Research Article

Gaussian Mixture Models Based on Principal Components and Applications

Nada A. Alqahtani and Zakiah I. Kalantan ⓘD

Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia

Correspondence should be addressed to Zakiah I. Kalantan; zkalanten@kau.edu.sa

Data scientists use various machine learning algorithms to discover patterns in large data that can lead to actionable insights. In general, high-dimensional data are reduced by obtaining a set of principal components so as to highlight similarities and differences. In this work, we deal with the reduced data using a bivariate mixture model and learning with a bivariate Gaussian mixture model. We discuss a heuristic for detecting important components by choosing the initial values of location parameters using two different techniques: cluster means, k-means and hierarchical clustering, and default values in the "mixtools" R package. The parameters of the model are obtained via an expectation maximization algorithm. The criteria from Bayesian point are evaluated for both techniques, demonstrating that both techniques are efficient with respect to computation capacity. The effectiveness of the discussed techniques is demonstrated through a simulation study and using real data sets from different fields.

1. Introduction

In real data such as engineering data, efficient dimension reduction is required to reveal underlying patterns of information. Dimension reduction can be used to convert data sets containing millions of functions into manageable spaces for efficient processing and analysis. Unsupervised learning is the main approach to reducing dimensionality. Conventional dimensional reduction approaches can be combined with statistical analysis to improve the performance of big data systems [1]. Many dimension reduction techniques have been developed by statistical and artificial intelligence researchers. Principal component analysis (PCA), introduced in 1901 by Pearson [2], is one of the most popular of these methods. The main purpose of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. Among the many PCA methods, singular value decomposition is used in numerical analysis and Karhunen–Loève expansion in electrical engineering. Eigenvector analysis and characteristic vector

analysis are often used in the physical sciences. In image analysis, the Hotelling transformation is often used for principal component projection.

In recent years, there has been increasing interest in PCA mixture models. Mixture models provide a useful framework for the modelling of complex data with a weighted component distribution. Owing to their high flexibility and efficiency, they are used widely in many fields, including machine learning, image processing, and data mining. However, because the component distributions in a mixture model are commonly formalized as probability density functions, implementations in high-dimensional spaces are constrained by practical considerations.

PCA mixture models are based on a mixture-of-experts technique, which models a nonlinear distribution through a combination of local linear submodels, each with a fairly simple distribution [3]. For the selection of the model, a PCA mixture model was proposed by Kim, Kim, and Bang [4], which has a more straightforward expectation maximization (EM) calculation, does not require a Gaussian error term for each mixture component, and uses an efficient technique for

model order selection. The researchers applied the proposed model to the classification of synthetic data and eye detection [4].

For multimode processes, the Gaussian mixture model (GMM) was developed to estimate the probability density function of the process data under normal operating conditions. However, in the case of high and collinear process variables, learning from process data with GMM can be difficult or impossible. A novel multimode monitoring approach based on the PCA mixture model was proposed by Xu, Xie, and Wang [5] to address this issue. In this method, first, the PCA technique is applied directly to each Gaussian component's covariance matrix to reduce the dimension of process variables and to obtain nonsingular covariance matrices. Then, an EM algorithm is used to automatically optimize the number of mixture components. A novel process monitoring scheme for the detection of multimode processes was developed using the resulting PCA mixture model. The monitoring performance of the proposed approach has been evaluated through case studies [5].

In recent years, hyperspectral imaging has become an important research subject in the field of remote sensing. An important application of hyperspectral imaging is the identification of land cover areas. The rich content of hyperspectral data enables forests, urban areas, crop species, and water supplies to be recognized and classified. In 2016, Kutluk, Kayabol, and Akan [6] proposed a supervised classification and dimensionality reduction method for hyperspectral images, using a mixture of probability PCA (PPCA) models. The proposed mixture model simultaneously allows the reduction of dimensionality and spectral classification of the hyperspectral image. Experimental findings obtained using real hyperspectral data indicate that the proposed approach results in better classification than the state-of-the-art methods [6].

In the field of face recognition, Ahmadkhani and Adibi [7] proposed a supervised version of the PPCA mixture model. This model provides a number of local linear underlying data samples. The underlying manifolds are used for face recognition applications to achieve dimensionality reduction without loss of information.

In this work, we reduce the dimensions of data by applying a PCA technique and then deal with the reduced data or principal components scores using one GMM. Then, we obtain estimates of the parameters using an EM algorithm. Finally, we compare the selection of initial values for location parameters in the mixture model using three different techniques: k-means, hierarchical clustering, and default values in the "mixtools" package.

The rest of this paper is organized as follows. Section 2 briefly defines the concept of PCA. In Section 3, the mixture densities are discussed. Section 4 provides the probability density function of the Gaussian mixture distribution and the EM algorithm that is used to estimate the mixture's parameters. Section 5 shows a PCA mixture model based on the proposed scenario. In Section 6, the experimental results are presented in three subsections, and the main conclusions are listed Section 7.

