A New Hypothesis Testing Based Technique for the Simultaneous Detection of Seismic Events

Erion-Vasilis M. 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*— One of the most difficult tasks in the solution of a hypothesis testing problem is the estimation of the probability density functions of the null and the alternative hypothesis. In this paper, the simultaneous detection of seismic events contained in a given data record is formulated as such a problem and an approximation of the probability density function under the null hypothesis, of a ratio based test statistic is proposed. By exploiting some interesting properties satisfied by the ratio of identical distributed Random Variables, as well as the sparsity of the seismic events in the data record, we succeed in obtaining such an approximation. From a series of experiments we have conducted in both synthetic and real seismic data, the effectiveness of the proposed technique is confirmed.**

I. Introduction

The determination of the arrival time of a seismic wave to a particular recording station is referred to as wave picking and automated procedures that address this problem as automatic pickers. 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 last two of the aforementioned problems require the detection and analysis of a very large number of small magnitude events implying that the automatic picking technique needs to be both robust, as microseisms produce signals with low SNR and computationally efficient. The problem becomes even more demanding if the case of multiple seismic events is considered, where both the detection of the events present in a given record, as well as the determination of the corresponding arrival times is required.

The solutions to the picking problem that have been proposed that can be roughly categorized in methods that are based on the ratio of a Short Term Average (STA) and a Long Term Average (LTA) of some Characteristic Function (CF) of the data [1], [2], [3], AR - modeling based methods [4], [5], [6], Discrete Wavelet Transform based methods [7], [8], Statistics based methods [9], [10], [11], [12], and combination of the above [13].

With the exception of the STA/LTA based methods and the method presented in [9], the above mentioned methods consider only the special case where the given record contains only one event, and the goal is the determination of the onset time of this event. The authors in [9] use a modification of the CUSUM algorithm, proposed in [14] 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. The authors stress the fact that the performance of the method is heavily influenced by the amplitude of the events and their ordering in the time axis. STA/LTA based methods on the other hand calculate the STA/LTA ratio at every point and by comparing the resulting value with an empirically predetermined threshold, decide whether an event is present in the examined area of the record, or not.

In this paper by exploiting the particular nature of the signals we are treating, and by using some interesting properties that obeys a ratio based test statistic as well as its ingredients, we propose a different approach, which is affected to a much lesser degree by the aforementioned factors, thus resulting to a more robust automatic simultaneous detection method.

The remaining of this paper is organized as follows. In Section II the problem formulation is presented and the behavior of a ratio test statistic in the different parts of a seismogram is considered in detail. In Section III the simultaneous detection of multiple seismic events problem is formulated as a hypothesis testing problem and by exploiting a useful property of the ratio statistic an efficient solution 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. Finally, Section V contains our conclusions.

II. Problem Formulation

Let us denote with x_n, $n = 0, 1, \cdots, T-1$, the record from a given station and let us also assume that during the recording interval occurred K seismic events. If we denote with s_n^k, $n = 0, 1, \ldots, T_k$, the signal produced by the k-th event and with n_k^1 the corresponding wave arrival time, then x_n can be expressed as:

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

where w_n is a noise process. Let us also consider the following transformed signal:

$$y_n = g(x_n),, \ n = 0, 1, \cdots, T-1 \tag{2}$$

where $g()$ is a nonlinear positive transformation acting on the available signal x_n.

The successful solution of the problem of identification of the seismic events contained in the given data record constitutes the basic ingredient in achieving successfully the ultimate goal of picking, that is the estimation of the arrival times n_k^1, $k = 1, 2, \cdots, K$. 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 time. 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 signal, 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 last factor vitally affects the quality of the solution, since if the separation in time of consecutive seismic events is not large enough, multiple seismic events are considered as a single one, thus deteriorating the performance of the detector and consequently the quality of the overall solution. This for example is very common in situations of

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. This work was financed by the University of Patras, "Karatheodori" research program, entitled "The relocation problem of seismic event hypocenter parameters".

very high seismicity where the appearance of consecutive events with limited separation in time from each other is a rule. On the other hand a general rule governing most well known detectors is that their discriminative ability is inversely proportional to the effective window length employed for the estimation of the value of the of the used statistic [3]. Specifically, windows of small effective length result to large discrimination power, but can lead to over-segmentation (or equivalently in an increased false alarms rate), while windows with large effective length result to the packing effect, that is multiple consecutive events are considered as a single one (or equivalently in an increased false misses rate).

All these factors reveal that we must treat the problem in a different manner. As we are going to see in the next paragraphs, we use the detection/segmentation as a vehicle for the identification of a "sufficient subset" of the noise segments contained in the record in order to be able to formulate the original problem in a hypothesis testing framework. Specifically, by exploiting the particular nature of the signals we are treating, and by properly selecting a test statistic, let us denote it by λ_n, we propose a different approach, which is affected to a much lesser degree by the aforementioned factors, thus resulting ultimately to a more robust automatic picking method. As we are going to see in the next section, instead of estimating the onset times from the extracted segments, we propose first to remove the parts of the test statistic sequence where the behavior of λ_n departs from its expected behavior in the noise parts of the record, and then use the remaining part of the sequence in order to obtain a more detailed estimation of the distribution of λ_n in the noise parts. Using this estimation, we return to the initial sequence and identify the areas of interest, i.e the "significant" values of λ_n, by reformulating the picking problem in a hypothesis testing framework. This makes it possible to solve the problem we are interested in by setting the probability of false alarm or false misses in the desired level, and even to appropriately update them if we like to use it in a real time base.

