Automatic Semblance Picking by a Bottom-up Clustering Method

Yuqing Chen[1] and Gerard Schuster[1]
[1]King Abdullah University of Science and Technology (KAUST)

ABSTRACT

Semblance picking is an important but tedious labor-intensive processing procedure in the petroleum industry. For a large 3D dataset, this task becomes extremely time-consuming. In this paper, we present an automatic semblance picking technique based on the K-means clustering algorithm. K-means clustering method can automatically partition different clusters of energy in the semblance spectrum into different groups. The centroid of each group is the automatically picked semblance point. A synthetic and field data example is shown in this paper to illustrate the effectiveness of this method.

INTRODUCTION

Velocity analysis is one of the most important routines of seismic data processing (Yilmaz, 1987), which relies on picking the maximum energy values in the semblance spectrum. However, semblance picking is often manually performed and extremely time-consuming for large data sets. For a large 3D dataset, it may take a geophysicist days or weeks to finish the work. To mitigates this problem we propose that the semblance spectrum can be picked by an unsupervised cluster analysis.

Cluster analysis is the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity (Jain, 2010). It is similar to logistic regression in grouping data into different classes, except the training data do not have to be labeled and there can be many different clusters. Thus, a clustering algorithm is an unsupervised learning method that avoids the cost of supervised labeling of large data sets.

One of the most widely used clustering algorithms is the K-means cluster method (Steinhaus, 1956; Ball and Hall Dj, 1965; MacQueen et al., 1967; Lloyd, 1982), which is simple and very effective in machine learning studies. The next section describes the bottom-up K-means clustering method.

THEORY OF K-MEANS CLUSTERING

Let $X = \{x^{(i)}\}, i = 1, 2, \ldots N$ be the set of $D-$dimensional feature vectors to be clustered into a set of K clusters, with the points in $C = \{c_k, k = 1, 2 \ldots K\}$. The mean centroid point of the k^{th} cluster is defined as

$$\mu_k = \frac{1}{N_k} \sum_{X \in c_k} x_i, \tag{1}$$

The sum of the squared distances between the k^{th} centroid point μ_k and the points in the k cluster is given as

$$d_k = \sum_{X \in c_k} ||x_i - \mu_k||^2. \tag{2}$$

The goal is to find the optimal distribution of cluster points c_k such that the objective function ϵ is minimized:

$$\epsilon = \sum_{k=1}^{K} \sum_{X \in c_k} ||x_i - \mu_k||^2. \tag{3}$$

There are two groups of clustering algorithms: hierarchical and partitional. For the hierarchical algorithm, there are two strategies.

- The bottom-up approach starts with each point being its own distinct cluster, and iteratively finds a smaller number of clusters that contain the points.
- The top-down approach, where all points belong to one super cluster, and smaller clusters are iteratively computed.

The partitioning clustering algorithm sets the number of clusters and finds the optimal arrangement of points

that minimize the misfit function. In our case, we use a top-down hierarchical approach where the number of clusters increases and we choose the cluster where the residual reduction becomes small.

There are four steps to the K-means clustering algorithm where the number of clusters is specified.

1. Select an initial configuration of K clusters. Compute the centroid point μ_k for each cluster.

2. Compute the distances $||x^{(i)} - \mu_k||$, and reassign points to the cluster of the centroid they are closest to. If the elements in each vector have different units then they should be normalized by the standard deviation.

3. Compute new centroids from the new clusters.

4. Repeat steps 2 and 3 until convergence. This procedure can be adapted to the hierarchical approach where an outer loop can be added where the number of clusters is either increased or decreased.

In this paper we will use a bottom-up approach where the number of clusters is increased at each iteration of the outer loop.

THEORY OF SEMBLANCE ANALYSIS

The goal of semblance analysis is to find the stacking velocity V_{stack} as a function of two-way zero-offset traveltime T_o such that the hyperbolas

$$t(x, V_{stack}) = \sqrt{\frac{x^2}{V_{stack}^2} + T_o^2}, \tag{4}$$

best fit the reflections in a CMP gather. Here, the midpoint coordinate is $x_o = (x_r - x_s)/2$ and the offset between the source and receiver is $x = x_r - x_s$. The primary reflections in each trace $d(x, t)$ approximately follow the traveltimes in equation 4 for the correct values of V_o and T_o, so the stacking procedure uses the semblance equation given by (Yilmaz, 1987; Luo and Hale, 2012)

$$S(T_o, V_{stack}) = \frac{\sum_x d(x, t(x, V_{stack}))}{\sum_x d(x, t(x, V_{stack}))^2}, \tag{5}$$

where $S(V_{stack}, T_o)$ is the semblance spectrum associated with the midpoint coordinate $x_o = (x_r - x_s)/2$. To generate a smooth semblance panel, the trace amplitudes are typically stacked together along a half-period wide strip centered about the moveout curve described in equation 4. A CMP gather is shown in Figure 1a and it's semblance spectrum is shown in Figure 1b, which can be seen as a combination of many small energy clusters. Therefore, hierarchical cluster analysis is used to identify different clusters of stacking velocities with high semblance values in the semblance spectrum. The centroids of each cluster are then connected to form the stacking velocity curve.