2. Principal Component Analysis

Suppose we have p-dimensional vectors and need to reduce them to a q-dimensional subspace. The reduction can be achieved by projecting the original vectors on to q dimensions, the principal components, which span the subspace. Suppose that X is a vector of p random variables. To find the principal components, we compute several linear functions of X with maximum variance; most of the variation in X will be accounted for by m principal components, where $m \ll p$. The PCA method determines the correlation between PCA components and data variables; a high correlation indicates important variables. Let Σ be the known covariance matrix for the random variable X. For $k = 1, 2, \ldots, p$, the kth PC is given by $z_k = \alpha_k x$, where α_k is an eigenvector corresponding to the kth largest eigenvalue λ_k. If α_k is chosen to have unit length, $\alpha_k \alpha_k = 1$ and $\text{var}(z_k) = \lambda_k$. Then, the technique of Lagrange multipliers can be used, by maximizing

$$\alpha_1^{'} \Sigma \alpha_1 - \lambda \left(\alpha_1^{'} \alpha_1 - 1 \right), \tag{1}$$

$$\alpha_1^{'} \Sigma \alpha_1 = \alpha_1^{'} \lambda \alpha_1 = \lambda \alpha_1^{'} \alpha_1 = \lambda, \tag{2}$$

and choosing λ to be as large as possible. Here, α_1 is the eigenvector corresponding to the largest eigenvalue of Σ, that is, the first principal component of x. In general, the kth principal component of x is $\alpha_k x$, and $\text{var}(\alpha_k x) = \lambda_k$. The second principal component, $\alpha_2 x$, maximizes $\alpha_2 \Sigma \alpha_2$, subject to being uncorrelated with $\alpha_1 x$. The uncorrelation constraint can be expressed using any of the following equations:

$$\text{cov} \left[\alpha_1^{'} x, t \alpha_2^{'} x \right] = \alpha_1^{'} \Sigma \alpha_2 = \alpha_1^{'} \Sigma \alpha_1 = \alpha_2^{'} \lambda_1 \alpha_1^{'} = \lambda_1 \alpha_2^{'}$$
$$\alpha_1 = \lambda_1 \alpha_1^{'} \alpha_2 = 0. \tag{3}$$

If we choose equation (3), we can write a Lagrange to maximize α_2 as follows:

$$\alpha_2^{'} \Sigma \alpha_2 - \lambda \left(\alpha_2^{'} \alpha_2 - 1 \right) - \phi \alpha_2^{'} \alpha_1. \tag{4}$$

Differentiation of this quantity with respect to α_2 gives us

$$\Sigma \alpha_2 - \lambda \alpha_2 - \phi \alpha_1 = 0. \tag{5}$$

Next, left multiplying $\alpha_1^{'}$ into this expression, we have

$$\alpha_1^{'} \Sigma \alpha_2 - \lambda \alpha_1^{'} \alpha_2 - \phi \alpha_1^{'} \alpha_1 = 0, \tag{6}$$

where, as mentioned above, the first two terms equal zero and $\alpha_1 \alpha_1 = 1$, resulting in $\phi = 0$. Therefore, $\Sigma \alpha_2 - \lambda \alpha_2 = 0$, or $(\Sigma - \lambda I_p)\alpha_2 = 0$, is another eigenvalue equation, and we use the same strategy of choosing α_2 to be the eigenvector associated with the second largest eigenvalue that yields the second principal component of x, namely, $\alpha_2 x$ [8, 9].

3. Mixture Densities

A mixture density is defined as a weighted sum of k component densities [9, 10]. Denote the jth component density by $p(x; \theta j)$, where θ_j indicates the component parameters. We use π_j to denote the weighting factor or "mixing proportions" of the jth component in the combination, with the constraints that $\pi_j \geq 0$ and $\sum_{j=1}^{K} \pi_j = 1$, and $p(j)$ represents the probability that a data sample belongs to the jth mixture component. A K component mixture density is then defined as

$$p(x) = \sum_{j=1}^{K} \pi_j p(x; \theta_j), \quad j = 1, \ldots, K. \tag{7}$$

The mixture model has a vector of parameters, $\underline{\theta} = \{\theta_1, \ldots, \theta_k, \pi_1, \ldots \pi_k\}$.

We consider a mixture density to model a process by selecting a "source" according to the multinomial distribution $\{\pi_1, \ldots \pi_k\}$ and then drawing a sample from the corresponding component density $p(x; \theta_j)$. Therefore, the probability of selecting source j and datum x is $\pi_j p(x; \theta_j)$. Equation (7) gives the marginal probability of selecting datum x. We can think of the source that generated a data vector x as "missing information"; that is, given a data point x, we want to infer which source it is likely to belong to. Section 4 presents the EM algorithm, which is used to iteratively estimate this missing information [11, 12].

In mixture models, we deal with hidden variables as a latent variable, denoted by Z. It takes values $\{1, \ldots, K\}$ as a discrete set satisfying $z_k \epsilon \{0, 1\}$ and $\sum_z z_k = 1$. We define the joint distribution $p(x, z)$ in terms of a marginal distribution $p(z)$ and a conditional distribution $p(x \mid z)$. Generally, in the mixture model, we first choose a sample z from a multinomial distribution and then draw observations for sample x from a distribution that depends on z, i.e.,

$$p(z, x) = p(z)p(x \mid z). \tag{8}$$

The marginal distribution over z is specified in terms of the mixing coefficients π_k, such that,

$$p(z_k = 1) = \pi_k. \tag{9}$$

4. Mixtures of Gaussians

The probability density function of X is defined as

$$p(x \mid \mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi |\Sigma^{-1}|}} \exp\left(-\frac{1}{2}(x - \mu_x)\Sigma_x^{-1}(x - \mu_x)^T\right), \tag{10}$$

where μ_x is a vector of means, $(\mu_{x1}, \ldots, \mu_{xN})$ and Σ_x is an $N \times N$ covariance matrix.