III. The Proposed Solution

Let us consider the set $\mathcal{T} = \{0, 1, \cdots, T-1\}$, where T is the duration of the record and the subset of $\mathcal{T}$, $\mathcal{N}$, containing all the time points n, such that the values of λ_n is calculated in noise intervals of the record, with $|\mathcal{N}| = N$ being its cardinality. Then, the set $\mathcal{E} = \mathcal{T} - \mathcal{N}$, will contain all the time points n, where the values of λ_n are affected by the presence of a seismic event in the record, with $|\mathcal{E}| = T - N$, being its cardinality.

Then, by denoting the probability distribution function (pdf) of λ_n as $f_{\lambda_n}(z)$, and the conditional pdfs of λ_n given each one of the above defined sets as $f_{\lambda_n}(z|\mathcal{N})$ and $f_{\lambda_n}(z|\mathcal{E})$, respectively, and using the Total Probability Theorem, we have that $f_{\lambda_n}(z)$ can be expressed by the following mixture:

$$f_{\lambda_n}(z) = p_0 f_{\lambda_n}(z|\mathcal{N}) + p_1 f_{\lambda_n}(z|\mathcal{E}), \tag{3}$$

where $p_0 \equiv \mathrm{P}\{\mathcal{N}\} = \frac{N}{T}$ and $p_1 \equiv \mathrm{P}\{\mathcal{E}\} = 1 - p_0$ are the corresponding a priori probabilities of occurrence of the sets.

It is clear that in order to be able to reformulate the original problem into a hypothesis testing framework and solve it, knowledge over the pdfs $f_{\lambda_n}(z|\mathcal{N})$ and $f_{\lambda_n}(z|\mathcal{E})$ of λ_n in the sets $\mathcal{N}$ and $\mathcal{E}$, respectively, as well as the a priori probabilities p_0 and p_1 is required. The latter pdf is in fact by itself a mixture of pdfs, determined by the number and the characteristics of the events contained in the record (e.g. the amplitudes of the first arrivals as well as the shapes and the durations of the events). Since we are not in place to make assumptions over these characteristics, the direct estimation of this pdf from the data record is in fact not feasible. On the other hand, since we are considering seismic records of arbitrary length and seismic events are by nature sparse signals, we can safely assume that $N \gg T/2$ (or equivalently $p_0 \gg p_1$). Under this assumption, it seems reasonable to follow a strategy that ensures that all needed estimations are based on the estimation of $f_{\lambda_n}(z|\mathcal{N})$ from the data. Indeed, based on certain assumptions for the stationarity of the noise process and by exploiting the properties of the proposed statistic λ_n, as we are going to see, such an estimation can be achieved.

It is obvious that the problem of estimation of the aforementioned pdf is related with the successful identification of set $\mathcal{N}$. Indeed, if set $\mathcal{N}$ is known, then the problem of estimating $f_{\lambda_n}(z|\mathcal{N})$ from the data becomes a trivial one. However this set is unknown and its identification constitutes in general a difficult task. Since, as it was already mentioned, we are unaware of either the number or the form of the recorded events, the discrimination between "signal" and "noise" can only be based on the available data record, as well as the values of the used test statistic λ_n. This last point makes it clear that the selection of a "proper" test statistic is vital for the successful solution of the problem. Namely, we are expecting that the better the ability of the used test statistic in discriminating between segments belonging to different population is, the better, and more reliable our estimations will become. Such a test statistic was proposed [15] for the solution of the single event P-phase picking problem. In the following subsection we are going to see that this test statistic, as well as its ingredients have some interesting properties, which permit us to successfully identify set $\mathcal{N}$ and to formulate the hypothesis test problem.

A. A Ratio Based Test Statistic

In [15] the use of the following test statistic

$$\lambda_n = \frac{L_n^{M^+}}{L_{n-1}^{M^-}},\ n = 0, 1, \cdots, T-1 \tag{4}$$

was proposed for the solution of the single event P-phase picking problem, where the quantities $L_n^{M^-}$, $L_n^{M^+}$ were defined as follows:

$$L_n^{M^+} = \frac{1}{M} \sum_{k=n}^{n+M-1} y_k\,, \quad L_n^{M^-} = \frac{1}{M} \sum_{k=n-M+1}^{n} y_k. \tag{5}$$

Note that if the noise process is i.i.d. and gaussian and for the special choice of $g(x_n) = x_n^2$, for each value of n the ratio defined in Equ. (4) can be considered as a special form of the well known $\mathcal{F}$-statistic with its numerator and denominator having equal degrees of freedom. However, as it was already mentioned such assumptions are too strong and rarely are valid. In practice, the noise, if it is stationary at all, is colored and its whitening or modeling constitutes a very difficult task [16]. Thus as a rule, for each value of n, the Random Variables (RV) defined in Equ. (5) are sums of correlated RVs. By imposing the constraint of equal window lengths we ensure that they are two identically distributed (i.d.) RVs. As we are going to see in the next paragraph the assumption of exchangeability [17] of such i.d. variables, permits us to solve efficiently the problem at hand.

