then, by selecting randomly K SNR values in the range $[\mathrm{SNR}_{min}, \mathrm{SNR}_{max}]$, as well as K onset times in the interval $[1, T]$, the resulting "recorded" signal x_n was calculated, by using Equ. (1). The synthetic signals s_n^k were created as follows. We started by low-pass filtering a window of white Gaussian noise. The resulting samples were shaped (i.e. multiplied) by using the corresponding values of a half-Gaussian window (of the same length) in order to model the exponential decay of the signal amplitude. Then the whole array was multiplied by constant gain, controlling the Signal to Noise Ratio (SNR). Note that for the estimation of the SNR, the signal power is calculated in a small window (100 samples) in the beginning of each signal. Due to the fading of the signal amplitude, this approach results in an overestimation of the mean SNR of the (whole) signal, but it represents a better estimation of the local SNR at the onset time, which is of more interest for the application at hand.



(a) (b) (c)

Fig. 3. (a)-(b): Histograms of true and estimated $p(m_L)$ values, respectively, for the IID and AR2 noise cases in Experiment I. (c): True and estimated noise histograms for the AR2 noise process case in Experiment II.

Experiment I: $p(\eta)$ and its estimation

The goal of this experiment is twofold. The first one is to asses the sensitivity of the probability $p(\eta)$ defined in Equ. (14), with respect to the assumption made for its derivation, namely the exchangeability property of the noise process. The second one is concerned with the performance evaluation of the $p(\eta)$ estimator defined by Equ. (49). For every synthetic record examined, we set $\eta = m_L$ (the median of $L_n, n = M, \ldots, T - M$) and estimated the value of $p(m_L)$, first in a straightforward way, based in the "noise only" w_n sequences (which were known in this case), and then by using Equ. (49), having the x_n sequences as input, therefore simulating a real-world scenario where the set $\mathcal{N}$ is unknown. These results are depicted as "true" and "estimated", respectively in Fig. 3 (a)-(b), where, for the sake of visibility, only the outcome for the two extreme scenarios, i.e. the IID and the AR2 noise processes, are shown.

As we can see in Fig. 3.(a), the insignificant drop in the $p(\eta)$ values despite the extreme correlation imposed by the AR2 noise model, reassures our belief that $p(\eta)$ constitutes a robust feature of the used test statistics. Moreover, the significant departure of $p(m_L)$ from 0.5, even in very unfavourable noise conditions, confirms our selection of m_L as the value of η in step S_1. Finally, even in critical noise conditions, where $p(m_L)$ could prove lower than expected, the approximation defined in Equ. (49) constitutes a very satisfactory estimation of $p(\eta)$ for our purposes, (as demonstrated in Fig. 3.(b)), thus providing automatic guidelines for the determination of η, in extreme cases.

Experiment II: Segmentation quality

In this experiment, we evaluate the quality of the segmentation results obtained by the thresholding and ordering scheme adopted in step S_1 of the proposed technique. To this end, we provided the true K in step S_2 (instead of relying on its estimation by Equ. (21)) and compare the average histograms of the true and estimated sets of noise samples, i.e. sets $\mathcal{N}$ and $\widetilde{\mathcal{N}}_K$, (see Equ. (17)), respectively.

The experimental set-up followed was identical to that of Experiment I and the results obtained under the worst case noise scenario AR2 model), are shown in Fig. 3.(c). As we can see, the almost perfect matching of the two histograms, constitutes a clear confirmation of the followed approach, as well as the used test statistic (i.e. L_n) for the segmentation needs of the problem at hand.

Experiment III: Detection performance

In this experiment, we evaluate the detection performance of the proposed technique with respect to the quality of the recorded signals (i.e the SNR value) under the different noise scenarios. To this end, for every SNR value and for every noise process considered, a separate (single SNR) data set of records was created. The detection curves depicted in Fig. 4.(a) constitute a clear demonstration of the high performance and the robustness of the proposed method, resulting in acceptable percentages even in the extremely unfavourable cases of negative SNRs (taking also into account the noise process used). For the moderate noise scenarios this percentage reaches values above 90% even for very low SNR values. Of high importance is also the fact that the number of false alarms, was very low in this experiment as it is clearly depicted in Fig. 4.(b).

Experiment IV: Comparison against $F-$test (STA/LTA)

For the special case of the white Gaussian noise process, we compare the performance of the proposed technique against the well known $F-$test [35], and the reason for this is twofold. On the one hand, under the aforementioned ideal noise conditions, the $F-$test constitutes a popular and widely used performance benchmark with mathematically tractable (and thus objectively evaluated) properties. On the other hand, since the lack of an objective threshold selection procedure for the STA/LTA technique, renders the direct comparison against the proposed one, in the general case, practically infeasible, by considering the aforementioned idealized noise case, the STA/LTA-based detection can be formulated as an F-test, meaning that the results obtained by the latter test act also as a theoretical performance ceiling for STA/LTA. In this experiment, for each of the considered SNR values, we fixed the probability of false alarm for the STA/LTA (F-test) technique to the value returned by the application of the proposed one for the IID data set (see Fig. 4.(b)), and subsequently compared the achieved probability of detection of the two methods. The STA and LTA durations were set to 200 (equal to M for the proposed technique) and 1000 samples respectively and the obtained results are depicted in Fig. 4.(c).