The Gaussian mixture distribution can be written as a linear superposition of Gaussians in the form

$$p(x) = \sum_{k=1}^{K} \pi_k p(x \mid \mu_k, \Sigma_k). \tag{11}$$

Now, the conditional distribution of x given a particular value of z is a Gaussian:

$$p(x \mid z_k = 1) = p(x \mid \mu_k, \Sigma_k), \tag{12}$$

$$p(x \mid z) = \prod_{k=1}^{K} p(x \mid \mu_k, \Sigma_k)^{z_k}. \tag{13}$$

The marginal distribution of x is obtained by summing the joint distribution over all possible states of z to give

$$p(x) = \sum_z p(z)p(x \mid z) = \sum_{k=1}^{K} \pi_k p(x \mid \mu_k, \Sigma_k). \tag{14}$$

An important derived quantity is the "posterior probability" on a mixture component for a given data vector [11]:

$$\gamma(z_k) \equiv p(z_k = 1 \mid x) = \frac{p(z_k = 1)p(x \mid z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x \mid z_j = 1)},$$

$$= \frac{\pi_k p(x \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j p(x \mid \mu_j, \Sigma_j)}. \tag{15}$$

In the example shown in Figure 1, the resulting distribution is bimodal, suggesting that the data come from two different sources. In the figure, the red and green lines indicate two components of the Gaussian mixture distribution.

4.1. EM for Gaussian Mixtures. The EM algorithm is an estimation method used to find the estimators of maximum likelihood when a data set has missing values or latent variables. In this work, we assume a GMM with a fixed number of k and that are known *a priori*.

The EM is obtained as follows:

(1) Initialize the means μ_k, covariances Σ_k, and mixing coefficients π_k and evaluate the initial value of the log-likelihood.

(2) E step: evaluate the responsibilities using the current parameter values:

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n \mid \mu_j, \Sigma_j)}. \tag{16}$$

(3) M step: re-estimate the parameters using the current responsibilities:

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n, \tag{17}$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{\text{new}})(x_k - \mu_k^{\text{new}})^T, \tag{18}$$
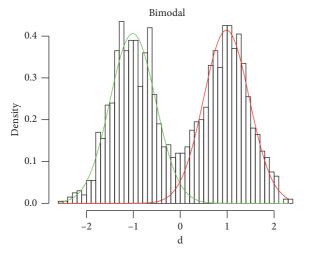
FIGURE 1: Example of density modelled by a mixture of two Gaussian probability density functions.

$$\pi_k^{\text{new}} = \frac{N_k}{N}, \tag{19}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \tag{20}$$

Evaluate the log-likelihood:

$$\ln p(X \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k) \right\}, \tag{21}$$

and check for convergence of either the parameters or the log-likelihood. If the convergence criterion is not satisfied, return to step 2 [13].

5. PCA of Gaussian Mixture Model

In this section, we present the steps of the proposed method. These steps are also illustrated in Figure 2.

(1) Use the PCA technique to reduce the dimensionality of a p-dimensional data set. To find the principal components, we first obtain the eigenvalues and eigenvectors of the covariance matrix. The eigenvalues are the principal components. The total number of principal components corresponds to the total number of variables in the data set.

(2) Choose the m eigenvectors that correspond to the m largest eigenvalues, where m is the number of dimensions of the new feature subspace ($m \leq p$).

(3) Transform the original data set to the lower-dimensional version.

(4) Use the k-means clustering method to partition the data set into clusters. When using k-means, it is important to determine the correct number of clusters for the data.

(5) The new data are modelled by a mixture of Gaussians with parameters set to assumed initial values.

(6) Use the EM algorithm to estimate the unknown parameters representing the mixing proportions between the Gaussians and the means and covariances of each.

(7) Use the Bayesian information criterion (BIC) test to assess the fit of the model; BIC is a model selection among a finite set of models, where the model with the lowest BIC is preferred [14].

6. Experimental Results

To study the effectiveness of the proposed method, we consider two scenarios. In the first scenario, the mixture model is fitted to the reduced data produced by the PCA method. In the second scenario, the clustering method is applied to the reduced data, and then the mixture model is fitted to the new data by taking the cluster means as initial values for the means in the mixture model. We use different types of data sets. The method is implemented on a scaled data set, and the results are illustrated in the following sections. We use the "stats", "mclust", and "mixtools" R packages to implement this method [15–17].

6.1. Simulation Case. We implement the proposed method on simulation data with different sample sizes: 50, 100, and 500. Consider a data set of four variables defined as follows:

$$X_1 \sim N(10, 1.5^2), \tag{22}$$

$$X_2 \sim N(0, 3^2), \tag{23}$$

$$X_3 = X_1 * X_2, \tag{24}$$

$$X_4 = 0.5Z_1 + 0.3Z_2 + 0.7Z_3 + N(0, 1), \tag{25}$$

where Z_1, Z_2, and Z_3 are scaled variables of X_1, X_2, and X_3, respectively. The simulated data consist of four variables, X_1, X_2, X_3, and X_4. The mean values for each variable in the simulated data are $\mu_1 = 10.003$, $\mu_2 = -0.1651$, $\mu_3 = -1.287$, and $\mu_4 = 0.032$.