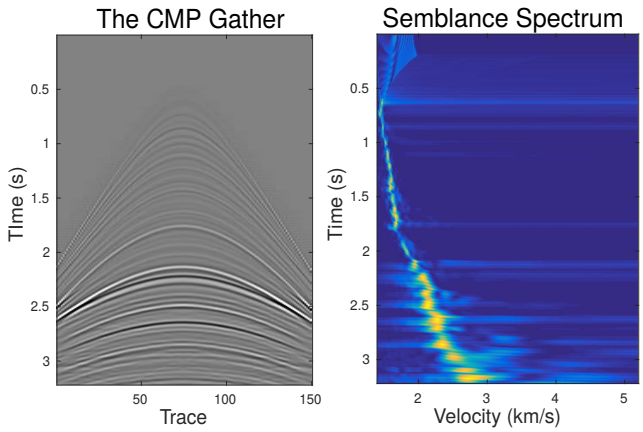

Figure 1: (a) A CMP gather and (b)it's semblance spectrum calculated by semblance analysis.

WORK FLOW

1. Threshold the semblance spectrum and only keep the points which have relatively large semblance values. Those points are viewed as the eligible points for clustering analysis.

2. Specify the initial number of clusters K. Assign the centroid point of the k cluster to be $(V_{stack}^{(k)}, T_o^{(k)})$. These K points are selected at equi-spaced T_o and V_{stack} intervals along an initial semblance curve.

3. Apply K-means clustering for several iterations until a stable group of K clusters has emerged.

4. Increment the value of K by 1 and repeat step 2. The value of K is incrementally increased until the residual misfit in equation 3 does not decrease by a significant amount.

5. The centroids of the final cluster are used to form the final semblance curve.

NUMERICAL RESULTS

The effectiveness of semblance picking by clustering analysis is now demonstrated with synthetic data generated from the Marmousi model and a Gulf of Mexico field data.

Marmousi Model

The automatic semblance picking method is now tested on the Marmousi Model. Figure 2 shows the true velocity model used for generating the observed data. A Ricker wavelet with a peak-frequency of 15 Hz is used as the source wavelet. A fixed-spread acquisition geometry is assumed where there are 400 sources evenly distributed on the surface at an interval of 20 m. The data are recorded by 800 receivers for each shot uniformly distributed every 10 m on the surface. For the CMP gathers at the location where the velocity model has most flat layers, the reflections in the CMP gather approximately follow

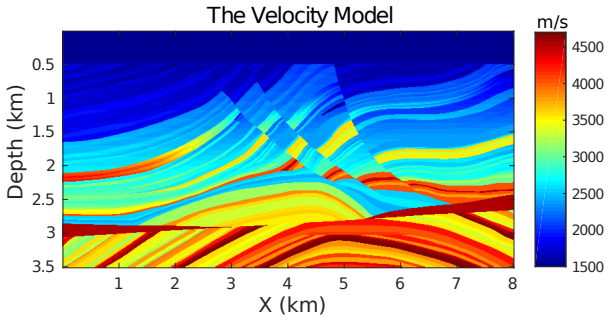

Figure 2: The marmousi velocity model.

the traveltimes in equation 4 for the correct values of V_o and T_o. By this procedure the energy clusters are more concentrated which result in a fairly accurate pick of the semblance curve as shown by the black curve in Figures 3b and 6b. The red stars and red crosses in Figures 3b and 6b are the initial and inverted centroids of each cluster. The centroids of the clusters are then connected by a black line to form the stacking velocity curve. However, for the CMP gathers at the location where the velocity model has complicated structure, the energy clusters are more diffused in the semblance spectrum, which results in a loss of picking accuracy results as shown in Figures 4 and 5.