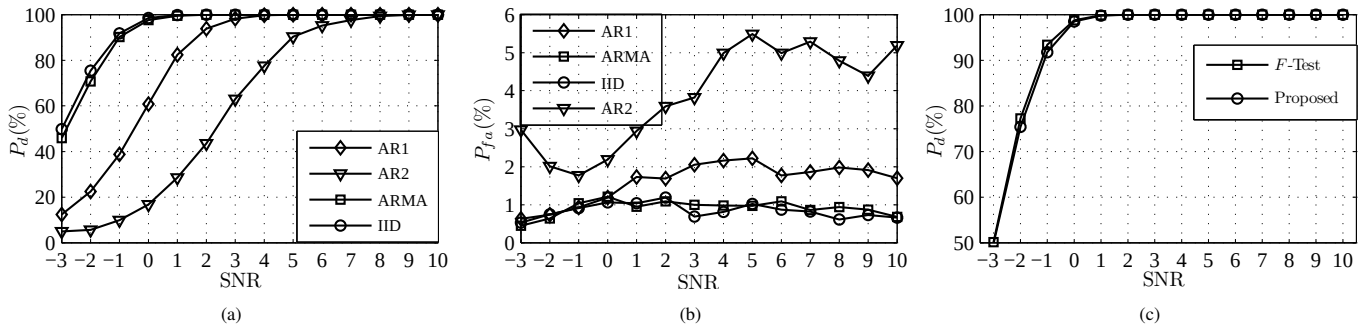


Fig. 4. (a)-(b): Detection (P_D) and false alarms (P_{FA}) ratios, respectively, as a function of SNR, in the synthetic data set of Experiment III. (c): Comparison of the detection ratios achieved by the proposed and the STA/LTA-based segmentation techniques, under the white Gaussian noise scenario, in Experiment IV.

Note that the proposed technique is able to effectively match the performance of the F-test, despite the fact that the latter method exploits in an optimal fashion the knowledge of the underlying noise process for the maximization of its detection rate, while the proposed one was blind to this exact knowledge, exploiting only the exchangeability property of the process.

To see this result from a different perspective, even in a case that is much more favorable to the STA/LTA-based detection than to the proposed approach, leading to the formulation of the former problem as an F-test, as well as to its optimal solution, the performance of the two rivals seems virtually identical. More importantly, while the F-test performance can be reached by the STA/LTA technique, only in the particular case of white Gaussian noise, the proposed one can maintain effectively the same performance (at least from a theoretical standpoint) in a vastly broader family of noise models, namely the ones ensuring the exchangeability property.

Experiment V: Application on real data set and comparison to STA/LTA

In this last experiment we evaluate the performance of the proposed method by applying it on real seismic data and compare it to the STA/LTA technique. The real data set, consisted of 120 pre-cut, 5 min records (T=30000 samples) of microseismic activity. The "true" number of events contained in the above mentioned records, counted by a human analyst, were approximately 1900. It should also be noticed that all waveforms were pre-filtered with the derivative-based filter defined in the next section. Note that in both cases, a detection was declared only in cases where at least 0.5 sec (or 50 samples) of the signal was discovered.

By using a window length of 1 sec (i.e. $M = 100$), the proposed detector was applied to the aforementioned data set and succeeded in identifying approximately 1650 events. This results in a detection rate of approximately 87%, while at the same time, the probability of false alarm was measured at 2.6%. The values used for the STA and LTA window parameters were 1 and 10 secs (or 100 and 1000 samples), respectively. We found that this combination led to the best balance concerning the sensitivity of the ratio. Then we tried to replicate the detection ratio achieved by the proposed technique in order to compare the corresponding false alarms. Specifically, the threshold value that yielded a detection ratio of 87% resulted also in a false alarms of approximately 9%, i.e. considerably higher than the one achieved by the proposed technique. By raising the selected threshold in order to lower the false alarms to a level of 2.6%, the detection ratio was also lowered to approximately 78%. An instance of Experiment V, justifying these findings is shown in Fig 5. The two aforementioned STA/LTA thresholds are shown in the middle and bottom plots of Fig 5, denoted as τ_D and τ_{FA}, respectively.

Another aspect of the comparison, concerns the segmentation capabilities of the two rivals, i.e. the their ability to estimate the time interval spanned by the recorded signal. For the proposed technique, this task is performed automatically in the thresholding step of the algorithm, namely step S_1. The accuracy of the obtained results is attributed the the combination of two factors. The first one is the employment of the signal envelope (i.e. L_n) as the used test statistic, and the second, the selection of a very low threshold value ($\eta = m_L$). In contrast, the same task for the STA/LTA technique is based on a set secondary parameters (e.g. detrigger threshold, pre- and post- event times and others), in addition to the basic ones mentioned above. Their selection is a demanding and unsafe task, since most of these parameters depend on uncontrollable signal features that cannot be assumed a priori (e.g. signal amplitude and shape, S-P interval), and ideally should be set on a signal-by-signal basis, which is infeasible. These segmentation-oriented issues are illustrated in the following example involving a real seismic record containing two events, shown in the top section of Fig. 6. The mid section of this figure shows the result obtained by the proposed technique, for a window length of 2 sec ($M = 200$ samples). Finally, the bottom section of Fig. 6, the results obtained by the STA/LTA technique, for STA=2 sec and two different LTA selections, are shown. More specifically, a selection of LTA=10 sec (solid lines), results in a successful detection of both events, but leads also to a significant underestimation of their time span, especially for the first event. Note also that the very different durations of the events, translates in very different requirements for the pre- and post- event time parameters. Increasing the LTA duration to 40 sec (heavy, dashed lines), improves the identification of the first event to a certain degree, but renders the second one undetectable, since the ratio remains below 1 throughout its duration. We must finally stress

