UTILIZING MODIFIED GUSTAFSON-KESSEL ALGORITHM TO ESTIMATE LOG DATA FROM SEISMIC ATTRIBUTES

AMIR NAGHIBI[1] and M. ALI RIAHI[2]

[1] *Amirkabir University of Technology, Iran. amir_naqibi87@yahoo.com*
[2] *Institute of Geophysics, University of Tehran, P.O. Box 14155-6, Tehran, Iran. mariahi@ut.ac.ir*

(Received September 21, 2010; revised version accepted July 6, 2011)

ABSTRACT

Naghibi, A. and Riahi, M.A., 2011. Utilizing modified Gustafson-Kessel algorithm to estimate log data from seismic attributes. *Journal of Seismic Exploration*, 20: 347-356.

One of the most effective methods for analyzing seismic attributes is Fuzzy C-Means (FCM) clustering. By extension of FCM, standard Gustafson-Kessel (G-K) algorithm is derived, which is a powerful tool for clustering analysis. However, G-K algorithm suffers from some shortcomings like singularity of the covariance matrix. By using different techniques for estimating covariance matrix, we can improve the performance of G-K algorithm and lessen the impacts of such pitfalls. Recently, standard G-K algorithm was used for estimation of log data from seismic attributes. In this article we applied the same procedure but instead of standard G-K algorithm we utilized modified G-K algorithm in which we employed two new formulas for further precise estimation of covariance matrix. Because of drawbacks of standard G-K algorithm due to covariance matrix, increasing the number of clusters is not possible in this estimation. Therefore, utilizing recent new techniques have assisted to overcome the drawbacks and improve log data estimation.

KEY WORDS: modified G-K algorithm, log data, seismic attributes, standard G-K algorithm.

INTRODUCTION

Clustering is the process of assigning a set of data into clusters so that the data in the same cluster are similar to some extent. Depending on the nature of the data, choosing a distance measurement [distance between the data set (Z_j) and cluster prototype (V_i)] is the critical step in most clustering procedures because the similarity of two elements will depend on this component.

0963-0651/11/$5.00

In fuzzy clustering, unlike the hard clustering, data components can belong to more than one cluster (Yang et al., 2006).

One of the most famous algorithms in fuzzy clustering is Fuzzy C-Means algorithm. FCM algorithm is a branch of criterion function minimization based on the Euclidean distance. Therefore this algorithm can be used to determine data classes with the spherical shapes merely (Hsiang-Chuan et al., 2009). The general form of FCM algorithm is

$$J(Z,U,V) = \sum_{i=1}^{K} \sum_{j=1}^{N} (\mu_{ij})^m D_{ij}^2 \quad , \tag{1}$$

where $Z \in \mathbb{R}^{nxN}$ is the data set, U is the partition matrix of the data Z and $U = [\mu_{ij}]$. μ_{ij} is the membership degree of data object Z_j in cluster i and it must meet the following conditions

$$\sum_{i=1}^{K} \mu_{ij} = 1, \forall \ j = 1,2,....,n \quad , \tag{2}$$

D_{ij} is the Euclidean distance between Z_j and V_i

$$D_{ij} = \| Z_j - V_i \| \quad . \tag{3}$$

The variable m is a weighting exponent, $m > 1$; it controls the fuzziness of clustering and K is the number of clusters (Hsiang-Chuan et al., 2009).

STANDARD G-K ALGORITHM

Gustafson and Kessel introduced G-K algorithm in 1979, which is derived from Fuzzy C-Means algorithm. To overcome the shortcoming due to Euclidean distance and also to detect clusters of different geometrical shapes, the distance measure in G-K algorithm is considered based on the Mahalanobis distance (MD). Thus, against the FCM algorithm, this adaptive distance norm can be used to detect data classes with non-spherical shapes (Hsiang-Chuan et al., 2009). G-K algorithm objective function is defined as

$$J(Z,U,V,A_i) = \sum_{i=1}^{K} \sum_{j=1}^{N} (\mu_{ij})^m D_{ij}^2 \quad . \tag{4}$$

In this method each cluster has its own norm-inducing matrix A_i, which yields the following distance

$$D^2_{ijA_i} = (Z_j - V_i)^T A_i (Z_j - V_i) \quad . \tag{5}$$

However, the objective function (J) cannot be minimized directly with respect to A_i. A_i must be constrained somehow in order to obtain a feasible solution. Constraining the determinant of A_i is a common way for this purpose, as follows

$$\rho_i = \| A_i \| \quad , \tag{6}$$

where ρ_i is the cluster volume of the i-th cluster (usually $\rho_i = 1$).