For the moment, let us concentrate ourselves on the behavior of statistics $L_n^{M^-}$, $L_n^{M^+}$ and λ_n in the different parts of a seismogram. To this end, let us consider that $n_k^i,\ i = 1,\ 2$ and T_k denote the onset time, the stopping time and the duration of the k-th, $k = 1,\ 2, \cdots,\ K$ seismic event contained in the given record, and discriminate the following three cases regarding the specific segments of record used for the computation of the above mentioned statistics.

$\mathcal{C}_1$: $n \in \mathcal{N}$. Set $\mathcal{N}$ can be defined from the union of the following $K+1$ subsets:

$$\mathcal{N}_k = \begin{cases} 0 \le n < n_k^1 - M, & k = 1 \\ n_{k-1}^2 + M < n \le\ n_k^1 - M, & k = 2, 3, \cdots, K \\ n_{k-1}^2 + M < n \le\ T, & k = K+1, \end{cases} \tag{6}$$

where without loss of generality we have assumed that the first and the last segments of the given record correspond to noise segments.

From the above definition it is clear that for each $n \in \mathcal{N}$, RVs $L_n^{M^-}$ and $L_n^{M^+}$, and consequently RV λ_n which is defined by their

ratio, are computed by using samples belonging into noise segments of the record. Moreover, assuming first order ergodicity for the noise process, RVs $L_n^{M^-}$ and $L_n^{M^+}$ constitute two different estimators of the same quantity, namely of the mean of RV $g(w_n)$ using two adjacent (disjoint) intervals of length M each. In addition, it is expected that as the length M of the window increases, the variance of the above mentioned estimators will decrease. It is clear that if set $\mathcal{N}$ was known and taking into account that its cardinality is large, by evaluating the ratio statistic of Equ. (4) for all the members of this set, we could use these values in order to obtain a good approximation of the desired conditional pdf. However set $\mathcal{N}$ is unknown and we are interested on identifying it. To this end we are going to exploit an interesting property satisfied by the distribution of a RV which is the ratio of two positive i.d. RVs whose joint pdf is bivariate symmetric [17].

Proposition 1 : *If the positive RVs $\mathcal{X}$, $\mathcal{Y}$ are exchangeable, i.e. their joint pdf obeys the following symmetry:*

$$f_{\mathcal{X}\mathcal{Y}}(x,y) = f_{\mathcal{X}\mathcal{Y}}(y,x), \tag{7}$$

then the pdfs of RVs $\mathcal{Z} = \mathcal{X}/\mathcal{Y}$ and $\mathcal{Z}' = 1/\mathcal{Z} = \mathcal{Y}/\mathcal{X}$, coincide, that is:

$$f_{\mathcal{Z}'}(z) \equiv f_{\mathcal{Z}}(z). \tag{8}$$

Note that the symmetry property expressed by Equ. (7) is satisfied always under the assumption of i.i.d RVs, while in the i.d. case, where the symmetry reduces to $f_{\mathcal{X}|\mathcal{Y}}(x) = f_{\mathcal{Y}|\mathcal{X}}(x)$, there are well known families of bivariate pdfs [18], [19] satisfying this property which are used for the approximation of pdfs of RVs whose analytical form is unknown. Although we can not prove that the joint pdf of the RVs defined in Equ. (5) obeys such a symmetry without imposing some assumptions on the noise process, we consider that in most cases it is true.

By using Proposition 1 we prove in the sequel a very useful property satisfied by the distribution of the ratio of such RVs which we are going to use in the next subsection for the successful identification of set $\mathcal{N}$.

Proposition 2 : *Let $\mathcal{Z}$ be the ratio of two i.d., exchangeable and positive RVs. Assuming that $f_{\mathcal{Z}}(z)$ is continuous, then for any $x_0 > 1$ its distribution function satisfies the following relation* $\mathrm{P}\{\mathcal{Z} \in [1, x_0]\} = \mathrm{P}\{\mathcal{Z} \in [\frac{1}{x_0}, 1]\}$. *In addition its median value is equal to unity.*

Proof: From the positiveness of $\mathcal{Z}$, $\mathcal{Z}'$ we can easily obtain the following identity:

$$\mathrm{P}\{1 \leq \mathcal{Z}' \leq x_0\} = \mathrm{P}\{1 \leq \frac{1}{\mathcal{Z}} \leq x_0\} = \mathrm{P}\{\frac{1}{x_0} \leq \mathcal{Z} \leq 1\}. \tag{9}$$

Then, by using Proposition 1 and Equ (9), we obtain the following relation:

$$\mathrm{P}\{1 \leq \mathcal{Z} \leq x_0\} = \mathrm{P}\{1 \leq \mathcal{Z}' \leq x_0\} = \mathrm{P}\{\frac{1}{x_0} \leq \mathcal{Z} \leq 1\}, \tag{10}$$

which is the desired one.