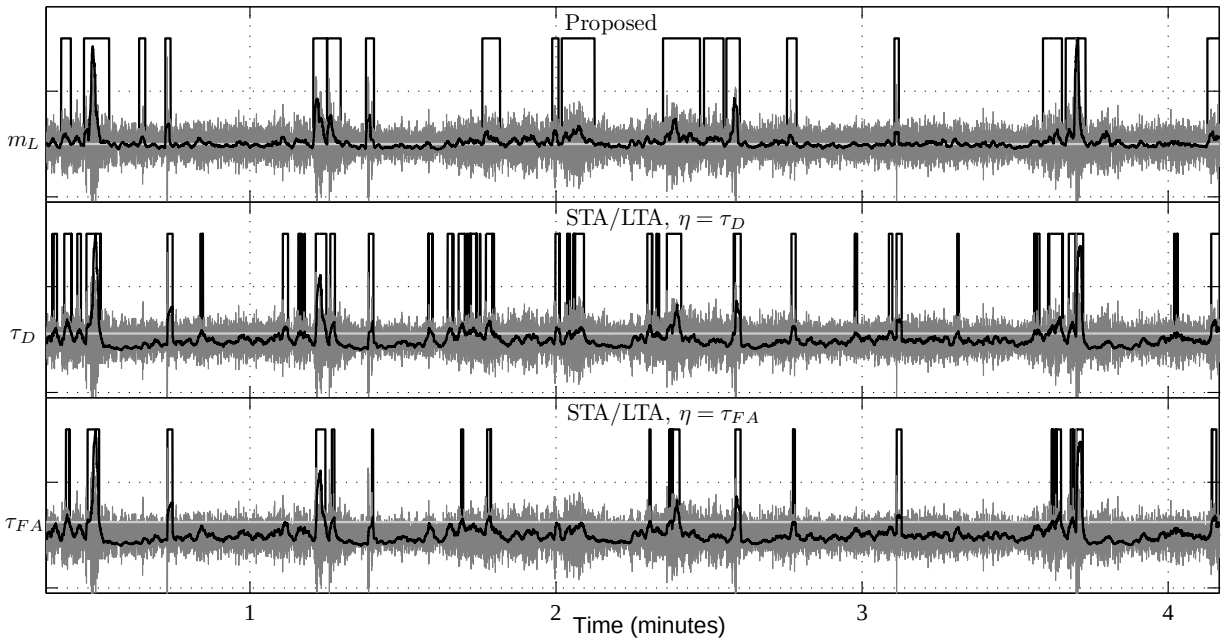


Fig. 5. An instance of Experiment V. The real record is shown in dark grey, the values of the test statistics are depicted by the black curves, the used thresholds are shown in light grey, while the detected intervals as rectangular pulses. Top: The result obtained by the proposed technique. Mid-Bottom: The results obtained by the STA/LTA technique for the two different thresholds used in the experiment. τ_D was able to yield the detection ratio of the proposed technique, but led to more false alarms. On the other hand, τ_{FA} lowered the false alarms to the same level with the proposed technique, but decreased also the detection ratio.

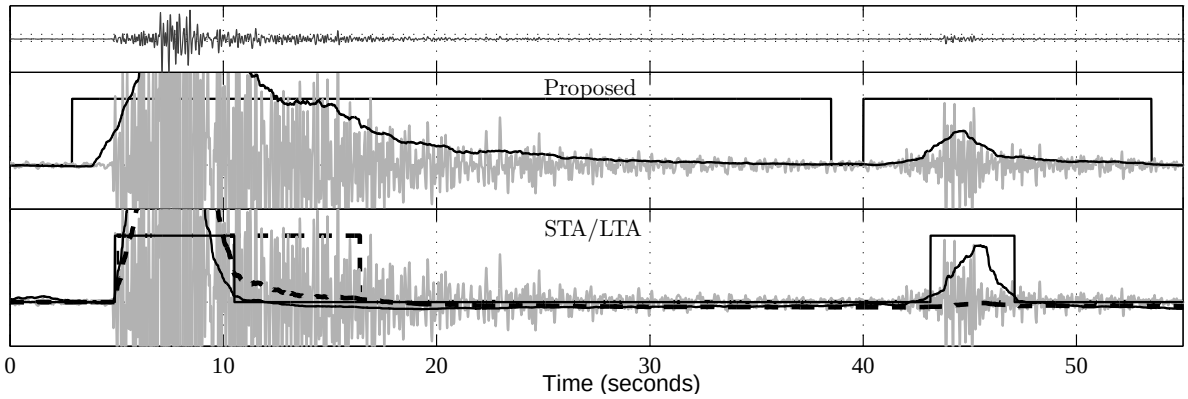


Fig. 6. A segmentation example. Top: A real seismic record containing two events. Middle: The results obtained by the proposed technique for $M = 200$ samples (2 sec). Bottom: The results obtained by the STA/LTA technique for STA=2 sec and LTA= 10 sec (solid lines), and 40 sec (dashed lines).

that, the selection of the trigger and detrigger thresholds for the STA/LTA technique were optimized for each particular case by hand and that all the depicted statistics were translated and normalized accordingly, in order to improve visibility. Finally, of special interest is the execution times of the two rivals in this particular experiment. First, it should be mentioned that the experiment was carried out in Matlab and that no particular care was given in the implementation of the algorithms towards speeding up their performance. The analysis of the whole data set using STA/LTA was completed in approximately 3 secs, while the proposed technique executed the same task in approximately 65 secs. As it is to be expected, due to step S_2, the proposed algorithm has an elevated complexity when compared to STA/LTA, which is executed in linear time. To be more specific, while step S_1 is executed in linear time, the most time consuming operation concerns the calculation of the histograms during the evaluation of C_l in step S_2. Although there are many ways to speed up the execution of S_2 (e.g. a more optimized implementation) the simplest strategy involves reducing the number of the executed iterations, which is equal to the number of intervals provided by S_1. A straightforward solution would be to discard the smallest of them, e.g. the ones containing 20 samples or less. Such short intervals comprise the great majority of the outcome of S_1, but their small duration classifies them with a very high probability as false alarms, meaning that their inclusion in S_2 is basically redundant. In any case, it has to be stressed that in absolute terms, the time requirements of the proposed technique remain relatively low, especially considering the processing power of modern computers.