Shape and orientation of the i-th cluster is determined by the matrix A_i. By using the Lagrange multiplier method, the following expression for A_i is obtained

$$\det(F_i)^{1/n} F_i^{-1} \quad , \tag{7}$$

where F_i is the fuzzy cluster covariance matrix of the i-th cluster.

Standard G-K algorithm sequential stages can be listed as follows:

Step 1: Choosing the number of clusters K, m-value (usually $m \geq 2$), and convergence error, $\varepsilon > 0$. Z should be given and randomly initialize the partition matrix $U^{(0)}$ (Babuska et al., 2002).

Step 2: Finding the cluster prototype:

$$V_i^{(l)} = \sum_{j=1}^{N} (\mu_{ij}^{(l-1)})^m Z_j \Big/ \sum_{j=1}^{N} (\mu_{ij}^{(l-1)})^m \; , \quad 1 \leq i \leq K \; . \tag{8}$$

Step 3: Generating the cluster covariance matrix:

$$F_i = \sum_{j=1}^{N} (\mu_{ij}^{(l-1)})^m (Z_j - V_i^{(l)})(Z_j - V_i^{(l)})^T \Big/ \sum_{j=1}^{N} (\mu_{ij}^{(l-1)})^m \; , \quad 1 \leq i \leq K \; . \tag{9}$$

Step 4: Calculation of the distance:

$$D^2_{ijA_i} = (Z_j - V_i^{(l)})[\rho_i \det(F_i)^{1/n} F_i^{-1}](Z_j - V_i^{(l)})^T \; , \; 1 \leq i \leq K, \; 1 \leq j \leq N. \tag{10}$$

Step 5: Updating the partition matrix:

for $1 \leq j \leq N$

if $D_{\overline{ijA_i}} > 0$ for $1 \leq \overline{i} \leq K$

$$\mu_{ij}^{(l)} = 1 / \sum_{h=1}^{K} (D_{\overline{ijA_i}} / D_{\overline{hjA_h}})^{2/(m-1)} \quad , \tag{11}$$

otherwise

$$\mu_{ij}^{(l)} = 0 \text{ if } D_{\overline{ijA_i}} > 0, \text{ and } \mu_{ij}^{(l)} \in [0,1]$$

$$\text{with} \sum_{i=1}^{K} \mu_{ij}^{(l)} = 1 \quad , \tag{12}$$

until $\| U^{(l)} - U^{(l-1)} \| < \varepsilon$.

METHODOLOGY

In early 1970's, seismic attributes were introduced and now they are critical tools in seismic interpretation purposes. Seismic attributes are elicited from seismic data. They are the constituents of geological and geophysical information and can be used for lithological and petrophysical prediction of reservoir properties. The authenticity of the interpretive use of seismic attributes depends on the discrimination of a set of them. Logical combinations of seismic attributes can be used to determine different lithologies and reservoir properties from seismic data (Taner, 2001).

Nowadays, in addition to these extensive applications of seismic attributes, log data estimation is considered as another important one. This can be achieved through analyzing seismic attributes by clustering methods. Recently, FCM and G-K (standard) algorithms applied this purpose; their results are shown in Figs. 1 and 2. G-K algorithm has shown a better performance than FCM; this claim is supported by comparing Figs. 1 and 2.

As mentioned before, because of standard G-K algorithm's drawback due to covariance matrix, increasing the number of clusters is not possible in this estimation. To overcome this problem and to improve the estimation of log data, we used the modified G-K algorithm. Two new techniques are introduced for computing covariance matrix in modified G-K, which make it possible to increase the number of clusters.

