A fuzzy c-means assisted AIC workflow for arrival picking on downhole microseismic data

Eduardo Valero Cano*, Jubran Akram, Daniel Peter, King Abdullah University of Science and Technology (KAUST) and Leo Eisner, Seismik

SUMMARY

We propose a workflow for automatic P- and S-wave arrival picking on downhole microseismic data. It uses conditional fuzzy c-means clustering to identify time intervals of possible wave arrivals. We classify the signal intervals as P- and S-wave using the first and second eigenvalues of the waveforms contained within. The Akaike information criterion (AIC) picker is then applied to the identified P- and S-wave intervals for arrival picking. Using real microseismic dataset examples, we show that the proposed workflow yields accurate arrival picks for both high and low signal-to-noise ratio waveforms. The identification of signal intervals, however, uses features based on amplitude, thus remains susceptible to high amplitude noise.

INTRODUCTION

Microseismic monitoring is a valuable tool for understanding fracture growth and behavior in unconventional reservoirs. The information obtained can also be used to optimize planning and stimulation design for future wells. One of the main objectives is to map the microseismicity, which is done using traveltime based methods for downhole arrays. Accurate arrival picking of P- and S-waves is required for use in the polarization analysis, velocity model calibration and in the estimation of hypocenter locations. However, this is a challenging task due to poor signal-to-noise ratio (SNR) of the waveforms and large dataset sizes (hours to days of continuous recordings).

Previously, numerous automatic arrival picking methods have been proposed (see, e.g., Akram and Eaton (2016) who compare different algorithms including the short- and long-term average ratio, STA/LTA; phase arrival identification-kurtosis; and cross-correlation pickers). The Akaike Information Criterion (AIC) proposed by Akaike (1998) is one of the frequently used arrival picking methods that works well when there is only a single arrival present within the selected window. For multiple arrivals, it is necessary to separate each arrival into non-overlapping time intervals.

Recently, supervised and unsupervised machine learning methods have gained considerable popularity for use in many geophysical applications. Although supervised machine learning methods like artificial neural networks have high success rate, these require a good training dataset. Conversely, unsupervised machine learning methods can yield good results even for low SNR waveforms without requiring any training dataset.

Here, we propose a two-step workflow that explores unsupervised machine learning methods to identify signal intervals for use as input to the AIC method. First, we apply FCM clustering to identify the time intervals containing possible wave

arrivals. We also use a modification of FCM, known as the conditional fuzzy c-means (CFCM) clustering to identify the time intervals. Second, we use the AIC method to pick the onset of P- and S-waves in the corresponding signal intervals. The proposed workflow is tested on a real microseismic dataset acquired with a downhole array.

THEORY

Fuzzy c-means clustering

The fuzzy c-means (FCM) is a soft partitional clustering method (Zadeh, 1977; Bezdek et al., 1984). It partitions a set of N points $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$, in a F -dimensional Euclidean space into C clusters by minimizing the objective function:

$$J(U, \mathbf{V}) = \sum_{k=1}^N \sum_{i=1}^C (u_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|^2 \quad (1)$$

where U is the partition matrix which elements $u_{ik} \in [0, 1]$ represent the degree of membership of the point $\mathbf{x}_k$ to the cluster i , $\mathbf{V} = \mathbf{v}_1, \dots, \mathbf{v}_C$ is a set of vectors $\mathbf{v}_i$ that represent the centroid of cluster i , $\mathbf{x}_k$ is a vector that represents the k -th point of the set $\mathbf{X}$, $m \in [1, \infty)$ is the controller of cluster fuzziness, and $\|\cdot\|$ is any norm.

One approach to minimize J is to update the set of centroids $\mathbf{V}$ and the partition matrix U via iterations of:

$$\mathbf{v}_i = \frac{\sum_{k=1}^N u_{ik}^m \mathbf{x}_k}{\sum_{k=1}^N u_{ik}^m}, \quad 1 \leq i \leq C \quad (2)$$

$$u_{ik} = \frac{1}{\sum_{j=1}^C \left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|} \right)^{\frac{2}{m-1}}}, \quad 1 \leq k \leq N; 1 \leq i \leq C \quad (3)$$

where equation (3) has the constraint $\sum_{i=1}^C u_{ik} = 1$ for all k .

