Journal of Applied Geophysics 102 (2014) 10–20

Contents lists available at ScienceDirect

Journal of Applied Geophysics

journal homepage: www.elsevier.com/locate/jappgeo

# Mixture decompositions and lithofacies clustering from wireline logs

Y. Zee Ma [a], Hongliang Wang [b,*], Jason Sitchler [a], Omer Gurpinar [a], Ernest Gomez [a], Yating Wang [a]

[a] Schlumberger, Denver, CO 80202, United States
[b] Energy Resource School, China University of Geosciences, Beijing 100083, China

ARTICLE INFO

*Article history:*
Received 18 July 2013
Accepted 19 December 2013
Available online 29 December 2013

*Keywords:*
Wireline logs
Rock properties
Mixture decomposition
Principal component analysis (PCA)
Rotating principal components
Hierarchical or multilevel clustering

ABSTRACT

Finite mixture models provide useful methods for modeling a wide variety of natural phenomena. Decomposing a mixture of distributions, however, has many difficulties, including separability of the component distributions, determination of the number of components, predictability of the clusters with realistic spatial patterns, and linkages between the component density functions and underlying physical processes. In this article, we use principal component analysis (PCA) to synthesize multiple wireline logs and decompose mixture populations for lithofacies clustering. The principal components of these logs characterize the rock physics; some of them contain essential information of lithofacies while others represent less relevant information or noise. In many cases, clustering based on one component is effective for decomposing the mixture and classifying lithofacies, although rotating principal components is often necessary to improve the lithofacies discrimination. In more complicated cases, PCA and direct mixture decomposition using histogram can be cascaded to decompose finite mixture models of a rock property. The proposed methodology combines the delicacy of probability theory and simplicity of linear transforms to classify lithofacies.



## 1. Introduction

Finite mixture models based on probability distributions provide a theoretical approach to modeling of a wide variety of random phenomena (McLachlan and Peel, 2000). Mixture models underpin a variety of statistical methods and play a useful role in neural networks (Bishop, 1995). Unlike traditional clustering analysis, the mixture models are based on decompositions of a histogram for classifications (Scott, 1992; Silverman, 1986). Although mixture models provide a convenient framework, decomposing a mixture of distributions can be difficult in several fronts, including separability of the component distributions, determination of the number of components, predictability of the clusters with realistic spatial patterns for geoscience applications, and linkages between the component density functions and underlying physical processes.

In resource characterization, traditional ad hoc methods, such as using cutoff or gated-logic techniques, are still commonly used for classifying rock types or lithofacies. These methods are statistically biased (Ma, 2011; Titterington et al., 1985), which is illustrated by an example of applying cutoffs on a Gamma Ray (GR) histogram (Fig. 1). The histogram shows two distinct modes at 62 and 120 API. In the cutoff method, sandy channel facies are generated from low GR values, shale-dominated floodplain from high GR values, and crevasse and splay, typically a mixture of sand and shale, from intermediate GR values (Fig. 1b). Applying cutoffs is rather arbitrary, creating demarcating "walls" in the lithofacies-component histograms. Other methods were proposed for clustering lithofacies using wireline logs, including clustering analysis (Wolff and Pelissier-Combescure, 1982), and neural network, with or without using principal component analysis (PCA) (Ma, 2011; Wang and Carr, 2012). However, these methods do not directly address the mixture decomposition of frequency distributions.

Although mixture decomposition has not drawn much attention of geoscientists, researchers have known the importance of modeling mixtures. Zhu and Journel (1991) simulated a mixture of two normal distributions and found that traditional geostatistical methods could go astray in a mixture of populations. This implies the importance of decomposing mixture populations prior to geostatistical modeling of rock properties. They also pointed out that the difficulty of inference was the main obstacle for broad applications of mixture models to rock physics and resource evaluation. Indeed, whereas the study by Zhu and Journel (1991) is forward modeling of mixtures, decomposition of a mixture is an inverse problem, of which inference is more challenging. Although several methods have been proposed for decomposing a mixture using histogram, the inference of mixture decompositions to petrophysical properties is lacking. Fig. 2 shows an example of mixture decomposition for clustering lithofacies using the kernel density technique (Chang et al., 2002; Hardle et al., 2004). While the component frequency distributions are quasi-normal – quite reasonable, the predicted clusters are not satisfactory because of the unrealistic randomness in spatial patterns of the rock formation.

In this article, we present methods for mixture decomposition based on multivariate analysis of wireline logs. We use principal component analysis (PCA), combined with geologic interpretations, to address inference problems of the mixture models for classifying lithofacies. We first review the finite mixture-modeling method, followed by presenting

* Corresponding author.

0926-9851/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jappgeo.2013.12.011

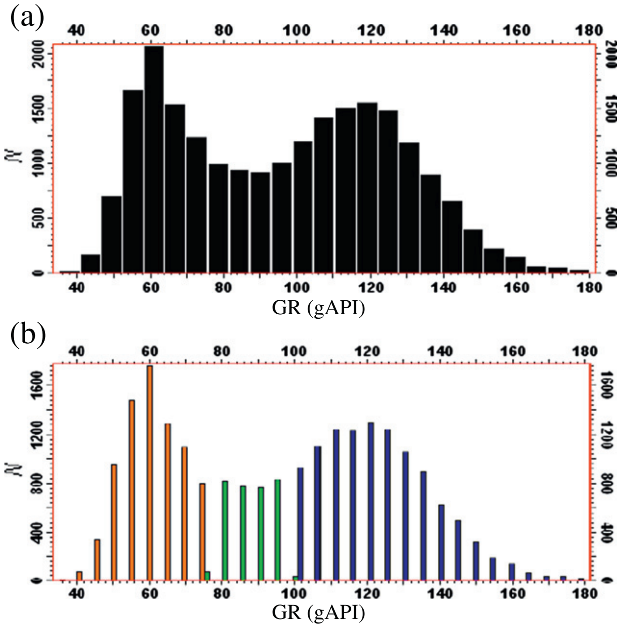


**Fig. 1.** (a) Histogram of Gamma Ray (GR) log with 21,854 samples from a Cretaceous formation in a Rocky Mountain basin. (b) Decomposition of the histogram in this figure into 3 histograms based on the GR cutoffs. Color codes: orange, channel facies; green, crevasse-splay; and black, overbank.