In conclusion, as demonstrated through various examples and experiments, the technique proposed in this work offers a better detection performance, as well as drastically improved segmentation capabilities compared to STA/LTA, while at the same time reducing to a minimum the employed set of parameters. This improvement in performance and automation comes of course at the price of increased computational complexity, which however remains relatively low in absolute terms.

V. Discussion

A. Practical issues

As it has become apparent by now, the technique presented in this paper is designed as an automatic tool, with minimal calibration requirements regarding the window length parameter M, which is basically the only parameter of the technique, since the selection of the threshold η is performed automatically (see also the corresponding paragraph in Section III). A simple rule for the selection of M is that a large window results in more smoothing and can give better detection results in poor noise conditions, while a smaller one results in better resolution (regarding the time separation of the events). It has to be also stressed that, by design, the overall performance of the technique is relatively insensitive to moderate changes in the value of this parameter, thus providing safe margins for its selection. This comes as a great advantage over thresholding-based detection methods (such as the STA/LTA), where the final outcome depends directly on the threshold value, thus making them very sensitive to the selection of this parameter. Moreover, we must mention that since the proposed technique is based on the estimation of statistical quantities, a sufficient sample of the input has to be provided in order to obtain the best possible results out of it. This means that the duration T of the record has to be sufficiently large, e.g. in the order of several minutes.

Another important fact is that the accurate estimation of the signal intervals in the first step of the technique, as shown in Fig. 1 and confirmed by Experiment II, enables the application of post-detection rules, in a much more natural and controllable way, as the following example demonstrates: of the intervals identified in step S_1, a) keep the ones that are longer than T_{max}, even if not detected in S_2 (potentially missed events), and b) discard those shorter than T_{min}, even if detected in S_2 (potentially false alarms due to man made noise), where the duration thresholds T_{min}, T_{max} are detected by the application at hand.

Finally, as already mentioned, an assumption made in this work for achieving the desired segmentation solution, is that the recorded events are separated by at least one window length, or M samples. The values used for this parameter are small to moderate, i.e. 0,5-4 secs, with the most common being equal to 1 sec. There are however problematic cases, where this assumption fails, resulting in unification of events during the segmentation (thresholding) step. In these cases, the proposed technique could benefit either from additional enhancement of the original data in the preprocessing step, or by the introduction of some post-detection logic. Regarding the latter scenario, we propose the investigation of techniques similar to the hill-clustering one [36], for the identification of "hills" and "valleys" in the detected L_n intervals in order to determine whether an interval contains more than one events.

B. Real time application

In this subsection we provide some simple guidelines for the application of the proposed technique in a real-time fashion, by taking into account the requirement of the technique for a sufficiently long record, ensuring the stability of the involved estimations. More specifically, we propose the use of two buffers, namely a long (e.g. several minutes long) "noise" one and a much shorter (several seconds long), "record" one, as well as the execution of the following steps, each time the record buffer is filled:

1) Concatenate the contents of noise and record buffers.
2) Apply the technique to the resulting waveform.
3) Update the contents of the noise buffer using the newly identified noise samples.

With the above proposed procedure, the requirements of the technique are fulfilled, while at the same time the introduced delay (which is determined by the length of the record buffer), is kept at a minimum (in the order of several seconds).

C. Preprocessing issues

As we already mentioned in Section II, the superposition of various noise processes that manifest themselves in both the low- and high- ends of the spectrum, results in noise dominated records, rendering in many cases the discrimination between noise an signal virtually impossible. For the detection task at hand, the effect of the low-frequency seismic noise is especially destructive, since it invalidates the most elementary assumption regarding the statistical properties of the signal, namely its (first order) stationarity. An example of such a poor quality record (from the data set of Experiment V) is shown in Fig. 7 (left, top section).

The use of derivative-based filters in signal detection problems has been well documented in the relative literature (including the specific segmentation problem at hand [12]). These filters have the ability to remove the non-stationary component of the signal, while at the same time preserving abrupt changes, which is highly desirable in problems of this nature. Note however that in the case of discrete time signals, there are three possible approximations of the signal derivative, namely the forward, the backward, and the forward-backward first-order difference operators. Although the first two are the most commonly used in signal detection problems, in this work we propose the use of the third one, i.e.:

$$f(x_n) = \frac{1}{2}(x_n - x_{n-2}), \tag{22}$$

which from the perspective of our goal, presents itself with two great advantages. Firstly, the above defined operator has the effect of a bandpass filter, suppressing simultaneously the low and high frequency content of the record, which makes it highly suitable for the application at hand. In contrast, the other two alternatives result in filters that suppress only the

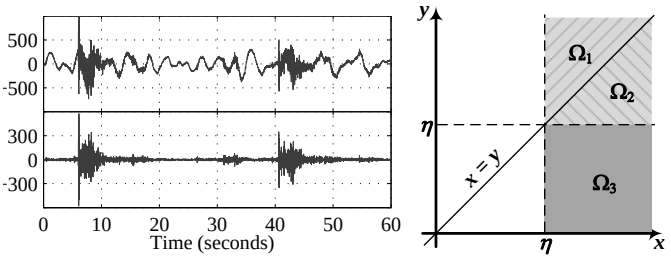


Fig. 7. Left: Example of the application of the proposed derivative-based filter. The recorded signal and its filtered version are shown on the top and bottom sections, respectively. Right: Partition of xy-plane

low frequency noise. Moreover, the filter defined in Equ. (22), constitutes the second scale high-pass filter of the Haar-based stationary wavelet transform, which was proposed in a number of works [8], [27], [28], for the preprocessing of the seismic signals in order to facilitate the solution of the P-phase picking problem.

