CHAPTER - 4
4.1 INTRODUCTION

Clustering is division of data into groups of similar objects. Representing data by fewer clusters necessarily achieves simplification and improves accuracy of model building. From a practical perspective, clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, database applications, web analysis, CRM, marketing, computational biology and many other areas. Clustering is the subject of active research in several fields such as statistics, pattern recognition and machine learning. A variety of algorithms have recently emerged that meet these requirements and were successfully applied to real-life data mining problems.

In chapter 3, a supervised learning algorithm for time series with bivariate AR(1) model is developed analyzed by considering number of classes in whole data set are known and fixed. However, there are some other time series data in which it is difficult to identify number of classes in the beginning and require a different approach for classifying time series data. In general, when the number of classes is unknown, the classification problem refers to unsupervised learning.

Anthony Bagnall and Gareth Janacek (2005) studied clustering time series from ARMA models by transforming real valued time series into a binary series through the process of clipping. Ashish Singhal and Dale E. Seborg (2005) developed a method that uses similarity factors to characterize the degree of dissimilarity between data sets. Jeroen Boets et al. (2005) performed clustering time series using

In all these papers, they considered that the attributes associated with time series data are univariate. Very little work has been reported in literature regarding clustering of bivariate time series data. The bivariate time series require segmentation based on both variables under study for efficient analysis. For example, in clustering regions according to ambient air quality, it is well known that two components Nitrogen dioxide (NO$_2$) and Sulphur dioxide (SO$_2$) in air over a period of time formulate a bivariate time series. The regions are to be segmented according to intensity of pollutants both put together in order to have optimal control and monitor on ambient air quality.

For these sorts of situations, the number of clusters in the whole data obtained over a period of time for various regions is not known and clustering is purely dependent upon the variables. Hence, for obtaining groups of regions one has to
utilize clustering algorithm. Therefore, in this chapter an unsupervised learning algorithm for bivariate time series data using bivariate AR(1) model is developed and analysed. It is assumed that the whole time series data in the domain of study are characterized by finite mixture of bivariate autoregressive process of order 1. In this mixture model the number of classes (components) plays a dominant role to classify observed time series. The K-means algorithm is utilized for obtaining number of clusters.

The model parameters are estimated by developing updated equations for EM algorithm. The initial estimates of parameters are obtained through optimal Least Squares estimates of bivariate AR(1) model. Using refined estimates the patterns for each region (segment) are estimated. The learning algorithm is obtained through maximizing the component likelihood function. The performance of developed algorithm is studied through quality metric misclassification rate. A comparative study of the developed algorithm with existing algorithm is also carried.

4.2 K-MEANS ALGORITHM

Let the time series are generated by M different bivariate AR(1) models which corresponds to M clusters of interest with weights \( P(\omega_1), P(\omega_2), ..., P(\omega_M) \). Let \( P(Z_t | \omega_k, \Phi_k) \) denote the conditional likelihood function of \( k^{th} \) cluster with \( \Phi_k \) as set of parameters for the model. Then the conditional likelihood function of mixture of finite bivariate AR process of order 1 can be expressed in the form of

\[
P(Z_t | \Theta) = \sum_{k=1}^{M} P(Z_t | \omega_k, \Phi_k) P(\omega_k)
\]  

(4.2.1)
where, \( \Theta = \{ \Phi_1, \Phi_2, ..., \Phi_M, \theta(\omega_1), \theta(\omega_2), ..., \theta(\omega_M) \} \) represents the set of model parameters for mixture model.

A bivariate time series \( Z_t \) is assigned to cluster \( \omega_k \) with posterior probability \( P(\omega_k|Z_t) \) such that \( \sum_{k=1}^{M} P(\omega_k) = 1 \). This model includes mixture of univariate autoregressive process of order 1 model when the model parameters \( \phi_{12}, \phi_{21} \) and \( \phi_{22} \) are zero for \( k=1,2,...,M \). If \( M=1 \), then this process becomes Bivariate AR(1) process. The feature vector representing \( \phi = (\phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}) \) for each bivariate time series is obtained through least squares method of estimation given by Douglas C. Montgomery, Lynwood A. Johnson and John S. Gardiner (1990).