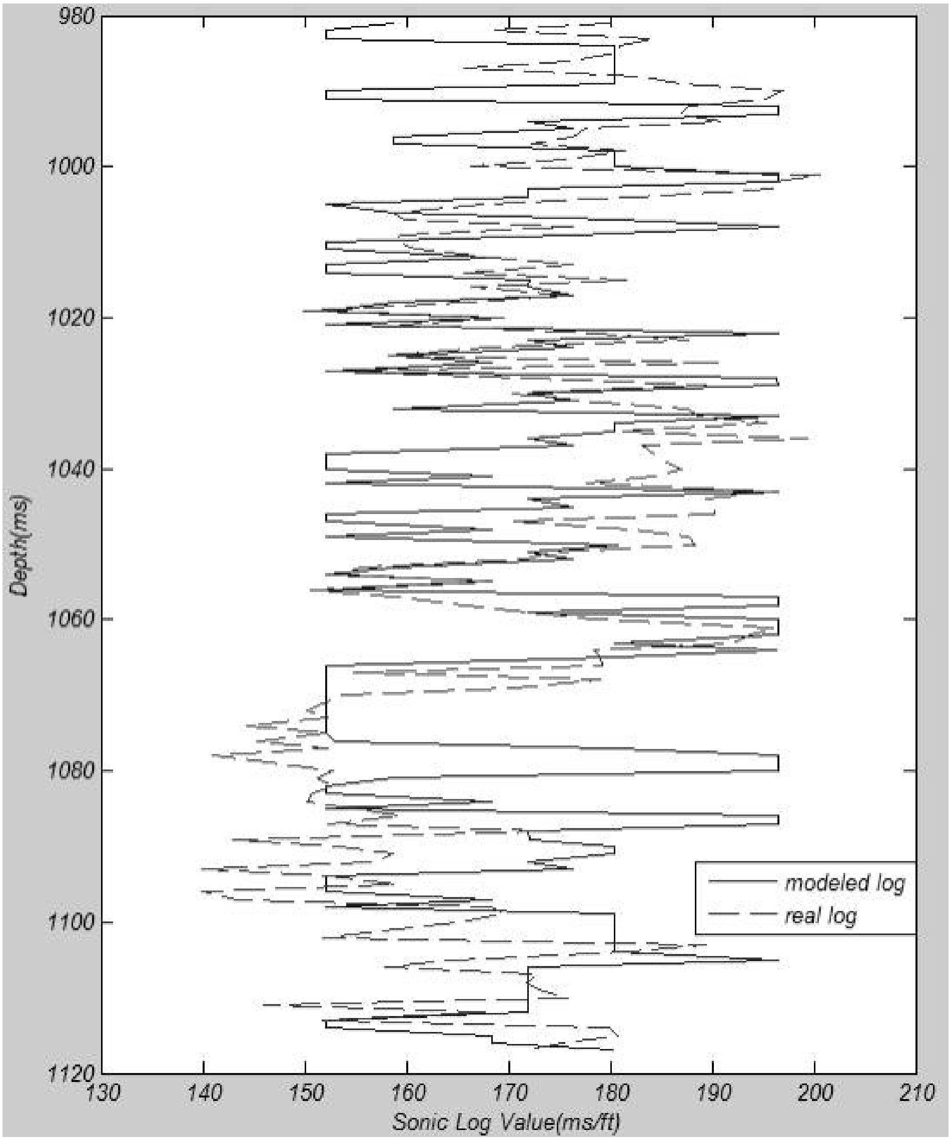

Fig. 1. Estimation of log data from seismic attributes using FCM algorithm (Eftekharifar and Riahi, 2009).