Finally, it should be noted that in cases where additional information on the frequency content of the signals is available, then classical bandpass filtering, which is very common in seismic applications [17], can lead to better results regarding the denoising task. However, common type filters have the effect of introducing artefacts (or "smearing") around the abrupt changes in the record, which especially for the task at hand is considered most unwanted. In this sense, the proposed filter can be considered as a compromise between two worlds. leading to a better representation of the original signal, but with the smallest possible smoothing effect near the onset times of the seismic events. An example of the results obtained from the application of this filter is shown in Fig. 7 (left, bottom section).

VI. Conclusions

In this paper, the automatic seismic signal detection problem was examined from a record segmentation perspective. The use of a two-step procedure based on two different, but functionally linked test statistics, was proposed for the solution of the problem at hand. The main motivations behind the proposed technique were drawn from the particularities of seismic signals, on the one hand, and the notion of exchangeable random variables, as well as their properties, on the other. The advantages of the proposed approach were confirmed through a series of experiments, where synthetic and real seismic records were used.

Appendix A. Proof of Proposition 1

Let $\mathcal{X}$, $\mathcal{Y}$, be two RVs and let $f(x) = f_{\mathcal{X}}(x) = f_{\mathcal{Y}}(x)$, $f_{\mathcal{X},\mathcal{Y}}(x,y)$, denote their (common) marginal pdf, and their joint pdf, respectively. Let us also consider that the following relations hold:

$$f(x) = 0, \quad x < 0 \qquad (\textit{Positiveness}) \tag{23}$$

$$f_{\mathcal{X},\mathcal{Y}}(x,y) = f_{\mathcal{X},\mathcal{Y}}(y,x) \qquad (\textit{Exchangeability}). \tag{24}$$

Let us now define the following auxiliary RVs:

$$\mathcal{Z} = \mathcal{X} - \mathcal{Y} \tag{25}$$

$$\mathcal{Z}' = \mathcal{Y} - \mathcal{X}. \tag{26}$$

Then, the pdfs of $\mathcal{Z}$, $\mathcal{Z}'$, denoted by $f_{\mathcal{Z}}(z)$, $f_{\mathcal{Z}'}(z)$, respectively, coincide (see e.g. [35], pp. 185-186), i.e.:

$$f_{\mathcal{Z}'}(z) = f_{\mathcal{Z}}(z). \tag{27}$$

On the other hand, since $\mathcal{Z}' = -\mathcal{Z}$, their pdfs are also related as follows:

$$f_{\mathcal{Z}'}(z) = f_{\mathcal{Z}}(-z). \tag{28}$$

Thus, by using (27) and (28), the following must hold:

$$f_{\mathcal{Z}}(z) = f_{\mathcal{Z}}(-z), \tag{29}$$

which concludes the proof. ■

Appendix B. Proof of Proposition 2

Using the partition of xy-plane shown in Fig. 7 (right), the probability defined in Equ. (12), can be expressed as follows:

$$d_{\mathcal{Z}}(\eta) = P(\mathcal{Z} > 0 | \mathcal{X} > \eta) \tag{30}$$

$$= \frac{\sum_{i=2}^{3} \int_{\Omega_i} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy}{\sum_{i=1}^{3} \int_{\Omega_i} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy}. \tag{31}$$

Note however that, due to the exchangeability property of $\mathcal{X}$, $\mathcal{Y}$ (Equ. (24)), their joint pdf is symmetric with respect to the line $x = y$. Therefore, the following equality must hold (see Fig. 7 (right)):

$$\int_{\Omega_1} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy = \int_{\Omega_2} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy. \tag{32}$$

Thus, combining (31), (32), we obtain:

$$d_{\mathcal{Z}}(\eta) = \frac{\int_{\Omega_2} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy + \int_{\Omega_3} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy}{2\int_{\Omega_2} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy + \int_{\Omega_3} f_{\mathcal{X},\mathcal{Y}}(x,y)dxdy}$$

$$= 1 - \frac{1}{2}\frac{P(\mathcal{Y} > \eta, \mathcal{X} > \eta)}{P(\mathcal{X} > \eta)}, \tag{33}$$

from which, (13) easily follows. ■

Appendix C. Estimation of $p(\eta)$

Let us begin by obtaining an alternative expression for the conditional probability $p(\eta)$ defined in Equ. (14), which, as it will become shortly apparent, will facilitate our estimation procedure. More specifically, by noticing that events $\{L_n \leq \eta\}$ and $\{L_n > \eta\}$, form a partition of the sample space, and by using the sum rule of probability, $p(\eta)$ can be expressed as follows:

$$p(\eta) = 1 - \frac{1}{2}\frac{P(L_{n-M} > \eta, L_n > \eta, n \in \mathcal{N})}{P(L_n > \eta, n \in \mathcal{N})}$$

$$= 1 - \frac{1}{2}\frac{P(L_{n-M} > \eta, n \in \mathcal{N}) - P(L_{n-M} > \eta, L_n \leq \eta, n \in \mathcal{N})}{P(L_n > \eta, n \in \mathcal{N})}.$$

Then, by exploiting the exchangeability of L_n, L_{n-M} in noise, i.e. equality $P(L_{n-M} > \eta, n \in \mathcal{N}) = P(L_n > \eta, n \in \mathcal{N})$, and after some simple mathematical manipulations, the following expression can be obtained:

$$p(\eta) = \frac{1}{2} + \frac{1}{2}\frac{P(L_{n-M} > \eta | L_n \leq \eta, n \in \mathcal{N})}{\frac{1}{P(L_n \leq \eta | n \in \mathcal{N})} - 1}. \quad (34)$$