The implementation includes a graphical plot of the data set as displayed in Figure 3, which shows a pairs plot of the simulation data and its three-dimensional surface. As a first step, we applied the PCA method; the results are summarized in Table 1, which presents the total variance of components. Practically, PCA describes the data in few variables without loss of information. As shown in Figure 4(a) and Table 1, two components explained 93% of the total variance. We considered those two components to comprise a new data set denoted RD data, which contained 93% of the information of the original data. The empirical distribution of the RD data is presented in Figure 4(b).

Hence, we fitted a two-component bivariate mixture model to the RD data. Figure 5(a) displays the fitting of the two components GMM on the new data (with 500 data points); the plot specifies each component's mean and sigma values. In order to estimate the density parameters for each component, we used an EM algorithm for mixtures of
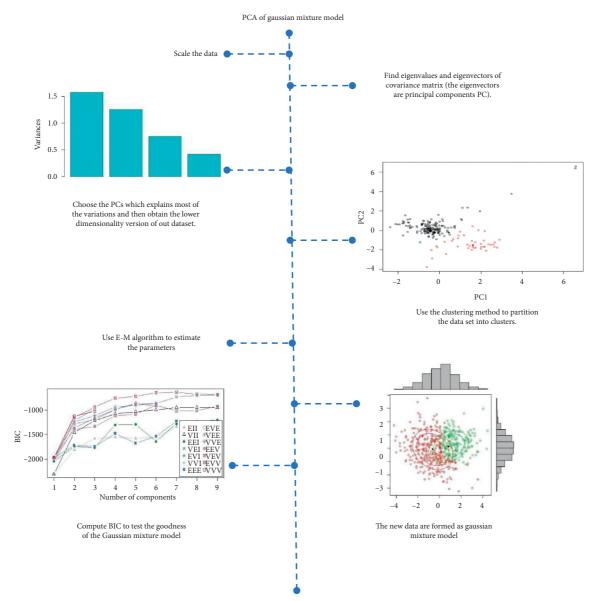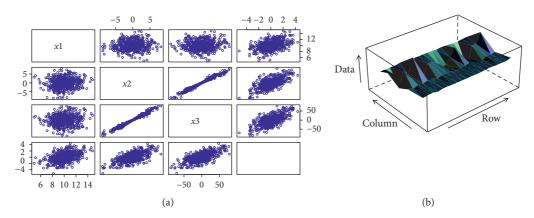
PCA of gaussian mixture model

Scale the data

Find eigenvalues and eigenvectors of covariance matrix (the eigenvectors are principal components PC).

Choose the PCs which explains most of the variations and then obtain the lower dimensionality version of out dataset.

Use the clustering method to partition the data set into clusters.

Use E-M algorithm to estimate the parameters

Compute BIC to test the goodness of the Gaussian mixture model

The new data are formed as gaussian mixture model

Figure 2: PCA of Gaussian mixture model.

(a)

(b)

Figure 3: Simulation case: (a) pairs plot of data and (b) three-dimensional surface of data.

TABLE 1: The results of PCA implementation on the simulation study.

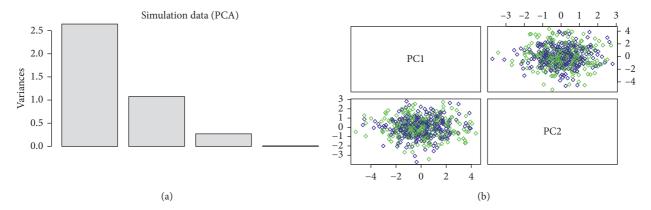| Importance of components:	PC1	PC2	PC3	PC4
Standard deviation	1.6105	1.0658	0.51073	0.09656
Proportion of variance	0.6484	0.2840	0.06521	0.00233
Cumulative proportion	0.6484	0.9325	0.99767	1.00000

(a)

(b)

FIGURE 4: Simulation case: (a) scree plot of PCA and (b) plot of reduced data.
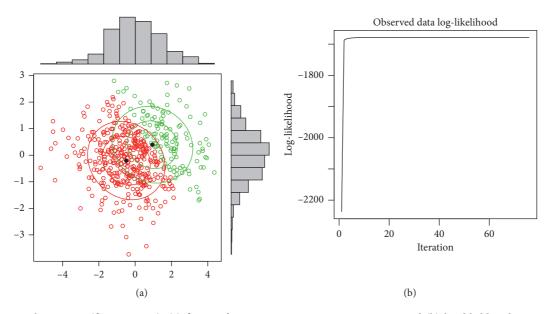

(a)

(b)

FIGURE 5: Simulation case (first scenario): (a) fitting of two-component Gaussian mixture and (b) log-likelihood versus iteration.

bivariate data. Table 2 presents the estimates of model parameters for each component; it also displays the estimates for different sample sizes. We computed the BIC of the model for the three cases presented in Table 2 and observed that the BIC value became large with increasing sample size. Figure 5(b) shows a plot of log-likelihood versus number of iterations; it is clear that the log-likelihood remained low when the number of iterations increased and that the EM method reached convergence.