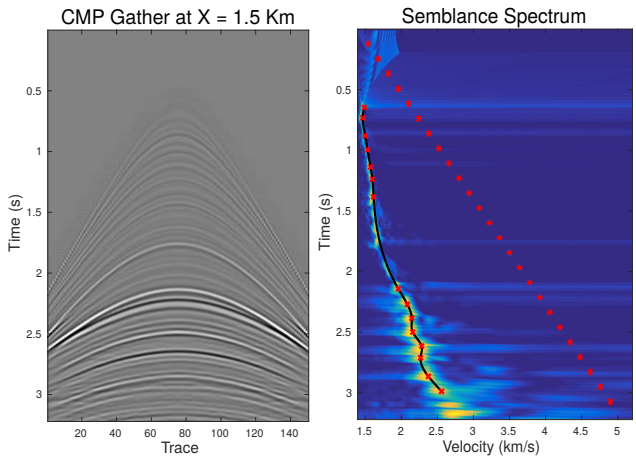

Figure 3: (a) A CMP gather at x = 1.5 km and (b)it's semblance spectrum with picked semblance curve. The red stars and red crossings are the initial and inverted centroid of each cluster. The black curve is automatic picked semblanced curve.

GOM Data

The proposed methods is tested on a 2D marine data set. There are 100 shots with a shot interval of 37.5 m, and each shot is recorded by a 6 km long cable with 480 receivers spaced with a 12.5 m receiver interval. The receiver offset from the source is 200 m by a 25-Hz Wiener

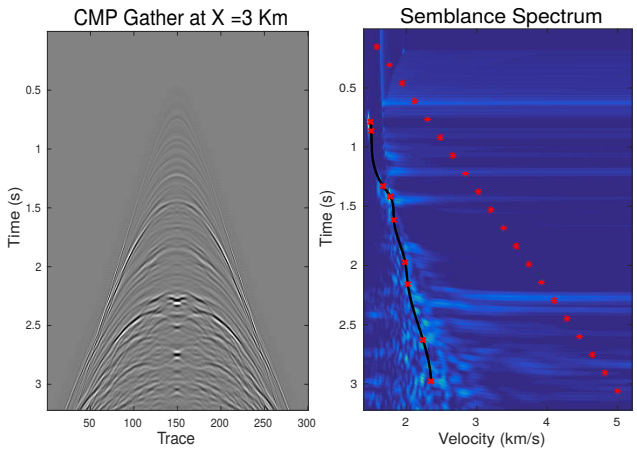

Figure 4: (a) A CMP gather at x = 3 km and (b)it's semblance spectrum with picked semblance curve. The red stars and red crossings are the initial and inverted centroid of each cluster. The black curve is automatic picked semblanced curve.

filter (Boonyasiriwat et al., 2009). Figures 7 and 8 shows the CMP gathers at the location of $x = 2.45$ km and $x = 6.1$ km, respectively. The energy clusters at the shallower parts of the semblance spectrums are more diffused. Therefore the inverted cluster's centroid result in a loss of accuracy. However, the accuracy in picking improves for the deeper parts where the energy clusters are more concentrated. The red stars and red crosses in Figures 7b and 8b are the initial and inverted centroid of each cluster. The centroids of each cluster are then connected to form the stacking velocity curve which is indicated by the black curve.

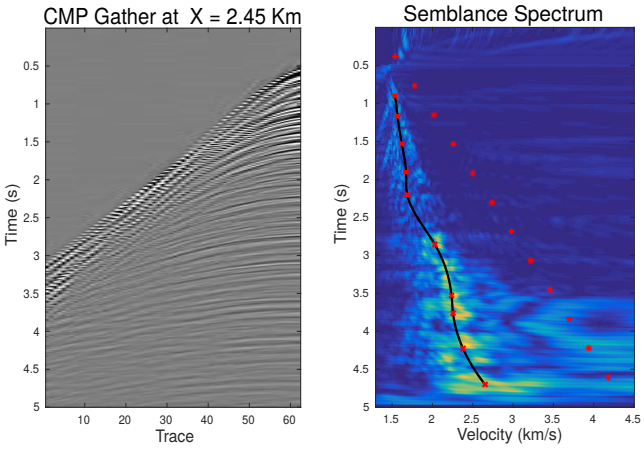

Figure 7: (a) A CMP gather at x = 2.45 km and (b)it's semblance spectrum with picked semblance curve. The red stars and red crossings are the initial and inverted centroid of each cluster. The black curve is automatic picked semblanced curve.