As it will be shown next, the probabilities involved in Equ. (34) can be approximated by sample averages, while the same is not true for the probabilities of Equ. (14). To this end, let us define the following sets of time points:

$$\mathcal{I}_{\mathcal{S}}^{X_n}(\eta^+) \equiv \{n \in \mathcal{S} : X_n > \eta\}, \quad (35)$$

$$\mathcal{I}_{\mathcal{S}}^{X_n}(\eta^-) \equiv \{n \in \mathcal{S} : X_n \leq \eta\}, \quad (36)$$

where $\mathcal{S} = \mathcal{T}, \mathcal{N}, \mathcal{E}$, and $X_n = L_n, L_{n-M}$, respectively. Note that in our case, only set $\mathcal{T}$ (i.e. the time points of the entire record) is completely known, meaning that the estimation can only be based on samples indexed in this set. Let us now state a number of useful properties satisfied by the sets defined in Equs. (35) and (36):

$$\mathcal{I}_{\mathcal{N}}^{X_n}(\cdot) \cup \mathcal{I}_{\mathcal{E}}^{X_n}(\cdot) = \mathcal{I}_{\mathcal{T}}^{X_n}(\cdot), \quad (37)$$

$$\mathcal{I}_{\mathcal{N}}^{X_n}(\cdot) \cap \mathcal{I}_{\mathcal{E}}^{Y_n}(\cdot) = \emptyset, \quad (38)$$

$$\mathcal{I}_{\mathcal{S}}^{X_n}(\eta^-) \cup \mathcal{I}_{\mathcal{S}}^{X_n}(\eta^+) = \mathcal{S}, \quad (39)$$

$$\mathcal{I}_{\mathcal{S}}^{X_n}(\eta^-) \cap \mathcal{I}_{\mathcal{S}}^{X_n}(\eta^+) = \emptyset, \quad (40)$$

where in Equ. (38), $Y_n = L_n, L_{n-M}$, and can be different from X_n. Let it now be reminded that, according to assumption $\mathcal{A}_1$ (Equ. (3)), for suitable (i.e. within the noise level of L_n) values of η, the following approximation can be obtained:

$$\mathcal{I}_{\mathcal{E}}^{L_n}(\eta^-) \approx \emptyset, \quad (41)$$

or equivalently, by using relation (37):

$$\mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-) \approx \mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-) \cup \mathcal{I}_{\mathcal{E}}^{L_n}(\eta^-) = \mathcal{I}_{\mathcal{T}}^{L_n}(\eta^-), \quad (42)$$

meaning that, for values of η within the range of interest, the unknown set $\mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)$ can be estimated by means of the known one $\mathcal{I}_{\mathcal{T}}^{L_n}(\eta^-)$. By using this approximation, the following relations can also be obtained (where $|\cdot|$ denotes set cardinality):

$$P(L_{n-M} > \eta | L_n \leq \eta, n \in \mathcal{N})$$

$$\approx \frac{|\mathcal{I}_{\mathcal{N}}^{L_{n-M}}(\eta^+) \cap \mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|}{|\mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|} \quad (43)$$

$$= \frac{|(\mathcal{I}_{\mathcal{N}}^{L_{n-M}}(\eta^+) \cap \mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)) \cup (\mathcal{I}_{\mathcal{E}}^{L_{n-M}}(\eta^+) \cap \mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-))|}{|\mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|} \quad (44)$$

$$= \frac{|(\mathcal{I}_{\mathcal{N}}^{L_{n-M}}(\eta^+) \cup \mathcal{I}_{\mathcal{E}}^{L_{n-M}}(\eta^+)) \cap \mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|}{|\mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|} \quad (45)$$

$$= \frac{|\mathcal{I}_{\mathcal{T}}^{L_{n-M}}(\eta^+) \cap \mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|}{|\mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|} \quad (46)$$

$$\approx \frac{|\mathcal{I}_{\mathcal{T}}^{L_{n-M}}(\eta^+) \cap \mathcal{I}_{\mathcal{T}}^{L_n}(\eta^-)|}{|\mathcal{I}_{\mathcal{T}}^{L_n}(\eta^-)|}, \quad (47)$$

where every quantity involved can be estimated from the record samples. In the above, note that (44) results from (43) by using relation (38), while the final expression in Equ. (47) is obtained by using the approximation in Equ. (42). Similarly, regarding the estimation of $P(L_n \leq \eta | n \in \mathcal{N})$, we have:

$$P(L_n \leq \eta | n \in \mathcal{N}) \approx \frac{|\mathcal{I}_{\mathcal{N}}^{L_n}(\eta^-)|}{|\mathcal{N}|} \approx \frac{|\mathcal{I}_{\mathcal{T}}^{L_n}(\eta^-)|}{|\mathcal{N}|}. \quad (48)$$

Thus, by using Equs. (34), (47) and (48), we maintain that, for values of η within the noise levels of the test statistic, the conditional probability $p(\eta)$ can be approximated as follows:

$$p(\eta) \approx \frac{1}{2}\left(1 + \frac{|\mathcal{I}_{\mathcal{T}}^{L_{n-M}}(\eta^+) \cap \mathcal{I}_{\mathcal{T}}^{L_n}(\eta^-)|}{|\mathcal{N}| - |\mathcal{I}_{\mathcal{T}}^{L_n}(\eta^-)|}\right), \quad (49)$$

which gives rise to a new problem, namely that of the estimation of the cardinality of the set of noise time points, $\mathcal{N}$. This estimation constitutes the topic of the following paragraph.