The second scenario involves the selection of the initial values for the means in the mixture model; this was done by applying the k-means method to the reduced data. Then, the centres (or the means) of the clusters were taken as initial

values. Thus, the RD data were partitioned into two clusters using the k-means method, denoted as PC1 and PC2, as shown in Figure 6(a). The cluster means for PC1 were 1.343257 and −1.210463, whereas those of PC2 were −0.02877005 and 0.02592586. In the next step, the bivariate GMM was fitted to the RD data using the cluster means as initial values. A visualization of the fitted bivariate GMM is presented in Figure 6(b). A summary of the results is given in Table 3. The resulting BIC value was 3389.031, which was the same as the BIC for GMM, as displayed in Table 2.

Parameter estimates were compared for the mixture and clustering methods. The k-means method computes the conventional Euclidean distance of given data, whereas

TABLE 2: Simulation case: parameter estimates of the bivariate GMM (first technique) (the log-likelihood, BIC, and number of iterations are provided).

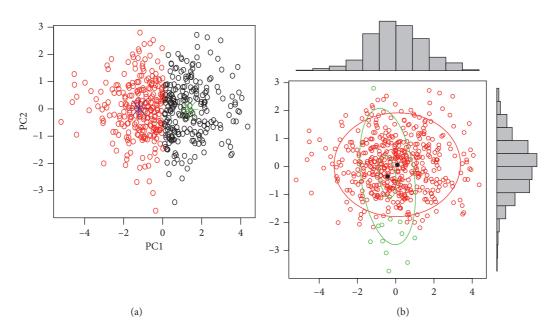| Sample size	First component					Second component				
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
50	0.53806	−0.19744	0.46146	2.68185	1.35652	0.46194	0.22997	−0.53750	2.35390	0.30614
	$\text{cov}(\sigma_1, \sigma_2) = 1.3565245$					$\text{cov}(\sigma_1, \sigma_2) = -0.30614414$				
	Log-likelihood = −331.7415, BIC = 704.9295									
	Number of iterations: 70									
100	0.38040	1.07288	−0.6829	2.33042	1.04702	0.61959	−0.6587	0.41927	1.46089	0.73289
	$\text{cov}(\sigma_1, \sigma_2) = 0.945838$					$\text{cov}(\sigma_1, \sigma_2) = 0.1452915$				
	Log-likelihood = −332.917, BIC = 707.2805									
	Number of iterations: 96									
500	0. 65985	−0.48985	−0.2038	2.10879	1.00732	0.34015	0.9503	0.39532	2.29243	0.96703
	$\text{cov}(\sigma_1, \sigma_2) = -0.2169786$					$\text{cov}(\sigma_1, \sigma_2) = -0.1483861$				
	Log-likelihood = −1678.241, BIC = 3379.508									
	Number of iterations: 77									

(a)

(b)

FIGURE 6: Reduced data (simulation data; second scenario): (a) estimated number of observations through PC1 and PC1 after applying the k-means method and (b) fitted GMM.

GMM computes the weighted distance by considering the variance in its measurement calculations. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centres of the latent Gaussians.

6.2. Forensic Glass Fragment Data. In this section, the proposed method is implemented using forensic glass fragment (FG) data (available in an R package). The FG data include 214 observations and 10 variables. Consider the four selected variables, manganese (Mg), aluminium (Al), silicon (Si), and potassium (K), and the types of fragments (WinF, WinNF, Veh, Con, Tabl, and Head). Note that the variable type is used only to classify the data. Figure 7(a) shows a pairwise plot for four measurements of the scaled data set.

As a first step, we applied the PCA method to the data to obtain a linear estimate of dimensionality. The results are summarized in Table 4. As shown in the scree plot in Figure 7(b), three components explained 89% of the total variance and two components explained 70% of it. We therefore consider the first two components in the following.

Hence, the two-component bivariate mixture model was fitted to the FG data, as shown in Figure 8(a). The mean and sigma values for each component are also shown in Figure 8(a). In order to estimate the density parameters for each component, we used the EM algorithm for mixtures of bivariate data. Note that the initial values were chosen using the "mixtools" package. Table 5 presents estimates of the model parameters for each component, as well as the BIC. Figure 8(b) shows a plot of log-likelihood versus number of iterations; it is clear that the log-likelihood remained low

TABLE 3: Simulation case: parameter estimates of the bivariate GMM (second technique), taking cluster means as initial values of model means.

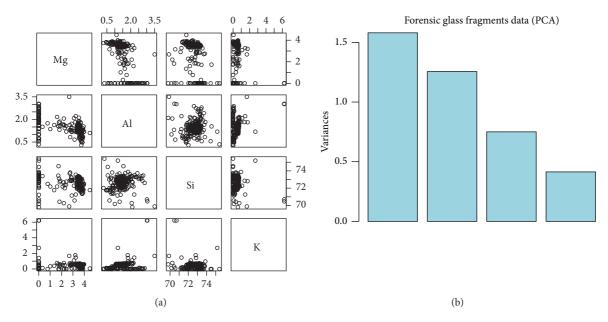| Sample size	Model parameters									
	First component					Second component				
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
	0.83586	0.08308	0.06959	2.99632	0.93938	0.16414	−0.4231	−0.35438	0.59144	1.61035
500	$\text{cov}(\sigma_1, \sigma_2) = 0.0173803$					$\text{cov}(\sigma_1, \sigma_2) = 1.6103530$				
	Log-likelihood = −1673.792, BIC = 3389.031									
	Number of iterations: 734									

FIGURE 7: (a) Pairs plot of data and (b) plot of PCA.

when the number of iterations increased and that the EM method reached convergence.