By substituting now in Equ. (10) $x_0 = \infty$ we can easily see that the median value of $\mathcal{Z}$ is equal to unity, and this concludes the proof of the proposition.

$\mathcal{C}_2$: $n \in \mathcal{E}$. Set $\mathcal{E}$ can be defined from the union of the following K subsets, created by the presence of the K events in the record:

$$\mathcal{E}_k = n_k^1 - M + 1 \leq n \leq n_k^2 + M,\ k = 1, 2, \cdots, K. \tag{11}$$

Contrary to $\mathcal{C}_1$, in this case the behavior of the statistics depends on uncontrollable factors such as the magnitude and the shape of the events, as well as their separation in time and as such, cannot be assessed in a statistical manner. It can be explained only intuitively, by taking into account the particular statistics used, as well as the general characteristics of the seismic signals, namely the fact that the energy of a seismic signal is higher than the energy of noise, and the "generic" amplitude envelope of these particular signals, i.e. their fading nature.

Specifically, let us first concentrate on the sequences $L_n^{M^+}$ and $L_n^{M^-}$, calculated for a given record. These sequences behave as smoothed positive "envelopes" of y_n, following conceptually its shape in intervals containing events, and attaining low values, around a constant level, in intervals of noise. Based on this, and due to the higher energy of the seismic signal compared to the energy of the noise, these sequences take higher values during the occurrence of an event, than they take in intervals of noise. Hence, in each $\mathcal{E}_k$, we expect both $L_n^{M^+}$, $L_n^{M^-}$ to take values that depart from their noise level, making these statistics an appropriate segmentation "tool".

On the other hand, the behavior of λ_n, is different in different parts of $\mathcal{E}_k$, and as such, can not be described "uniformly" for the whole interval. In order to be able to describe its behavior, we can divide $\mathcal{E}_k$ in the following three subintervals:

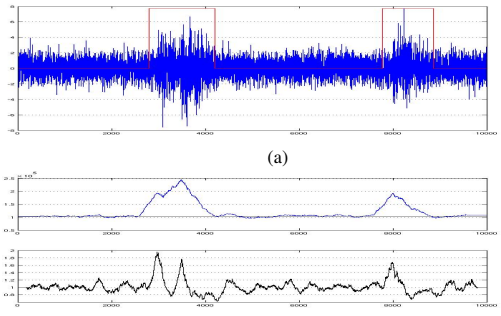
- $\mathcal{E}_k^1 = n_k^1 - M + 1 \leq n \leq n_k^1$. In this interval, the window corresponding to $L_n^{M^+}$ will gradually cover the beginning of the seismic signal s_n^k, thus causing the values of $L_n^{M^+}$ to grow, while $L_{n-1}^{M^-}$ will still account only for noise. This results in a gradual rise in the values of λ_n, attaining ideally its maximum at the onset time, $n = n_k^1$, where the whole right-hand window is placed over signal and the whole left-hand one is placed over noise. Depending on the amplitude of the first arrival samples, the values of λ_n in this interval can be very high near the onset time (i.e. near the top of the peak).
- $\mathcal{E}_k^2 = n_k^1 < n \leq n_k^1 + M$. In this case, as the samples from the beginning of s_n^k will enter the denominator of the ratio, and as the corresponding numerators values become gradually lower, λ_n will exhibit a steep drop from its previous large values.
- $\mathcal{E}_k^3 = n_k^1 + M < n \leq n_k^2 + M$. Finally, in this interval, the windows corresponding to the numerator and the denominator of the ratio, will gradually transit from both covering signal parts, to both covering noise. Although the behavior of λ_n in this interval is basically unpredictable, due to aforementioned inherent fading nature of seismic events, the numerator of the ratio is expected to attain systematically lower values than the denominator, thus in this interval, λ_n will be biased towards values lower than 1. Note also that, λ_n, being a ratio statistic, is basically insensitive to the amplitude of the event in these intervals. Taking this into account, and also the fact that the fading of the event amplitudes takes place gradually, as opposed to its beginning where the change is more radical, we can say that λ_n is not expected to take "extreme" values in this interval,

Thus concerning λ_n, the presence of the k^{th} event in the record, i.e. for $n \in \mathcal{E}_k$, is reflected by the presence of a narrow peak (of approximately $2M$) samples in the beginning of such intervals, where λ_n takes "significantly" large values (depending on the signal-to-noise ratio), followed by a large interval (of approximately T_k samples), of "non-significant" values that are biased towards a level lower than unity.

The aforementioned behavior of the used statistics in a synthetic and a real data record, is displayed in Figs. 1 and 5.c respectively.

By taking into account the above analysis, it becomes apparent that in order to solve the picking problem, in the general setting presented in this paper, the most crucial step is the identification of the "significant" peaks of the calculated λ_n sequence, under the assumption that each one of them denotes the beginning of an event.