Estimation of $|\mathcal{N}|$

Let τ be a threshold value for which the approximation in (42) holds. Then, by using the properties expressed in Equs. (37) - (40), we can estimate the cardinality of $\mathcal{N}$, by using the following relation:

$$\begin{aligned} |\mathcal{N}| &= |\mathcal{I}_{\mathcal{N}}^{L_n}(\tau^-)| + |\mathcal{I}_{\mathcal{N}}^{L_n}(\tau^+)| \\ &\approx |\mathcal{I}_{\mathcal{T}}^{L_n}(\tau^-)| + |\mathcal{I}_{\mathcal{N}}^{L_n}(\tau^+)| \\ &= |\mathcal{I}_{\mathcal{T}}^{L_n}(\tau^-)| + \left(|\mathcal{I}_{\mathcal{T}}^{L_n}(\tau^+)| - |\mathcal{I}_{\mathcal{E}}^{L_n}(\tau^+)|\right). \end{aligned} \quad (50)$$

As we have already mentioned, while we are not in place to assume any prior knowledge regarding the presence of events in the record, we can safely infer that their existence, will manifest itself in set $\mathcal{I}_{\mathcal{E}}^{L_n}(\tau^+)$. Therefore, the estimation of the second term on the right-hand side of Equ. (50), is not feasible by simply considering sample averages of record values. Instead, a different approach is required that will be able to discard the contents of the unknown set $\mathcal{I}_{\mathcal{E}}^{L_n}(\tau^+)$ from $\mathcal{I}_{\mathcal{T}}^{L_n}(\tau^+)$ and therefore, provide an estimation of the desired set $\mathcal{I}_{\mathcal{N}}^{L_n}(\tau^+)$. To this end, let us assume that for a specific value of τ, $\mathcal{I}_{\mathcal{T}}^{L_n}(\tau^+)$ is composed of the following R intervals:

$$i_1(\tau^+), i_2(\tau^+), \ldots, i_R(\tau^+), \quad (51)$$

K of which, comprise the unwanted set $\mathcal{I}_{\mathcal{E}}^{L_n}(\tau^+)$ of signal intervals. Notice now that, although the latter intervals are unknown, due to sparsity and for a sufficiently low τ, their number (which is independent of the threshold value), can be considered much less than the number of noise intervals in the above defined sequence, i.e. $K \ll R$ can be assumed to hold. This condition will be satisfied with a very high probability, if τ lies in the middle of the noise distribution of L_n (i.e. in the vicinity of m_L), with the specific selection having no particular importance. In order to exploit this condition and achieve the desired estimation, we follow a bootstrapping-like approach [37] and form $Q \gg R$ interval sequences of length R, as follows:

$$i_{r_1(q)}(\tau^+), i_{r_2(q)}(\tau^+), \ldots, i_{r_R(q)}(\tau^+), \; q = 1, \ldots, Q, \quad (52)$$

where $r_m(q) \in \{1, 2, \ldots, R\}$, $m = 1, \ldots, R$, by choosing each time randomly and with replacement, R intervals from the sequence defined in (51). Then, since $K \ll R$ is assumed, the adopted randomization procedure has the effect of greatly reducing the probability of signal intervals existing in the newly formed (random) sequences. Moreover, by taking into account the fact that the signal intervals are expected to be much longer than the noise ones (see also the example on Fig. 1 (b)), we maintain that an estimation of the cardinality of the unknown set $\mathcal{I}_{\mathcal{N}}^{L_n}(\tau^+)$, can be obtained as follows:

$$|\widehat{\mathcal{I}}_{\mathcal{N}}^{L_n}(\tau^+)| \approx \min_q \sum_{m=1}^{R} |i_{r_m(q)}(\tau^+)|. \tag{53}$$

Using this estimation, as well as the relation of Equ. (50), we can now approximate the cardinality of $\mathcal{N}$, as follows:

$$|\mathcal{N}| \approx |\mathcal{I}_{\mathcal{T}}^{L_n}(\tau^-)| + |\widehat{\mathcal{I}}_{\mathcal{N}}^{L_n}(\tau^+)| \tag{54}$$

and this concludes our derivation.

REFERENCES

[1] Y. Vaezi and M. van der Baan, "Analysis of instrument self-noise and microseismic event detection using power spectral density estimates," *Geophys J. Int.*, 2014.

[2] R. Allen, "Automatic earthquake recognition and timing from single traces," *Bull. Seism. Soc. Am.*, vol. 68, pp. 1521–1532, 1978.

[3] M. Baer and U. Kradolfer, "An automatic phase picker for local and teleseismic events," *Bull. Seism. Soc. Am.*, vol. 77, pp. 1437–1445, 1987.

[4] T. Takanami and G. Kitagawa, "A new efficient procedure for the estimation of onset times of seismic waves," *J. Phys. Earth*, vol. 36, pp. 267–290, 1988.

[5] M. Leonard and B. Kennett, "Multi-component autoregressive techniques for the analysis of seismograms," *Phys. Earth Planet. Int.*, vol. 113, pp. 247–264, 1999.

[6] C. Saragiotis, L. Hadjileontiadis, and S. Panas, "Pai-s/k: A robust automatic seismic p phase arrival identification scheme," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, pp. 1395–1404, 2002.

[7] H. Zhang, C. Thurber, and C. Rowe, "Automatic p-wave arrival detection and picking with multiscale wavelet analysis for single-component recordings," *Bull. Seism. Soc. Am.*, vol. 93, pp. 1904–1912, 2003.

[8] J. Galiana-Merino, J. Rosa-Herranz, and S. Parolai, "Seismic p phase picking using a kurtosis-based criterion in the stationary wavelet domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, pp. 3815–3825, 2008.