The K-means algorithm is a faster method to perform clustering of time series data. In K-means algorithm, continuous reassignments of objects into different clusters are done so that within cluster distance is minimized. For utilizing the K-means algorithm, each time series is reduced to the object vector of \( \Phi_t = (\phi_{11t}, \phi_{12t}, \phi_{21t}, \phi_{22t}) \) where \( \phi_{ij}, i=1,2; j=1,2 \) and \( t=1,2,...,N \) are the parameters of bivariate AR(1) model which are obtained using ordinary least square method of estimation. If \( \Phi_t = (\phi_{11t}, \phi_{12t}, \phi_{21t}, \phi_{22t}) \), \( t=1,2,...,N \) is an object and \( C_k \) is the center of the \( k^{th} \) cluster, then K-means algorithm attempts to minimize the objective function

\[
F = \sum_{t=1}^{N} \sum_{k=1}^{M} (\Phi_{tk} - C_k)^2
\]

The K-means algorithm also requires initial number of classes, which can be obtained by plotting a bivariate scatter surface for all time series of training data.
Based on the number of surfaces visualized in scatter surface plot, the initial value of $M$ is taken.

The K-means algorithm for obtaining the number of clusters and classes is as follows:

Step 1: Identify the value of K.

Step 2: Initialize K cluster centers.

Step 3: Decide the class memberships of $N$ objects by assigning them to nearest Cluster center.

Step 4: Re-estimate the K cluster centers by assuming the membership found above as correct.

Step 5: If none of the N objects changed membership in last iteration, then exit otherwise go to step 3.

For classifying the training time series data into M clusters the residual variances of each class, correlation coefficient between the attributes in each class and model parameters $\phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}$ within each class are obtained using clustered time series of that class and taking average over the number of time series in that class.

4.3 ESTIMATION OF PARAMETERS THROUGH EM ALGORITHM

In this section, estimation of parameters using Expectation- Maximization algorithm is presented. Here, without loss of generality, it is assumed that the time series values are standardized by taking each individual observation from their mean value and hence the mean function of each time series is considered as zero. Therefore the problem of estimation is to obtain refined estimates of the
parameters $P(\omega_k), \sigma^2_{\omega_k}, \sigma^2_{\varepsilon_k}, \phi_{11}, \phi_{12}, \phi_{21}, \phi_{22}$ and $\rho$ for each class. For estimating the parameters through EM algorithm, one can use likelihood function of the sample.

Let $D = \{Z_1, Z_2, ..., Z_N\}$ are $N$ time series and assuming that these are conditionally independent under the given model, then the likelihood function of realization is

$$P(D | \Theta) = \prod_{i=1}^{N} P(Z_i | \Theta)$$

$$= \prod_{i=1}^{N} \sum_{k=1}^{M} P(Z_i | \omega_k, \Phi_k) P(\omega_k)$$

The Log likelihood function of realization $D$ is

$$\ln P(D | \Theta) = \sum_{i=1}^{N} \ln \left( \sum_{k=1}^{M} P(Z_i | \omega_k, \Phi_k) P(\omega_k) \right)$$

In EM algorithm, the first step is to obtain expectation of log likelihood. Given the observed data set $D$ and current parameter estimate $\Theta(r)$, the expected value of log likelihood for $M$ clusters of time series each having $n$ points of observations can be expressed as

$$Q(\Theta | \Theta(r)) = E(\ln P(D | \Phi))$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{M} P(\omega_k | Z_i, \Theta(r)) \ln P(Z_i | \omega_k, \Phi_k) + \sum_{i=1}^{N} \sum_{k=1}^{M} P(\omega_k | Z_i, \Theta(r)) \ln P(\omega_k)$$

(4.3.1)

where, the posterior probabilities $P(\omega_k | Z_i, \Theta)$ are computed using Bayes rule as

$$P(\omega_k | Z_i, \Theta) = \frac{P(Z_i | \omega_k, \Phi_k) P(\omega_k)}{\sum_{q=1}^{M} P(Z_i | \omega_q, \Phi_q) P(\omega_q)}; \quad t = 1, 2, ..., N \text{ and } k = 1, 2, ..., M.$$  