In order to achieve this goal we reformulate this identification problem in a hypothesis testing framework. More specifically, we make the assumption that any "significantly" high value of λ_n belongs to a peak, thus constituting the following one-sided hypothesis testing problem (since the peak values will manifest themselves on the right



(a)

(b)

Fig. 1. Behavior of the statistics λ_n and $L_n^{M^+}$ in the presence of seismic events. (a): A synthetic record. (b): Sequence of the statistics λ_n (red) and $L_n^{M^+}$ (blue).

tail of $f_{\lambda_n}(z)$):

$$\mathcal{H}_0 : \lambda_n \text{ corresponds to noise}$$
$$\mathcal{H}_1 : \lambda_n \text{ corresponds to a peak.}$$

If $f_{\lambda_n}(z|\mathcal{N})$ is known, then we can solve this problem by selecting a statistical significance level (or probability of false alarm) α thus determining a threshold T_α such that $P(\lambda_n > T_\alpha|\mathcal{N}) < \alpha$, and then decide $\mathcal{H}_1$ for all values of λ_n which exceed this threshold.

However $f_{\lambda_n}(z|\mathcal{N})$ is unknown and we must somehow estimate it. This is exactly the goal of the next subsection.

B. Estimation of conditional pdf $f_{\lambda_n}(z|\mathcal{N})$

In this Section, we treat the problem of estimation of $f_{\lambda_n}(z|\mathcal{N})$ from data by identifying set $\mathcal{N}$. Since the direct estimation of $\mathcal{N}$ from the data is not straightforward, in order to achieve our goal we are going to follow a two step procedure. Specifically, in the first step we exploit:

- the appropriateness of statistic $L_n^{M^+}$ for its use as a segmentation "tool", and
- the fact that the values of the ratio test are biased in intervals containing seismic events,

and we obtain a "gross" segmentation of the record, by selecting a very "conservative" threshold, thus making sure that all the signal intervals are selected.

In the second step using

- the symmetry imposed to the desired conditional pdf by Proposition 2

we identify the desired set by solving a well defined optimization problem.

Let us analyze in detail each one of the above step starting from step one, that is the "gross" segmentation of the record.

Let us consider the following sequence of intervals

$$\widehat{\mathcal{E}}_1, \widehat{\mathcal{E}}_2, \cdots, \widehat{\mathcal{E}}_L. \tag{12}$$

resulting from the sequence $L_n^{M^+},\ n = 0, 1, \cdots, T-1$ by sequentially performing the following actions:

- $\mathcal{A}_1$: find the intervals where the values of $L_n^{M^+}$ are greater than m_L,
- $\mathcal{A}_2$: discard all intervals whose the median of test statistic λ_n is lower than unity, and
- $\mathcal{A}_3$: sort the interval according their cardinality, i.e. $|\widehat{\mathcal{E}}_1| \geq |\widehat{\mathcal{E}}_2| \geq \cdots \geq |\widehat{\mathcal{E}}_L|$.

where m_L denotes the median value of sequence $L_n^{M^+},\ n = 0, 1, \cdots, T-1$, which by taking into account the sparsity of the seismic events in the data record constitutes a good estimator of the median of this statistic under the assumption of noise ($n \in \mathcal{N}$), $m_{L|\mathcal{N}}$.

Since the selected threshold is indeed a very conservative one, we anticipate that the union of all intervals formed from the application of action $\mathcal{A}_1$, will contain all the segments of the signal intervals of the record (i.e. all $\mathcal{E}_k$), as well as a great number of other segments, containing only noise, due to the random fluctuations of $L_n^{M^+}$ around m_L, in the noise intervals. Based on the fact that in intervals containing seismic events, the median of λ_n should be lower than unity, according to the analysis of the previous subsection, we limit the number of intervals of the latter type, by performing action $\mathcal{A}_2$, that is by discarding all the intervals where the median value of λ_n is greater than unity. Finally, we sort the remaining intervals according to their length, in descending order, making the inherent assumption that the intervals corresponding to seismic events are by nature longer than the ones formed by the random oscillations of $L_n^{M^+}$. This way, the signal intervals are anticipated to be in the beginning of the resulting interval sequence, followed by the ones that contain only noise.

Finally, we must stress at this point that since our ultimate goal is the identification of $\mathcal{N}$ and not the accurate detection of the intervals $\mathcal{E}_k$ (as would be the case in a common detection - picking approach), a one-to-one correspondence of each $\mathcal{E}_k$ to one of the members of the above sequence is neither assumed nor needed for the success of our method. For example, in the case of events occurring very close in time (as the case depicted in Fig.1), it is possible that the noise intervals between them (if any) are not sufficiently long to permit the values of $L_n^{M^+}$ to drop below m_L, between the events. If that is the case, then all the consecutive events will be packed in one single interval $\widehat{\mathcal{E}}_l$ of the above sequence. What we are stressing out at this point is that, while situations like this (that occur very often in seismic records) would pose a serious problem if the goal was the successful identification of each event interval separately (i.e. the identification of each $\mathcal{E}_k$), for the proposed approach are not considered as problematic since the particular way the signal intervals are removed are of no importance to the method.