[9] W. Freiberger, "An approximate method in signal detection," *Jour. Applied Math*, vol. 20, pp. 373–378, 1963.

[10] B. Ursin, "Seismic signal detection and parameter estimation," *Geophysical Prospecting*, vol. 27, pp. 1–15, 1979.

[11] I. Nikiforov and I. Tikhonov, "Application of change detection theory to seismic signal processing," *Detection of Abrupt Changes in Signals and Dynamical Systems*, pp. 355–373, 1986.

[12] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes - Theory and Application*. Prentice-Hall, Inc., 1993.

[13] V. Pisarenko, A. Kushnir, and I. Savin, "Statistical adaptive algorithms for estimation of onset moments of seismic phases," *Phys. Earth Planet. Int.*, vol. 47, pp. 4–10, 1987.

[14] R. Roberts, A. Christoffersson, and F. Cassidy, "Real time event detection, phase identification and source location estimation using single station three component seismic data," *Geophysical Jour.*, vol. 97, pp. 471–480, 1989.

[15] C. Young, J. Beiriger, M. Harris, and J. Trujillo, "Waveform correlation event detection," *A collaborative project effort between the Geophysics Group at New Mexico Tech, Sandia National Laboratories, and UC San Diego*, 1996.

[16] Z. Der, W. McGarvey, and R. H. Shumway, "Automatic interpretation of regional seismic signals using the cusum-sa algorithms," *21st Seismic Research Symposium*, pp. 393–403, 1999.

[17] Z. Der and R. Shumway, "Phase onset time estimation at regional distances using the cusum algorithm," *Phys. Earth Planet. Int.*, vol. 113, pp. 227–246, 1999.

[18] C. Inclan and G. Tiao, "Use of cumulative sums of squares for retrospective detection of changes of variance," *J. Amer. Statist. Assoc.*, vol. 89, pp. 913–923, 1994.

[19] P. Chung, M. Jost, and J. Boehme, "Estimation of seismic-wave parameters and signal detection using maximum-likelihood methods," *Computers & Geosciences*, vol. 27, pp. 147–156, 2001.

[20] S. J. Gibbons, F. Ringdal, and T. Kværna, "Detection and characterization of seismic phases using continuous spectral estimation on incoherent and partially coherent arrays," *Geophys J. Int.*, vol. 172, pp. 405–421, 2008.

[21] N. Langet, A. Maggi, A. Michelini, and F. Brenguier, "Continuous kurtosis-based migration for seismic event detection and location, with application to piton de la fournaise volcano, la reunion," *Bull. Seism. Soc. Am.*, vol. 104, pp. 229–246, 2014.

[22] S. R. Taylor, S. J. Arrowsmith, and D. N. Anderson, "Detection of short time transients from spectrograms using scan statistics," *Bull. Seism. Soc. Am.*, vol. 100, no. 5A, pp. 1940–1951, 2010.

[23] S. J. Arrowsmith, R. Whitaker, S. R. Taylor, R. Burlacu, B. Stump, M. Hedlin, G. Randall, C. Hayward, and D. ReVelle, "Regional monitoring of infrasound events using multiple arrays: application to utah and washington state," *Geophys. J. Int.*, vol. 175, pp. 291–300, 2008.

[24] M. Withers, R. Aster, C. Young, J. Beiriger, M. Harris, S. Moore, and J. Trujillo, "A comparison of selected trigger algorithms for automated global seismic phase and event detection," *Bull. Seism. Soc. Am.*, vol. 88, pp. 95–106, 1998.

[25] R. Allen, "Automatic phase pickers: their present and future prospect," *Bull. Seism. Soc. Am.*, vol. 68, pp. S225–S242, 1982.

[26] R. Sleeman and T. van Eck, "Robust automatic p-phase picking: an online implementation in the analysis of broadband seismogram recordings," *Phys. Earth Planet. Int.*, vol. 113, pp. 265–275, 1999.

[27] F. Botella, J. R. Herranz, J. Giner, S. Molina, and J. Galiana-Merino, "A real-time earthquake detector with prefiltering by wavelets," *Computers & Geosciences*, vol. 29, pp. 911–919, 2003.

[28] S. V. Baranov, "Application of the wavelet transform to automatic seismic signal detection," *Physics of the Solid Earth*, vol. 43, pp. 177–188, 2007.

[29] C. A. Rowe, R. J. Stead, M. L. Begnaud, and E. A. Morton, "Seismic signal analysis for event detection and categorization," *2012 Monitoring Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, pp. 312–320, 2012.

[30] A. Trnkoczy, "Understanding and parameter setting of sta/lta trigger algorithm," *In: Bormann, P. (Ed.), New Manual of Seismological Observatory Practice 2 (NMSOP-2), Potsdam: Deutsches GeoForschungsZentrum GFZ*, pp. 1–20, 2012.

[31] M. Hollander, "A nonparametric test for bivariate symmetry," *Biometrica*, vol. 58, pp. 203–212, 1971.

[32] N. Balakrishnan and C. D. Lai, *Continuous Bivariate Distributions*. Springer Verlag, 2009.

[33] N. Gumbel, "Bivariate exponential distributions," *J. Amer. Stat. Assoc.*, vol. 55, pp. 698–707, 1960.

[34] S. Nadarajah and A. Gupta, "Intensity-duration models based on bivariate gamma distribution," *Hiroshima Math. J.*, vol. 36, pp. 387–395, 2006.

[35] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes*. Mc Graw-Hill, 2002.

[36] D. M. Tsai and Y. H. Chen, "A fast histogram-clustering approach for multi-level thresholding," *Pattern Recognition Letters*, vol. 13, pp. 245–252, 1992.

[37] B. Efron and R. Tibshirani, *An introduction to the bootstrap*. Chapman & Hall, 1993.