mixture frequency distributions that are commonly encountered in wireline logs, including univariate and multivariate frequency distributions. Typically, the different lithofacies exhibit significant overlaps in individual wireline logs, but the bias due to the overlaps can be mitigated by using two or more wireline logs. Because PCA enables synthesizing the information from multiple logs, it is used to facilitate introducing geological prior knowledge for clustering lithofacies. Rotating a principal component allows using several components and weighing their relative importance. In more complicated cases, PCA and direct mixture decomposition are cascaded to decompose the mixture.

## 2. Finite probabilistic mixture models

Finite mixture models can be formulated as a probabilistic distribution that is a convex combination of several component probabilistic distributions (McLachlan and Peel, 2000), such as:

$$f(x) = \sum_{i=1}^{g} \omega_i \ f_i(x) \tag{1}$$

where $g$ is the number of components, $f(x)$ is the mixture density, $f_i(x)$ is the component densities of the mixture, and $\omega$ represents the mixing weights such as:

$$\omega_i > 0, \text{and } \ \omega_i + \ldots + \omega_g = 1$$
$$f_j(x) = \ > 0, \text{and}$$
$$\Sigma f_j(x) \ dx = 1.$$

For mixtures of univariate normal distributions, Eq. (1) can be expressed as:

$$f(x) = \sum_{i=1}^{g} \omega_i \ N_i(\mu_i; \sigma_i) \tag{2}$$

where $N(\mu_i; \sigma_i)$ is a normal density for each component $i$ and $\omega_i$ is its proportional weight.

When their variances are identical, the component distributions are referred to as homoscedastic; otherwise, they are said heteroscedastic. For example, a mixture of two normal homoscedastic components has the form:

$$f(x) = \omega_1 \phi\left(x; \mu_1, \sigma^2\right) + \omega_2 \phi\left(x; \mu_2, \sigma^2\right) \tag{3}$$

where $\phi(x; \mu, \sigma)$ is a normal density with mean μ and standard deviation σ.

Heteroscedastic mixtures are much more common in practice because of the different variances in the components of mixtures. The GR histogram discussed earlier (Fig. 1a) can be decomposed into three normal or quasi-normal component histograms (Fig. 2a). These include: channel facies having a normal distribution of $N(61.5, 8.6)$, crevasse-splay facies having a normal distribution of $N(84.2, 10.9)$, and overbank facies having a normal distribution of $N(117.4, 19.1)$, respectively. Therefore, the GR log is a mixture of three heteroscedastic components. It is noteworthy that the means of the GR in the channel and overbank facies are slightly different than the two modes because the modes are impacted by the overlapped GR values in the crevasses-splay.

Normal distributions are often used as kernel densities to analyze probability mixture distributions. For instance, the following equation

$$\Delta = |\mu_1 - \mu_2| / \sigma \tag{4}$$

is the Mahalanobis distance between the two homoscedastic normal densities (Titterington et al., 1985). Few studies have been conducted for measuring separability of heteroscedastic densities. Here, we extend the Mahalanobis distance in Eq. (4) to the heteroscedastic case, such as:

$$\Delta = |\mu_1 - \mu_2| / (0.5^*(\sigma_1 + \sigma_2). \tag{5}$$

In other words, the Mahalanobis distance is defined as a function of the means and their standard deviations of the component distributions. The greater the difference in mean between the component distributions and the smaller the difference in standard deviations are, the greater the separation between the component distributions will be. When it is greater than 3, the separation of the two components is generally straightforward. On the other hand, when it is less than 1, separating them can be quite challenging.

The histogram in Fig. 3d shows two distinct modes in the lognormal scale of a resistivity log and the two components are approximately lognormal. The Mahalanobis distance based on Eq. (5) with their means and standard deviations in logarithmic scale from the histogram is 6 (i.e., |0.80 − 0.02| / [(0.18 + 0.08) × 0.5]), and the two component distributions are separable simply using a cutoff. On the other hand, the component histograms in Fig. 2a are much more difficult to separate using the GR log alone. In fact, the Mahalanobis distance is 2.3 (i.e., |61.5 − 84.2| / [(10.9 + 8.6) × 0.5]) for the channel and crevasse-splay distributions based on their means and standard deviations; and it is 2.2 (i.e., |84.2 − 117.4|/[(10.9 + 19.1) × 0.5]) for the overbank and crevasse-splay distributions. Moreover, the conditions for appearance of bimodality and separation of the components also depend on their proportions, i.e., $\omega_1$ and $\omega_2$ in Eq. (3). Detail on the boundary conditions for bimodality of normal mixtures can be found in Eisenberger (1964).

A normal density is commonly found as a component distribution in natural phenomena, from Gauss' geodesic measurements to Laplace's analysis of errors of experiments. That explains why it is extensively used as a kernel density (Chang et al., 2002; Efron and Tibshirani, 1993). The kernel density method is frequency-based in that it fits the frequency in the histogram using kernel densities. However, component histograms do not need to be normally distributed. In particular, decomposition of a mixture using PCA is totally distribution free (Comaniciu ad Meer, 1999). Moreover, some components can still represent a mixture of subpopulations, and the mixture may or may not be