Let us now proceed to the second step of the identification of the desired set.

To this end let us define the following sequence of sets:

$$\widetilde{\mathcal{N}}_l = \widetilde{\mathcal{N}}_{l-1} \cap \widehat{\mathcal{E}}_l',\ \widetilde{\mathcal{N}}_0 = \mathcal{T},\ l = 1, \cdots, L, \tag{13}$$

where $\widehat{\mathcal{E}}_l'$ denotes the complement of set $\widehat{\mathcal{E}}_l$. It is obvious that for the defined sequence of sets, the following relations hold:

$$\widetilde{\mathcal{N}}_l = \mathcal{T} \cap \left(\bigcup_{i=1}^{l} \widehat{\mathcal{E}}_i \right)', \tag{14}$$

with $|\widetilde{\mathcal{N}}_l| = |\widetilde{\mathcal{N}}_{l-1}| - |\widehat{\mathcal{E}}_l|$.

Note now that each set $\widetilde{\mathcal{N}}_l$ can be used for the approximation of the following pdf:

$$f_{\lambda_n}(z|\widetilde{\mathcal{N}}_l) = \frac{|\widetilde{\mathcal{N}}_{l-1}|}{|\widetilde{\mathcal{N}}_l|} f_{\lambda_n}(z|\widetilde{\mathcal{N}}_{l-1}) - \frac{|\widehat{\mathcal{E}}_l|}{|\widetilde{\mathcal{N}}_l|} f_{\lambda_n}(z|\widehat{\mathcal{E}}_l). \tag{15}$$

If there exists a member $\widetilde{\mathcal{N}}_{l^\star}$ of the above defined sequence that approximates set $\mathcal{N}$, then the corresponding pdf $f_{\lambda_n}(z|\widetilde{\mathcal{N}}_{l^\star})$ can be considered as a good approximation of the ideal one. Thus, our goal now is to identify this member of the sequence defined in Equ (13). In order to solve this problem we resort to the results of Proposition 2. Specifically, we define the following cost function:

$$C(\widetilde{\mathcal{N}}_l) = \int_1^\infty \left[\int_1^x f_{\lambda_n}(z|\widetilde{\mathcal{N}}_l)dz - \int_{1/x}^1 f_{\lambda_n}(z|\widetilde{\mathcal{N}}_l)dz \right]^2 dx, \tag{16}$$

and solve, in an iterative fashion, the following minimization problem:

$$l^{\star} = \arg\min_{l} C(\widetilde{\mathcal{N}}_l). \tag{17}$$

In the l-th iteration of the above procedure we consider the values of λ_n, $n \in \widetilde{\mathcal{N}}_l$ as a sample drawn from the distribution governed by the pdf $f_{\lambda_n}(z|\mathcal{N})$ and obtain an estimation of the latter, by a detailed histogram of the sample. Having this estimation, we evaluate the cost function defined in Equ. (16), in order to assess the degree by which the estimated pdf exhibits the symmetry property imposed by Proposition 2 to $f_{\lambda_n}(z|\mathcal{N})$. Let us now assume that the first K elements of the sequence defined in Equ. (12), are intervals that correspond to seismic events, while the rest $L-K$ correspond to noise. Since the behavior of λ_n is totally different in signal and noise intervals, which in turn is reflected by the assumed symmetry of $f_{\lambda_n}(z|\mathcal{N})$, the first K iterations of the algorithm will lead to estimated pdfs that are increasingly more symmetric (by removing intervals of λ_n values corresponding to signal intervals), thus resulting in a decreasing sequence of $C(\widetilde{\mathcal{N}}_l)$ values. After that, for $l = K+1, \cdots, L$, although the intervals that are removed belong to the set $\mathcal{N}$, by the construction of the sequence in Equ. (12), the values of λ_n that belong to these intervals are systematically biased (recall that the median of λ_n, $n \in \widehat{\mathcal{E}}_l$ is lower than unity, which represents the theoretical value of the median of $f_{\lambda_n}(z|\mathcal{N})$). Subsequently, the pdfs $f_{\lambda_n}(z|\widetilde{\mathcal{N}}_l)$, $l = K+1, \cdots, L$ will exhibit the symmetry property to an increasingly lesser degree, thus resulting in increasing the values of sequence $C(\widetilde{\mathcal{N}}_l)$, $l = K+1, \cdots, L$. Consequently, by calculating the function defined in Equ. (16) for all the members of pdf sequence defined in Equ (12), and by taking into account the systematic way we followed for the construction of this sequence, we are expecting that $C(\widetilde{\mathcal{N}}_l)$ will attain its minimum value for the member of the sequence that we are interested in identifying ($l^{\star} = K$) and thus giving the following estimation for $f_{\lambda_n}(z|\mathcal{N})$:

$$f_{\lambda_n}(z|\widetilde{\mathcal{N}}_{l^{\star}}) \approx f_{\lambda_n}(z|\mathcal{N}). \tag{18}$$