The similarity metric of the points and the shape of the clusters depend on the choice of norm $\|\cdot\|$. Here, we use the L2 norm, which induces a similarity metric based on Euclidean distance and clusters of hyperspherical shape.

Conditional fuzzy c-means clustering

The conditional fuzzy c-means (CFCM) is a modified version of the FCM method where an auxiliary variable f_k , called condition, is used to introduce *a priori* information to the clustering process (Pedrycz, 1996). In the CFCM method, each point $\mathbf{x}_k$ is associated with a condition f_k and the partition matrix U is updated via:

$$u_{ik} = \frac{f_k}{\sum_{j=1}^C \left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|} \right)^{\frac{2}{m-1}}}, \quad 1 \leq k \leq N; 1 \leq i \leq C \quad (4)$$

Arrival picking using a fuzzy c-means assisted AIC workflow

with the constraint $\sum_{i=1}^C u_{ik} = f_k$ where $f_k \in [0, 1]$.

This method controls the influence of the point $\mathbf{x}_k$ in the clustering process by modifying its degree of membership to all clusters i via the condition f_k . If a point $\mathbf{x}_k$ has a condition less than one, its degree of membership to all clusters will decrease, minimizing its influence to the centroids of all clusters. However, if the condition of k is equal to one, its degree of membership to all clusters will be unaffected, preserving its influence to the centroids of all clusters.

AIC based wave arrival picker

The Akaike Information Criterion (AIC) is a model selection technique developed by Akaike (1998). The AIC can be used for wave arrival onset picking on a single component trace (Sleeman and Eck, 1999). It assumes that a seismic trace can be divided into locally stationary segments where each is modeled as an autoregressive process. The onset time of a wave arrival separates two different segments and is associated with the minimum of the AIC values.

Usually, the autoregressive process coefficients are needed to calculate the AIC function, however, Maeda (1985) calculates it directly from the trace as follows:

$$\text{AIC}(k) = k \log(\text{var}\{x(1, k)\}) + (N - k - 1) \log(\text{var}\{x(k + 1, N)\}) \quad (5)$$

where x is a trace of N samples, k ranges from 1 to N , and $\text{var}\{x\}$ is the variance function. In this study, we use equation (5) to pick the onset of P- and S-wave arrivals.

CFCM assisted AIC based wave arrival picker

Assuming that each sample of a microseismic trace represents only noise or seismic signal – from here on referred to as signal, we consider the identification of wave arrivals as a binary clustering problem (Chen, 2017). Following this definition, we define the samples $k = 1, \dots, N$, of a trace $d(k)$ as a set of points $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$, in a F -dimensional Euclidean space (Figure 1), where $\mathbf{x}_k$ is a feature vector which elements represent the value of some feature of $d(k)$ (e.g. mean) at the sample k . F is the number of features, and N is the number of samples. Moreover, we set the number of clusters C equal to two, denote the cluster number by i , where $i = 1$ is the noise cluster and $i = 2$ is the signal cluster, and denote the vector that represents the centroid of cluster i by $\mathbf{v}_i$.

Let us define the feature vectors $\mathbf{x}_k$ that comprise the following features of $d(k)$:

Mean of the absolute value of the amplitude:

$$M(k) = \frac{1}{N} \sum_{k-w}^{k+w} |d(k)| \quad (6)$$

Peak power spectral density:

$$P(k) = \max(|D(k, \omega)|^2) \quad (7)$$

Short- and long-term average ratio:

$$R(k) = \frac{STA}{LTA} = \frac{\frac{1}{SW} \sum_{j=k-SW}^k |d(j)|}{\frac{1}{LW} \sum_{j=k-LW}^k |d(j)|} \quad (8)$$

where in equation (6), the constant w is half the length of the window around k and set to half the dominant period of the signal T_{dom} , in equation (7), $D(k, \omega)$ is the modulus of the discrete short-time Fourier transform of $d(k)$, and in equation (8), SW and LW are the lengths of the short- and long-term windows, respectively.