Next, we studied the selection of initial values for the location parameters in the mixture model based on the centres of clusters obtained with different clustering methods: k-means and hierarchical clustering. First, the k-means method was applied to the reduced FG data, and then the centres (or the means) of clusters were taken as initial values. A visualization of the resulting data for PC1 and PC2 is given in Figure 9(a). The cluster centres (means) for PC1 were −0.36697 and 1.45934 and those for PC2 are 0.35596 and −1.41558. Then, the bivariate GMM was fitted to the reduced FG data using the cluster means as initial values. The results are summarized in Table 6, and a plot of the fitted mixture model is shown in Figure 9(b). The resulting BIC was 1174.485.

Second, the initial values for the location parameters were selected using hierarchical clustering with the "mclust" package. As a first step, the hierarchical clustering was applied to the reduced FG data. A visualization of the resulting data for PC1 and PC2 is given in Figure 10(a). The resulting centres were 1.49372 and −0.51297 for PC1, while those for PC2 were −0.58542 and 0.20104. In the next step, the bivariate GMM was fitted to the reduced FG data using

TABLE 4: Forensic glass fragments data: summary of PCA implementation on the data.

Importance of components:	PC1	PC2	PC3	PC4
Standard deviation	1.257	1.1205	0.8661	0.6439
Proportion of variance	0.395	0.3139	0.1875	0.1037
Cumulative proportion	0.395	0.7088	0.8963	1.00000

the cluster means as initial values, as shown in Figure 10(b). Table 7 shows the results for the mixture model, and the resulting BIC was 1111.483.

We observe that the selection of initial values for location parameters using clustering methods provided good results, similar to those obtained by selection of initial values with the "mixtools" package.

6.3. Applications to Real Data. In this section, the proposed method is implemented on a real data set obtained from the "Knoema" website [18]. Knoema is one of the most comprehensive sources of global decision-making data. The data set for cancer incidence in 100 countries in a specific year (2016) includes 3168 observations and 32 variables and

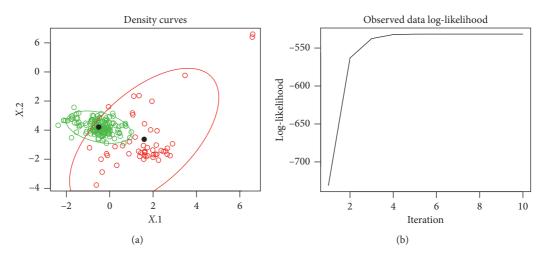(a)   (b)

Figure 8: Forensic glass fragments data: (a) fitting two-component Gaussian mixture nd (b) log-likelihood versus iteration.

Table 5: Forensic glass fragments data: parameter estimates and BIC of the bivariate GMM (first technique).

Values	Model parameters									
	First component					Second component				
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
	0.24311	1.58580	−0.62982	1.97795	3.97284	0.75689	−0.50935	0.20229	0.37507	0.20665
Parameter estimation	$\text{cov}(\sigma_1, \sigma_2) = 1.63918$					$\text{cov}(\sigma_1, \sigma_2) = -0.10266$				
	Log-likelihood = −531.5947, BIC = 1111.483									
	Number of iterations: 27									

(a)   (b)
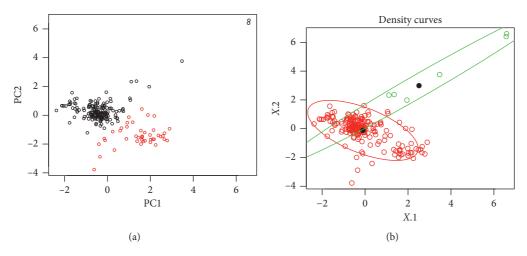
Figure 9: Forensic glass fragments data: (a) cluster criterion and centres and (b) fitted GMM.

covers different cancer types. In the first step, we used the PCA approach to approximate the dimensionality of the data in a linear manner. The results of the first five main components are summarized in Table 8. As shown in the scree plot in Figure 11, 93% of the overall variance was described by the first two components. We thus consider the first two components in the following.

The data were thus fitted by the two-component bivariate mixture model, as shown in Figure 12(a), which gives the mean and sigma values for each component. We used the EM algorithm for bivariate data mixtures to determine density parameters for each component. The initial values were as suggested by the "mixtools" package. The estimates of model parameters are provided in Table 9 for each component, as well as the BIC. Figure 12(b) shows a plot of log-likelihood versus number of iterations; the log-likelihood remained low as the number of iterations increased, and the EM method reached convergence.

TABLE 6: Forensic glass fragments data: parameter estimates and BIC of the bivariate GMM (second technique) by taking cluster means as initial values of model means, and the initial values for PC1 were −0.36697 and 1.45934 and those for PC2 are 0.35596 and −1.41558.