Let us comment on some advantages of the proposed approach. To this end, let us consider the two "error" type cases, namely $\widetilde{\mathcal{N}}_{l^{\star}} \subset \mathcal{N}$ and $\widetilde{\mathcal{N}}_{l^{\star}} \supset \mathcal{N}$. The first one means the solution of optimization problem defined in (17) returned a number greater than the actual number signal intervals ($l^{\star} > K$) leading to false identification of noise segments as signal ones, (meaning that the estimation of $f_{\lambda_n}(z|\mathcal{N})$ will be based on a biased sample) while the second ($m^{\star} < K$) leads to signal segments being considered as noise ones (meaning that the estimation of $f_{\lambda_n}(z|\mathcal{N})$ will be based on a "mixed" sample). In both cases the quality of the final solution is affected only by the amount the estimation of $f_{\lambda_n}(z|\mathcal{N})$ is affected. Considering the fact that this estimation will be based on a large sample (assuming a relatively long seismic record), the proposed approach is more robust and "forgiving" since such small contamination problems affect the quality of the overall solution to a much lesser degree than the alternative approaches, where the above mentioned problems will certainly lead to false alarms (first case) or false misses (second one).

Concluding, having estimated the desired conditional pdf, we can easily solve the problem of hypothesis testing defined in the previous subsection resulting in the identification of the "significant" peaks, and then on a second step we can apply the picking method proposed in [15] to each one of them, for the total solution of the automatic picking problem, i.e. the estimation of the arrival times n_k^1, $k = 1, 2, \cdots, K$.

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. The synthetic seismic signals, described by Equ. (1), were modeled as low-pass filtered Gaussian noise, multiplied by a half-Gaussian window for the effect of amplitude shaping, and a constant gain, controlling the signal-to-noise ratio (SNR). In order to construct a data record, first the noise process w_n was created and then a value for the number of events K was selected randomly in a range $[K_{min}, K_{max}]$. Then the synthetic signals s_n^k were created as described above, by selecting randomly K respective SNR values in a range $[SNR_{min}, SNR_{max}]$. Finally, the onset times were obtained by random selection in the interval $[1, T]$ and the resulting signal x_n was calculated, by using Equ. (1).

A. Experiment I

In the first experiment, we evaluate the performance of the proposed procedure in Section III for the estimation of the conditional pdf $f_{\lambda_n}(z|\mathcal{N})$ when its theoretical counterpart is known. To this end we used as the noise process white Gaussian noise, i.e. $w_n \sim N(0,1)$ and as $g(.)$ the squared norm of the signal, i.e. $g(x_n) = x_n^2$. By making these selections, it is well known that $f_{\lambda_n}(z|\mathcal{N}) \sim F(M, M)$, where $F(n_1, n_2)$ is the F distribution with n_1 and n_2 degrees of freedom and M is the window size ($M = 100$ in this experiment). We constructed a synthetic data set of 1000 records x_n of duration $T = 30000$ samples, each containing a varying number of seismic events between 5 and 10 with their SNRs between 5 and 15 dB. Then, we applied the proposed method to this data set and calculated the mean of all the estimated $f_{\lambda_n}(z|\mathcal{N})$.

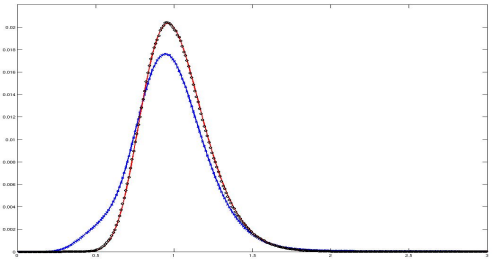


Fig. 2. Mean histogram of the ratio sequence (blue), mean of estimated histograms obtained from the application of the proposed fitting procedure (black) and its theoretical counterpart (red).

The results we obtained from the application of the proposed procedure are shown in Fig. 2. For comparison purposes, we also plotted the mean of all the $f_{\lambda_n}(z)$. As we can see from this figure the estimated histogram is perfectly matching the theoretical one, thus revealing its effectiveness.

B. Experiment II

In the second experiment the noise process we used is a superposition of a second order AR process, modelling the low-frequency seismic noise, as well as Gaussian and uniform noise, where $g_n \sim N(0,1)$ and $u_n \sim U(-0.5, 0.5)$. All the other parameters of the model were identical to the one used in the first experiment. An example of a synthetic record of the data set containing 3 events with SNRs 7,7 and 6, respectively is shown in Fig. 3. Finally, a more appropriate function $g()$ that is based on the notion of the length of the seismic curve proposed in [15] is used instead of the squared norm of the signal.

We then tested our method using this data set and calculated the percentage of the successfully detected events as well as the percentage of false alarms, for different values of α. We considered as successful detections only the cases where the estimated onset time was in a reasonable neighborhood of the real one (we used a threshold of 50 samples). Fig. 4 shows a plot of the obtained results for values of the significance level α ranging from 2% to 0. The selected window size was $M = 100$. From this figure it is clear the

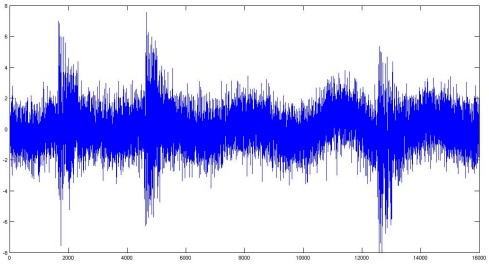


Fig. 3. Example of a synthetic seismogram used in Experiment II.