Using this definition, identification of wave arrivals is a clustering problem with highly unbalanced size clusters (in terms of their numbers of points). The reason is that seismic signals only represents a small segment of a trace, making the signal cluster much smaller than the noise cluster. In this situation, FCM performs poorly because the centroids of small clusters tend to drift towards adjacent large clusters, leading to a non-optimal partitioning (Noordam et al., 2002). To alleviate this problem, the CFCM algorithm is used to identify wave arrivals. We assume that the trace features tend to vary less for noise than for signals, and therefore, the spread of the noise cluster is less than the spread of the signal cluster. Based on this assumption, we define the condition f_k as:

$$X_i = \{\forall k \mid u_{ik} > \bar{u}_i\}, \quad 1 \leq k \leq N; 1 \leq i \leq C \quad (9)$$

$$\sigma_i = \frac{\text{var}\{M(k)\} + \text{var}\{P(k)\} + \text{var}\{R(k)\}}{3}, \quad k \in X_i \quad (10)$$

$$f_k = \frac{\sigma_i}{\sigma_{max}}, \quad 1 \leq k \leq N; 1 \leq i \leq C \quad (11)$$

where in equation (9), X_i is the set of samples k that are member of cluster i and $\bar{u}_i$ the mean of u_{ik} over all k , in equation (10), σ_i is the variance of cluster i , and in equation (11), σ_{max} is the highest σ_i of all clusters.

With this definition, f_k will be low for the noise samples, decreasing their degrees of membership to both clusters and minimizing their contribution to the centroid of the signal cluster, similar to the cluster size insensitive FCM method proposed by Noordam et al. (2002). We use a condition f_k that depends on the spread of the clusters instead of one that depends on their number of samples because, in the last case, the high number of noise samples will cause their conditions to be extremely low, decreasing their degree of membership to both clusters almost to zero. This would cause the centroid of the noise cluster to drift towards the signal cluster, leading to a poor partitioning of the samples.

Figure 2 summarizes the proposed workflow. To identify the intervals containing the wave arrivals on a microseismic trace, we calculate the input features as described in equations (6) to (8). Min-max normalization is applied to the features. To avoid that CFCM gets stuck in a local minima, we use the results obtained from the FCM method as input to the CFCM method (Noordam et al., 2002). CFCM is implemented as follows:

1. Obtain a partition matrix U using FCM.

Arrival picking using a fuzzy c-means assisted AIC workflow

2. Assign each sample k to a cluster i via de-fuzzification of the partition matrix U using equation (9).
3. Calculate the conditions f_k using equations (10) and (11).
4. Update the partition matrix U using equation (4).
5. Update the set of centroids $\mathbf{V}$ using equation (2).
6. Repeat steps 2 to 5 until changes in U are less than a predefined stop criterion (set to 10^{-4} in this study).

This process is repeated for the components c of each receiver, and assuming that the signal is arriving on all 3 components simultaneously, their $u_{i=2,k}$ values are stacked to highlight the wave arrivals:

$$u_s = \frac{1}{3} \sum_{c=1}^3 u_{c,i=2,k} \quad (12)$$

where u_s is the stacked degree of membership to the signal cluster.

Values of u_s will suddenly increase in time intervals corresponding to the wave arrivals. To identify these intervals, a threshold β must be defined. Any continuous interval where u_s is greater than β for at least one dominant period T_{dom} duration is considered to be containing a possible wave arrival T_A . Once the signal intervals are identified, we distinguish between P- and S-waves arrivals based on their polarization properties that are obtained from the eigenvalue analysis of the covariance matrix of 3C waveforms in moving windows. Considering the three eigenvalues ($\lambda_1 \geq \lambda_2 \geq \lambda_3$), the signal interval that exhibits the maximum value of λ_1 is deemed to contain the P-wave arrival whereas the signal interval with maximum value of λ_2 is deemed to contain the S-wave arrival. To pick the onset of P- and S-wave arrivals, the AIC picker is used on the corresponding intervals that are extended from their start by two times T_{dom} . Since the arrivals are picked on each component, these picks can be different especially for noisy datasets. To get a representative arrival time, a weighted average of the arrival times from three components is calculated where the SNR in the arrival window is taken as weight w_c .

APPLICATION TO REAL DATA

We apply the proposed workflow on real microseismic data examples. The microseismic data were acquired by a downhole array of 20 three-component receivers during a hydraulic fracturing operation. Figure 1 shows the point representation of the samples of the trace shown in Figure 3(a), along with the results of CFCM, where the trace shows a relatively high SNR vertical component trace of recorded particle velocity and its features.