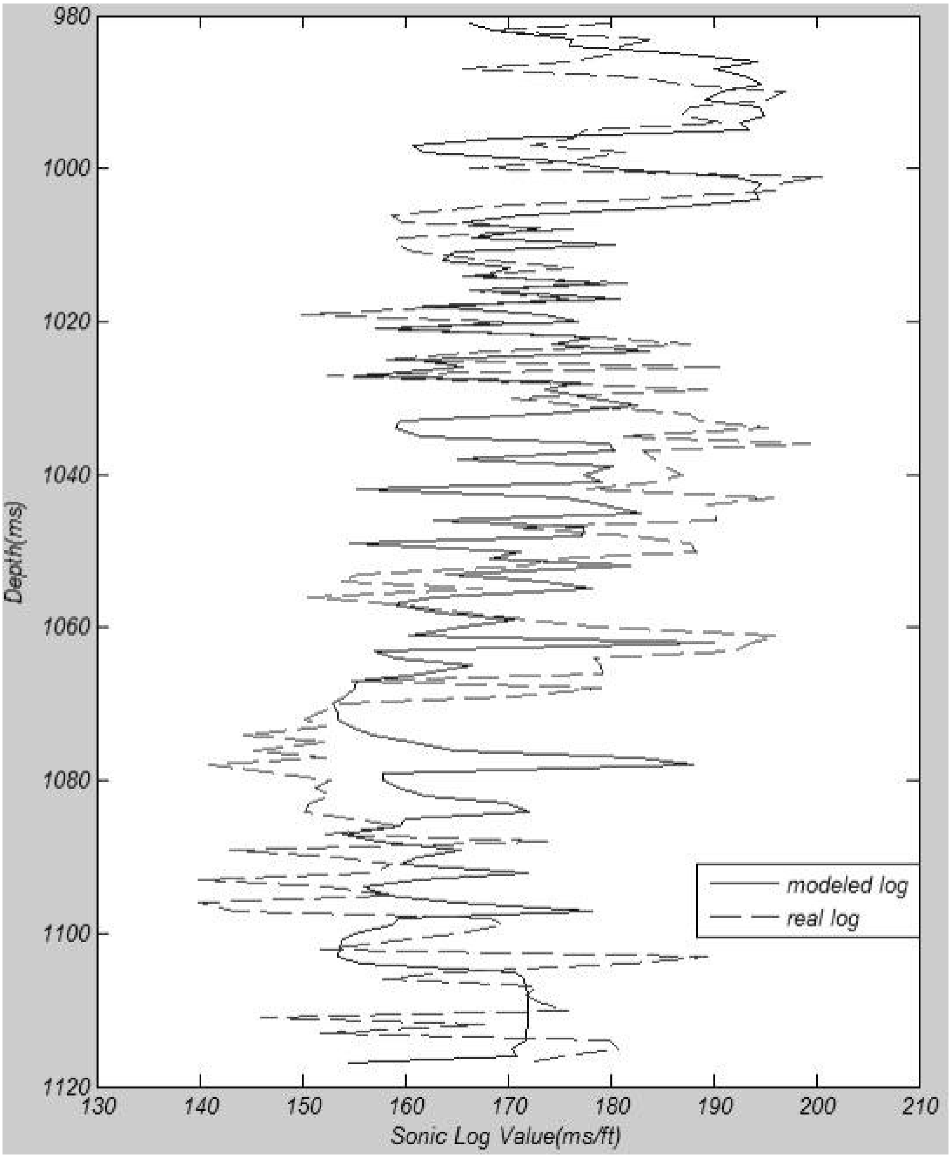

Fig. 2. Estimation of log data from seismic attributes using standard G-K algorithm (Eftekharifar and Riahi, 2009).

MODIFIED G-K ALGORITHM

When the number of data samples is small or when the data within a cluster is linearly correlated, the covariance matrix becomes singular. In such a case, the following formula cannot be adapted

$$\det(F_i)^{1/n}F_i^{-1} \quad , \tag{13}$$

and cannot be inverted to compute the norm-inducing matrix (step 4) either. By constraining the ratio between the maximal and minimal eigen values which should be smaller than some predefined threshold and then reconstructing covariance matrix using the following formula, we can avoid singularity problem,

$$F = \Psi X \Psi^{-1} \quad , \tag{14}$$

where X is a diagonal matrix containing the limited eigen values and Ψ is a matrix whose columns are the corresponding eigen vectors. In fact, the shape and orientation of the clusters are set out with the eigen values and eigen vectors (Babuska et al., 2002).