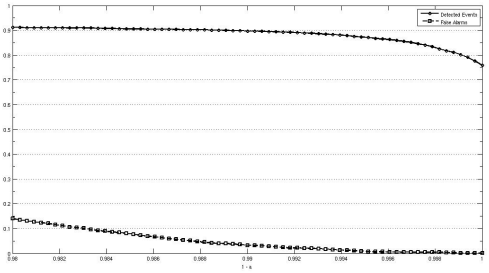


Fig. 4. Successful detections and false alarms for the experiment using the synthetic data set.

high performance of the proposed method, resulting in a successful identification of approximately 90% of the events with a percentage of false alarm in the neighborhood of 0.5%, for values of α in the neighborhood of 0.05%.

C. Experiment III

In this last experiment we evaluate the performance of the proposed method by applying it in real seismic data. The real data set, was comprised by 300 pre-cut records of size T=60000 samples (10 min) each. The "true" number of events, counted by a human analyst contained in the above mentioned records were 2312, with different amplitudes and durations. By using $M = 100$ (1 sec), the proposed detector was applied to the above data set with the probability of false alarm ranging from 0.2-0.5$\%$ and succeeded in identifying 2098 events. This results in an successful identification in approximately 91% of the cases, thus confirming its appropriateness for the problem at hand. This is also evident in Fig. 5.c where the results of the solution to the detection problem for the record of Fig. 5.a are shown, as well as from Fig. 5.d where the curve shown in Fig. 5.b (after its multiplication with the sign of the backward differences of the signal for a more comprehensive view) and the results of the identification are superimposed.

V. Conclusions

In this paper the problem of simultaneous detection of seismic events contained in a given data record was examined. The problem was formulated as a hypothesis testing one and an approximation of the probability density function under the null hypothesis of a ratio based test statistic was proposed. The effectiveness of the proposed technique was confirmed from its application in solving the detection problem on a series of experiments, where synthetic and real seismic records were used.

References

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

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

[3] G. Xiantai.

[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.

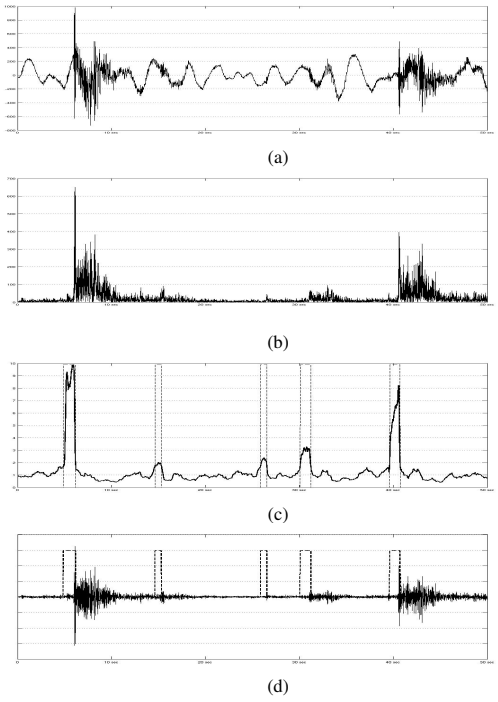


Fig. 5. Example of the application of the proposed method on the real data set. (a): The recorded signal. (b): Values of $g(x_n)$. (c): Identification of significant lobes of λ_n. (d): Plot of the identified areas of interest against the signal for a more comprehensive view.

[5] ——, "Estimation of the arrival times of seismic waves by multivariate time series models," *Ann. Inst. Stat. Math.*, vol. 43, pp. 407–433, 1991.

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

[7] J. E. P. Gendron and D. Manolakis, "Rapid joint detection and classification with wavelet bases via bayes theorem," *Bull. Seism. Soc. Am.*, vol. 90, pp. 764–774, 2000.

[8] C. T. H. Zhang 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.

[9] 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.

[10] L. H. C.D. Saragiotis 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.

[11] J. R.-H. J.J. Galiana-Merino 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.

[12] Y. O. S. Nakamula, M. Takeo and M. Matsuura, "Automatic seismic wave arrival detection and picking with stationary analysis: Application of the KM_2O - langevin equations," *Earth Planets Space*, vol. 59, pp. 567–577, 2007.

[13] E. K. T. Diehl, N. Deichmann and S. Husen, "Automatic s-wave picker for local earthquake tomograph," *Bull. Seism. Soc. Am.*, vol. 99, pp. 1906–1920, 2009.

[14] 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.

[15] E. Pikoulis and E. Psarakis, "A ratio test for the accurate automatic p - wave onset detection," *In Proceedings of the ICSP-2010*, pp. 2621–2624, Beijing, 2010.

[16] M. Basseville, "Detecting changes in signals and systems - a survey," *Automatica*, vol. 24, pp. 309–326, 1988.

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

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

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