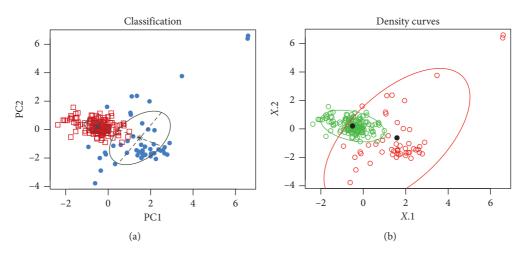| Values | Model parameters | | | | | | | | |
| | First component | | | | Second component | | | | |
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
	0.96102	−0.10113	−0.12063	1.10542	0.72645	0.03898	2.49307	2.97325	6.61789	4.94843
Parameter estimation	$\text{cov}(\sigma_1, \sigma_2) = 0.54182$					$\text{cov}(\sigma_1, \sigma_2) = 5.64324$				
	Log-likelihood = −563.0958, BIC = 1174.485									
	Number of iterations: 28									

FIGURE 10: (a) Cluster criterion and centres (using "mclust"), and (b) fitted Gaussian mixture model.

TABLE 7: Forensic glass fragments data: parameter estimates and BIC of the bivariate GMM (second technique) by taking cluster means as initial values of model means, and the initial values for 1.49372 and −0.51297 for PC1, while those for PC2 were −0.58542 and 0.20104.

| Values | Model parameters | | | | | | | | |
| | First component | | | | Second component | | | | |
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
	0.24293	1.58683	−0.63017	1.97751	3.97491	0.75707	−0.50919	0.20222	0.37519	0.20682
Parameter estimation	$\text{cov}(\sigma_1, \sigma_2) = 1.64088$					$\text{cov}(\sigma_1, \sigma_2) = -0.10269$				
	Log-likelihood = −531.5947, BIC = 1111.483									
	Number of iterations: 7									

TABLE 8: Summary of PCA implementation on cancer incidence data.

Importance of components	PC1	PC2	PC3	PC4	PC5
Standard deviation	5.1588	1.7982	1.01642	0.57151	0.47353
Proportion of variance	0.8317	0.1011	0.03228	0.01021	0.00701
Cumulative proportion	0.8317	0.9327	0.96499	0.97520	0.98220

FIGURE 11: Cancer incidence data: plot of PCA.

Then, we used the centres of clusters with different clustering techniques (k-means and hierarchical clustering) to analyse the choice of initial parameters for the location parameters of the mixture model. First, the k-means method was applied to the reduced data, and then the centres (or the
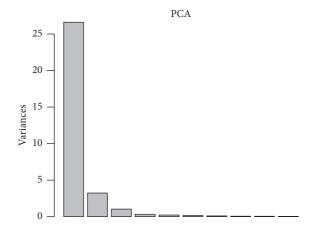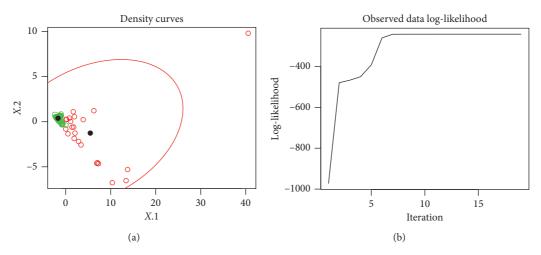
(a)

(b)

FIGURE 12: Cancer incidence data: (a) Fitting of the two-components Gaussian mixture, (b) log-likelihood versus iteration.

TABLE 9: Cancer incidence data: parameter estimates and BIC of the bivariate GMM (first technique).

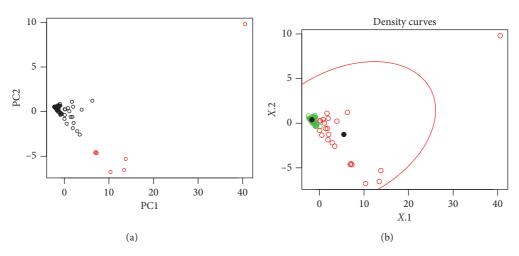| Values	Model parameters									
	First component					Second component				
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
	0.23754	5.47758	−1.26673	70.72932	11.1328	0.76245	−1.70651	0.39464	0.25654	0.07405
Parameter estimation		$\text{cov}(\sigma_1, \sigma_2) = 9.3649$					$\text{cov}(\sigma_1, \sigma_2) = -0.0824$			
	Log-likelihood = −242.2074, BIC = 556.9625									
	Number of iterations: 30									

(a)

(b)

FIGURE 13: Cancer incidence data: (a) cluster criterion and centres and (b) fitted GMM.

TABLE 10: Cancer incidence data: parameter estimates and BIC of the bivariate GMM (second technique) by taking cluster means as initial values of model means.

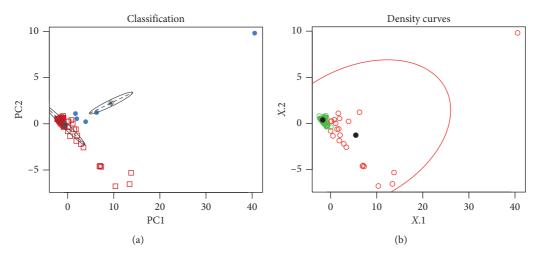| Values	Model parameters									
	First component					Second component				
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
	0.23784	5.47040	−1.26496	70.67932	11.1214	0.76216	−1.7071	0.39475	0.25557	0.07398
Parameter estimation		$\text{cov}(\sigma_1, \sigma_2) = 9.34287$					$\text{cov}(\sigma_1, \sigma_2) = -0.0824$			
	Log-likelihood = −242.2075, BIC = 556.9627									
	Number of iterations: 30									

(a)

(b)

FIGURE 14: Cancer incidence data: (a) cluster criterion and centres (using "mclust") and (b) fitted GMM.