Another technique which we used to improve the performance of covariance matrix was adding a scaled identity matrix to the covariance matrix. When the number of data points in a cluster becomes too low, the computed covariance matrix is not a valid estimate of the underlying data distribution. This technique avoids overfitting the problem and is based on the following formula

$$F_i^{new} = (1 - \delta)F_i + \delta\det(F_0)^{1/n}I \quad , \tag{15}$$

where $\delta \in [0,1]$ is a tuning parameter and F_0 is the covariance matrix of the whole data set. Depending on the value of δ, the clusters are forced to have a more or less equal shape. When δ is 1, all the covariance matrices are equal ($\det(F_0)^{1/n}I$) and have the same size which of course limits the possibility of the algorithm to properly identify clusters (Babushka et al., 2002).

The number of clusters does not control the value of F because it depends on the entire data set. Also the volumes of F_i decrease with increasing number of clusters. This indicates that the clusters become rounder by increasing number of clusters (Babushka et al., 2002).

With respect to the clarified formulas, step 3 in G-K algorithm computing, becomes as follows:

Step 3: Generating covariance matrix

$$F_i = \sum_{j=1}^{N} (\mu_{ij}^{(l-1)})^m (Z_j - V_i^{(l)})(Z_j - V_i^{(l)})^T / \sum_{j=1}^{N} (\mu_{ij}^{(l-1)})^m \ , \quad 1 \leq i \leq K \ . \qquad (16)$$

Adding a scaled identity matrix

$$F_i = (1 - \delta)F_i + \delta \det(F_0)^{1/n} I \ . \qquad (17)$$

Extracting eigen values λ_{ij} and eigen vectors φ_{ij} from F_i.

Assigning $\lambda_{i\,max} = \max_j \lambda_{ij}$ and setting:

$$\lambda_{ij} = \lambda_{i\,max}/\beta \quad \forall j \text{ for which } \lambda_{i\,max}/\lambda_{ij} > \beta \ , \qquad (18)$$

Reconstructing F_i

$$F_i = [\varphi_{i1} \ldots\ldots \varphi_{in}] \operatorname{diag}(\lambda_{i1},\ldots,\lambda_{in})[\phi_{i1} \ldots\ldots \phi_{in}]^{-1} \ . \qquad (19)$$

The result of modeling using modified G-K algorithm is shown in Fig. 3, where the modeled log (blue) and the real log(red) are compared, like Figs. 1 and 2. The correlation between real log values and modeled log values indicates that clustering by modified G-K algorithm yields better result than FCM and standard G-K algorithms, as the correlation coefficient for FCM and standard G-K are 36% and 54% respectively, but this magnitude for the modified G-K is 98%.

CONCLUSIONS

Nowadays estimation of log data is considered as one important application of seismic attributes. Log properties can be estimated throughout the 3D seismic cube using a number of complex seismic attributes and limited log data. In this paper, we tried to improve this application. Recently, this estimation was performed using standard G-K algorithm. Thus, we used modified G-K algorithm to enhance the performance of standard G-K and also to yield better results than standard G-K.

Two new techniques in covariance matrix computing, are used in modified G-K by which the singularity of covariance matrix and overfitting problems are eliminated almost completely. Due to these problems, an increasing number of clusters is not feasible in estimating log data using standard G-K. This purpose is accessible when we use modified G-K. Increasing the number of clusters would result in yielding better outputs as correlation coefficient is greatly increased. Correlation coefficient for standard G-K was about 54% whereas this value was 98% for modified G-K.