and $P(\omega_k) = \frac{1}{N} \sum_{i=1}^{N} P(\omega_k | Z_i, \Theta(r))$ (4.3.2)
This implies, $Q(\Theta|\Theta(r)) = \frac{\sum_{i=1}^{N} \sum_{k=1}^{M} P(\omega_k|Z_t, \Theta(r)) \left( (n-1) \ln \left( 2\pi \sigma_{e_k} \sigma_{e_k} \sqrt{1-\rho_k^2} \right) + \frac{1}{2} \sum_{u=1}^{n} \frac{e_{x_{t,u}}^2}{\sigma_{e_k}^2} - \frac{2\rho_k e_{x_{t,u}} e_{x_{t,u-1}}}{\sigma_{e_k} \sigma_{e_k}} + \frac{e_{x_{t,u-1}}^2}{\sigma_{e_k}^2} \right)}{1}$

$$- \sum_{i=1}^{N} \sum_{k=1}^{M} P(\omega_k|Z_t, \Theta(r)) \ln \left( \frac{P(Z_t|\omega_k, \Phi_k) P(\omega_k)}{\sum_{q=1}^{M} P(Z_t|\omega_q, \Phi_q) P(\omega_q)} \right)$$

(4.3.3)

since residual terms in the sample parts of time series starts from $u=2$ in bivariate AR(1) process, i.e. $e_{x_{t,u}} = X_{t,u} - \phi_1 X_{t,u-1} - \phi_2 Y_{t,u-1}$ and $e_{y_{t,u}} = Y_{t,u} - \phi_1 Y_{t,u-1} - \phi_2 X_{t,u-1}$.

In the M-step of EM algorithm, one has to find estimates of parameters by maximizing $Q(\Theta|\Theta(r))$.

The updated equation $P(\hat{\omega}_k)$ of the model is

$$P(\hat{\omega}_k) = \frac{1}{N} \sum_{i=1}^{N} P(\omega_k|Z_t, \Theta(r)) \quad \text{for} \quad k = 1, 2, ..., M \quad (4.3.4)$$

For estimating $\sigma_{e_1}^2, \sigma_{e_2}^2, \phi_{11k}, \phi_{12k}, \phi_{21k}, \phi_{22k}$ and $\rho_k$, the updated equations respectively are

$$\frac{\sum_{i=1}^{N} P(\omega_k|Z_t, \Theta(r)) \left( (n-1) - \frac{1}{(1-\rho_k^2)} \sum_{u=2}^{n} \left[ \frac{X_{t,u}^2}{\sigma_{e_k}^2} + \left( \frac{e_{x_{t,u}}^2}{\sigma_{e_k}^2} - \frac{2\rho_k e_{x_{t,u}} e_{x_{t,u-1}}}{\sigma_{e_k} \sigma_{e_k}} + \frac{e_{x_{t,u-1}}^2}{\sigma_{e_k}^2} \right) \right] \right)}{1}$$

$$\frac{\sum_{i=1}^{N} P(\omega_k|Z_t, \Theta(r)) \left( (n-1) - \frac{1}{(1-\rho_k^2)} \sum_{u=2}^{n} \left[ \frac{e_{y_{t,u}}^2}{\sigma_{e_k}^2} - \frac{2\rho_k e_{y_{t,u}} e_{y_{t,u-1}}}{\sigma_{e_k} \sigma_{e_k}} + \frac{e_{y_{t,u-1}}^2}{\sigma_{e_k}^2} \right) \right] \right)}{1}$$

(4.3.5)
\[
\sum_{i=1}^{N} P(\omega_{k} | Z_{i}, \Theta (r)) \left\{ (n-1) - \frac{1}{(1 - \rho_{k}^{2})} \sum_{u=2}^{n} \frac{Y_{i,u}^{2}}{\sigma_{\varepsilon_{k}^{2}}} + \left( \frac{\phi_{12k}^{2}}{\sigma_{\varepsilon_{k}^{2}}^{2}} - \frac{\rho_{k} \phi_{11k} \phi_{21k}}{\sigma_{\varepsilon_{k}^{2}} \sigma_{\varepsilon_{h}^{2}}} \right) X_{i,u-1}^{2} + \left( \frac{\phi_{21k}^{2}}{\sigma_{\varepsilon_{h}^{2}}} - \frac{\rho_{k} \phi_{12k} \phi_{11k}}{\sigma_{\varepsilon_{h}^{2}} \sigma_{\varepsilon_{k}^{2}}} \right) \left( \frac{2 \phi_{21k} \phi_{12k}}{\sigma_{\varepsilon_{k}^{2}}^{2}} - \frac{\rho_{k} \phi_{11k} \phi_{21k}}{\sigma_{\varepsilon_{k}^{2}} \sigma_{\varepsilon_{h}^{2}}} \right) \right) Y_{i,u} \right\} = 0 \quad \text{for} \quad k = 1, 2, \ldots, M
\]