In this example, all three input features distinguish the signal intervals from noise intervals. High values of all features represent the signal intervals whereas low feature values represent noise intervals. Consequently, the samples of signal arrivals are far from the rest of the samples in the F -dimensional Euclidean space (Figure 1), allowing CFCM to identify time intervals of the wave arrivals. In Figure 3(a), we identify two corresponding signal intervals with high degree of membership

that are later categorized as P-wave (red) and S-wave (blue). Because of the high SNR of input waveforms, the AIC picker is able to pick accurate arrival times for both P- (green dot) and S-wave (orange dot).

Figure 1: Point representation of the samples of the trace (shown in Figure 3(a)). Each point represents one sample of the trace. Different colors indicate the degree of membership to each cluster. The black and red stars represent the centroids of the signal and noise clusters, respectively.

Figure 3(b) shows a relatively low SNR vertical component trace of recorded particle velocity and its features. In this noisy example, high values of the features do not always correspond to the signal arrivals. It is, therefore, difficult to distinguish the signal from the noise intervals, especially on the STA/LTA. Because of this, CFCM identifies several potential intervals containing signal. Still, using the polarization properties enabled us to select P- and S-wave intervals correctly.

DISCUSSION AND CONCLUSIONS

The CFCM method correctly identifies signal intervals as long as the input features exhibit a distinguishing pattern for the signal and noise. Furthermore, the AIC is a robust onset picker as long as it is used on an interval that contains the arrival with an amplitude higher than the noise. The AIC method picks the sample at global minimum, and cannot pick multiple arrivals in the selected time interval. It is, therefore, important to identify the P- and S-wave intervals before applying the AIC method.

Unsupervised machine learning methods such as (conditional) fuzzy c-means clustering do not require any training datasets. These methods can be used effectively to identify the time intervals where P- and S-waves can be picked by the AIC method. These signal intervals can be classified into P- and S-wave using the first and second eigenvalues of the waveforms contained within. The proposed method successfully identifies the P- and S-wave intervals for even low signal-to-noise traces such as illustrated in Figure 3(b). Since we used features that are based on amplitude, the identification of signal intervals however remains susceptible to high amplitude noise.

Arrival picking using a fuzzy c-means assisted AIC workflow

Figure 2: Proposed workflow. Gray boxes indicate the steps associated with CFCM and the AIC based wave arrival picker.

Figure 3: Relative high SNR trace (a) and low SNR trace (b). Top panel: vertical component trace of recorded particle velocity. The green dot represents the P-wave (red) arrival onset. The orange dot represents the S-wave (blue) arrival onset. The following trace features are shown in the figure panels from the top: mean (second panel), peak power spectral density (third panel), STA/LTA (fourth panel). The bottom panel represents the degree of membership to the signal cluster of the trace.

REFERENCES

- Akaike, H., 1998, Information theory and an extension of the maximum likelihood principle, *in* Selected papers of hirotugu akaike: Springer, 199–213.
- Akram, J., and D. Eaton, 2016, A review and appraisal of arrival-time picking methods for downhole microseismic data: *GEOPHYSICS*, **81**, KS67–KS87.
- Bezdek, J. C., R. Ehrlich, and W. Full, 1984, Fcm: The fuzzy c-means clustering algorithm: *Computers & Geosciences*, **10**, 191–203.
- Chen, Y., 2017, Automatic microseismic event picking via unsupervised machine learning: *Geophysical Journal International*, **212**, 88–102.
- Maeda, N., 1985, A method for reading and checking phase time in auto-processing system of seismic wave data: *Zisin (Journal of the Seismological Society of Japan. 2nd ser.)*, **38**, 365379.
- Noordam, J., W. van den Broek, and L. Buydens, 2002, Multivariate image segmentation with cluster size insensitive fuzzy c-means: *Chemometrics and Intelligent Laboratory Systems*, **64**, 65–78.
- Pedrycz, W., 1996, Conditional fuzzy c-means: *Pattern Recogn. Lett.*, **17**, 625–631.
- Sleeman, R., and T. V. Eck, 1999, Robust automatic p-phase picking: an on-line implementation in the analysis of broadband seismogram recordings: *Physics of the Earth and Planetary Interiors*, **113**, 265275.
- Zadeh, L. A., 1977, Fuzzy sets and their application to pattern classification and clustering analysis, *in* Classification and clustering: Elsevier, 251–299.