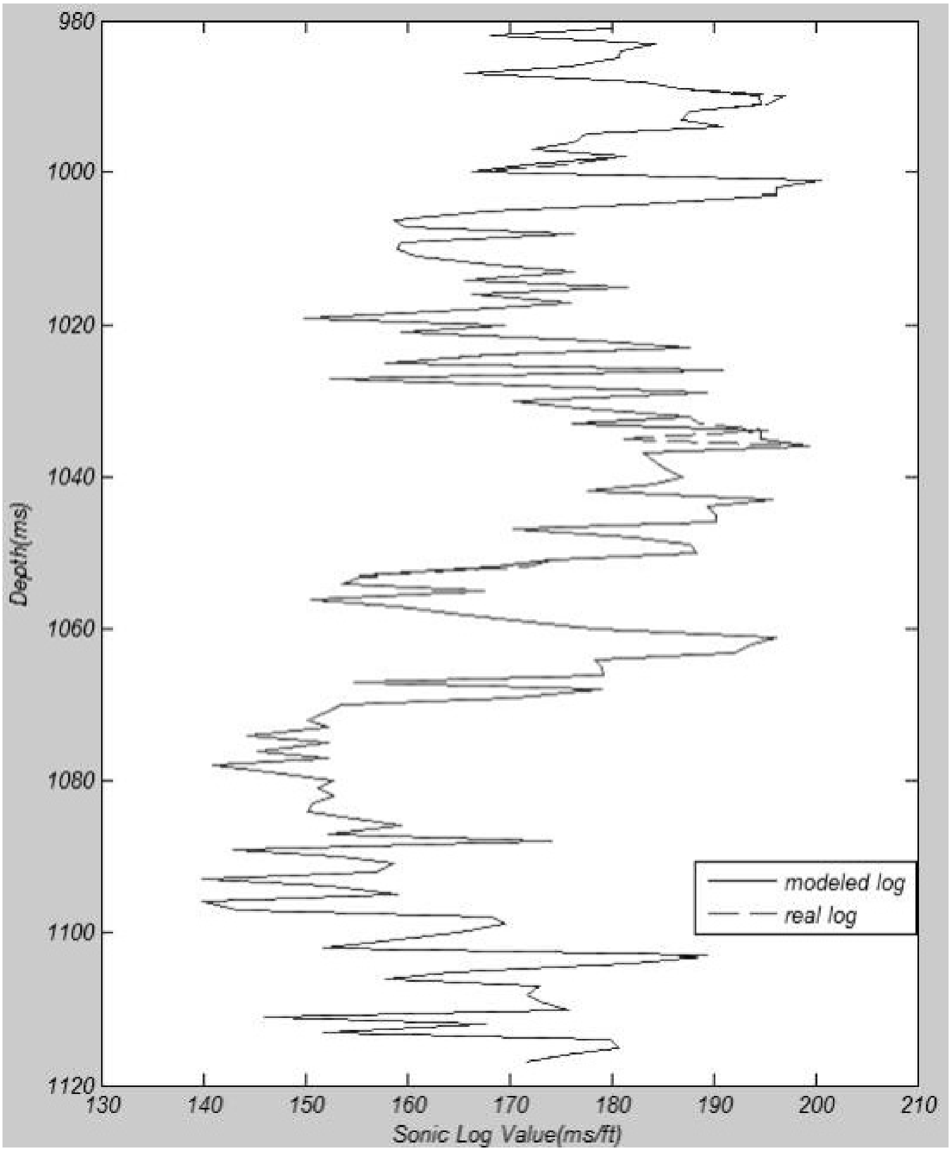

Fig. 3. Estimation of log data using modified G-K algorithm.

REFERENCES

Babuska, R., Van der Veen, P.J. and Kaymak, U., 2002. Improved Covariance Estimation for Gustafson-Kessel Clustering. IEEE, Honolulu, Hawaii.

Eftekharifar, M., Riahi, M.A. and Kharrat, R., 2009. Integration of Gustafson-Kessel Algorithm and Kohonen's self-organizing maps for unsupervised clustering of seismic attributes. J. Seismic Explor., 18: 315-328.

Liu, H.-C., Yih, J.-M., Lin, W.-C. and Liu, T.-S;, 2009. Fuzzy C-Means algorithm based on PSO and Mahalanobis Distance. Internat. J. Innovat. Comput., Informat. Control, 5: 5033-5040.

Taner, M.T., 2001. Seismic Attributes. CSEG Recorder, 9: 48-56.

Yang, P., Yin, X. and Zhang, G., 2006, Seismic Data Analysis Based on Fuzzy Clustering, IEEE, Gulin, China.