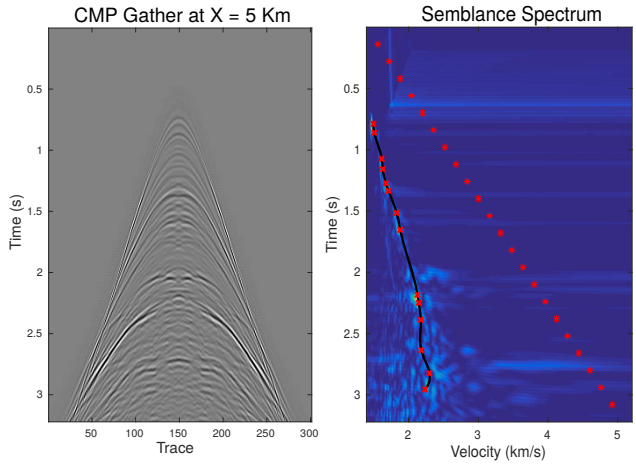

Figure 5: (a) A CMP gather at x = 5 km and (b)it's semblance spectrum with picked semblance curve. The red stars and red crossings are the initial and inverted centroid of each cluster. The black curve is automatic picked semblanced curve.

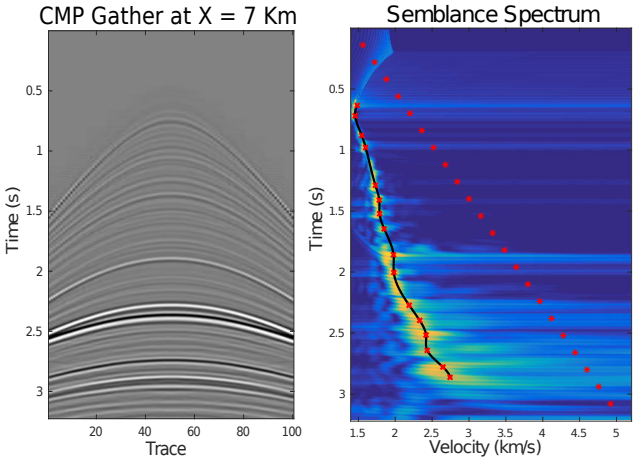

Figure 6: (a) A CMP gather at x = 7 km and (b)it's semblance spectrum with picked semblance curve. The red stars and red crossings are the initial and inverted centroid of each cluster. The black curve is automatic picked semblanced curve.

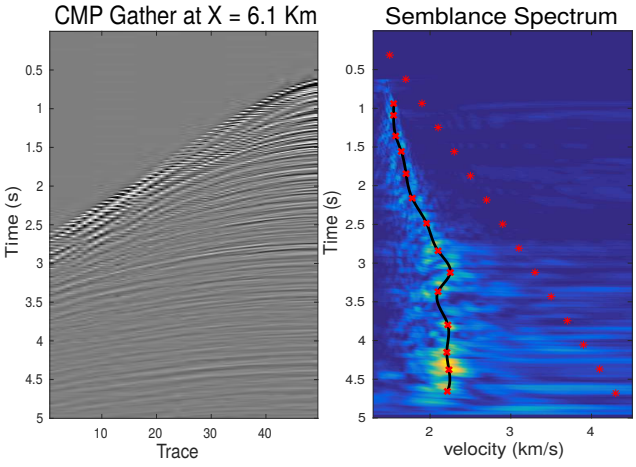

Figure 8: (a) A CMP gather at x = 6.1 km and (b)it's semblance spectrum with picked semblance curve. The red stars and red crossings are the initial and inverted centroid of each cluster. The black curve is automatic picked semblanced curve.

SUMMARY

An automatic semblance picking method is presented which use clustering analysis to compute for the centroid of each energy cluster in a semblance spectrum. The centroids of the clusters are then connected to form the stacking velocity curve. Both synthetic and real data tests show that the automatic picking method performs well when the energy clusters in the semblance spectrum are more concentrated. However, the picking accuracy decreases when the energy cluster become more diffused. Therefore, to get a fairly reliable stacking-veloity curve, more constraints are needed.

REFERENCES

Ball, G., and I. Hall Dj, 1965, A novel method of data analysis and pattern classification. isodata, a novel method of data analysis and pattern classification. tch. report 5ri, project 5533.

Jain, A. K., 2010, Data clustering: 50 years beyond k-means: Pattern recognition letters, **31**, 651–666.

Lloyd, S., 1982, Least squares quantization in pcm: IEEE transactions on information theory, **28**, 129–137.

Luo, S., and D. Hale, 2012, Velocity analysis using weighted semblance: Geophysics, **77**, U15–U22.

MacQueen, J., et al., 1967, Some methods for classification and analysis of multivariate observations: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, **1**, 281–297.

Steinhaus, H., 1956, Sur la division des corp materiels en parties: Bull. Acad. Polon. Sci, **1**, 801.

Yilmaz, O., 1987, Seismic data processing: Investigations in geophysics.