TABLE 11: Cancer incidence data: the estimate results of fitting of the bivariate mixture model by taking cluster means as initial values of model means.

Values	Model parameters									
	First component					Second component				
	π_1	μ_1	μ_2	σ_1^2	σ_2^2	π_2	μ_1	μ_2	σ_1^2	σ_2^2
	0.23719	5.48573	−1.26877	70.7862	11.1459	0.76281	−1.70579	0.39453	0.25765	0.07416
Parameters estimation	$\text{cov}(\sigma_1, \sigma_2) = 9.39014$					$\text{cov}(\sigma_1, \sigma_2) = -0.08263$				
	Log-likelihood = −242.2076, BIC = 556.9629									
	Number of iterations: 13									

means) of clusters were taken as initial values. A visualization of the resulting data for PC1 and PC2 is given in Figure 13(a). The cluster centres (means) for PC1 were 14.19303 and −1.0799, while those of PC2 were −3.22915 and 0.24569. Then, the bivariate GMM was fitted to the reduced data using the cluster means as initial values. The results are summarized in Table 10, and a plot of the fitted mixture model is shown in Figure 13(b). The resulting BIC was 556.9627.

Second, a hierarchical clustering procedure was used to specify the initial values of the location parameters; this was achieved using the "mclust" package. The hierarchical clustering of the reduced data was implemented as a first step. Figure 14(a) provides a visualization of the data obtained for PC1 and PC2: the resulting centres were 9.39997 and −0.58877 for PC1, while those for PC2 were 2.25038 and −0.14095. In the next step, the bivariate GMM was fitted on the reduced data using the cluster means as initial values; the fitted data are presented in Figure 14(b). Table 11 shows the results of the mixture model and the BIC of 556.9629.

We observe that the selection of initial values for location parameters using clustering methods provided good results, similar to those obtained when the initial values were selected using the "mixtools" package.

7. Conclusions

This work aimed to study the applications of PCA in mixture models. First, we discussed the use of the well-known PCA technique for dimension reduction and applied it to high-dimensional data sets. Then, in the reduced data (which contained only two variables), we dealt with the two variables together and fitted a two-component bivariate GMM to the data. We used an EM algorithm to estimate the model parameters. This approach is suitable for large data sets with high dimension and can solve the problem of overfitting. We compared three different techniques for the selection of initial values of location parameters in the mixture model: two clustering methods, k-means and hierarchical clustering, and default values from the "mixtools" package. With all three techniques, EM convergence was reached and similar BIC values were obtained.

Data Availability

The data were taken from Knoema website, which is one of the most comprehensive sources of global decision-making data in the world (world and regional statistics, national data, maps, rankings) (retrieved from https://knoema.com/Atlas).

Disclosure

This paper was a component of a Masters' thesis by the first author under the supervision of the second author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] M. H. Ur Rehman, C. S. Liew, A. Abbas, P. P. Jayaraman, T. Y. Wah, and S. U. Khan, "Big data reduction methods: a survey," *Data Science and Engineering*, vol. 1, no. 4, pp. 265–284, 2016.

[2] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[3] Z. Jin, F. Davoine, and Z. Lou, "An effective EM algorithm for PCA mixture model," in *Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 626–634, Springer, Lisbon, Portugal, August 2004.

[4] H.-C. Kim, D. Kim, and S. Y. Bang, "An efficient model order selection for PCA mixture model," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1385–1393, 2003.

[5] X. Xu, L. Xie, and S. Wang, "Multimode process monitoring with PCA mixture model," *Computers & Electrical Engineering*, vol. 40, no. 7, pp. 2101–2112, 2014.

[6] S. Kutluk, K. Kayabol, and A. Akan, "Classification of hyperspectral images using mixture of probabilistic PCA models," in *Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1568–1572, IEEE, Budapest, Hungary, September 2016.

[7] S. Ahmadkhani and P. Adibi, "Face recognition using supervised probabilistic principal component analysis mixture model in dimensionality reduction without loss framework," *IET Computer Vision*, vol. 10, no. 3, pp. 193–201, 2016.

[8] I. Jolliffe, "Principal component analysis," in *International Encyclopedia of Statistical Science*, pp. 1094–1096, Springer, Berlin, Germany, 2011.

[9] Z. I. Kalantan and N. A. Alqahtani, "A study of principal components analysis for mixed data," *International Journal of Advanced and Applied Sciences*, vol. 6, no. 12, pp. 99–104, 2019.

[10] G. J. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, Inc, Hoboken, NJ, USA, 2000.

[11] Z. I. Kalantan and J. Einbeck, "Quantile-based estimation of the finite cauchy mixture model," *Symmetry*, vol. 11, no. 9, p. 1186, 2019.

[12] J. Verbeek, "Mixture models for clustering and dimension reduction," Universiteit Van Amsterdam, Amsterdam, Netherlands, Doctoral dissertation, 2004.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Berlin, Germany, 2006.

[14] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018.

[16] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models," *The R Journal*, vol. 8, no. 1, pp. 205–233, 2017.

[17] T. Benaglia, D. Chauveau, D. Hunter, and D. Young, "Mixtools: an R package for analyzing finite mixture models," *Journal of Statistical Software*, vol. 32, no. 6, pp. 1–29, 2009.

[18] World Data Atlas, "World and regional statistics, national data, maps, rankings," https://knoema.com/atlas.