(4.3.6)

\[
\frac{1}{\sigma_{\varepsilon_{x}^{2}}} \sum_{i=2}^{n} X_{i,u} - \frac{\rho_{k}}{\sigma_{\varepsilon_{k}^{2}}} \sum_{u=2}^{n} X_{i,u-1} Y_{i,u} - \left( \frac{\phi_{11k}}{\sigma_{\varepsilon_{x}^{2}}} - \frac{\rho_{k} \phi_{21k}}{\sigma_{\varepsilon_{h}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1}^{2} Y_{i,u-1} - \left( \frac{\phi_{12k}}{\sigma_{\varepsilon_{x}^{2}}} - \frac{\rho_{k} \phi_{11k}}{\sigma_{\varepsilon_{x}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1}^{2} Y_{i,u-1} - \left( \frac{\phi_{21k}}{\sigma_{\varepsilon_{h}^{2}}} - \frac{\rho_{k} \phi_{12k} \phi_{11k}}{\sigma_{\varepsilon_{x}^{2}} \sigma_{\varepsilon_{h}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1}^{2} Y_{i,u-1} - \left( \frac{2 \phi_{21k} \phi_{12k}}{\sigma_{\varepsilon_{k}^{2}}^{2}} - \frac{\rho_{k} \phi_{11k} \phi_{21k}}{\sigma_{\varepsilon_{k}^{2}} \sigma_{\varepsilon_{h}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1}^{2} Y_{i,u-1} = 0 \quad \text{for} \quad k = 1, 2, \ldots, M
\]

(4.3.7)

\[
\frac{1}{\sigma_{\varepsilon_{x}^{2}}} \sum_{i=2}^{n} X_{i,u} Y_{i,u-1} - \frac{\rho_{k}}{\sigma_{\varepsilon_{k}^{2}}} \sum_{u=2}^{n} Y_{i,u-1} Y_{i,u} - \left( \frac{\phi_{11k}}{\sigma_{\varepsilon_{x}^{2}}} - \frac{\rho_{k} \phi_{21k}}{\sigma_{\varepsilon_{h}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1}^{2} Y_{i,u-1} - \left( \frac{\phi_{21k}}{\sigma_{\varepsilon_{h}^{2}}} - \frac{\rho_{k} \phi_{12k} \phi_{11k}}{\sigma_{\varepsilon_{x}^{2}} \sigma_{\varepsilon_{h}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1}^{2} Y_{i,u-1} = 0 \quad \text{for} \quad k = 1, 2, \ldots, M
\]

(4.3.8)

\[
\frac{1}{\sigma_{\varepsilon_{x}^{2}}} \sum_{i=2}^{n} X_{i,u} - \frac{\rho_{k}}{\sigma_{\varepsilon_{k}^{2}}} \sum_{u=2}^{n} X_{i,u-1} Y_{i,u} - \left( \frac{\phi_{11k}}{\sigma_{\varepsilon_{x}^{2}}} - \frac{\rho_{k} \phi_{21k}}{\sigma_{\varepsilon_{h}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1} Y_{i,u-1} - \left( \frac{\phi_{21k}}{\sigma_{\varepsilon_{h}^{2}}} - \frac{\rho_{k} \phi_{12k} \phi_{11k}}{\sigma_{\varepsilon_{x}^{2}} \sigma_{\varepsilon_{h}^{2}}} \right) \sum_{u=2}^{n} X_{i,u-1} Y_{i,u-1} = 0 \quad \text{for} \quad k = 1, 2, \ldots, M
\]

(4.3.9)
\[
\sum_{i=1}^{N} P(\omega_k | Z_i, \Theta(r)) \left[ \frac{\rho_k}{\sigma_{\omega_k}} \sum_{u=2}^{n} X_{t,u} - \frac{1}{\sigma_{\omega_k}} \sum_{u=2}^{n} Y_{t,u-1} Y_{t,u} \right]
\]
\[
- \left( \frac{\rho_k \phi_{11k} - \phi_{21k}}{\sigma_{\omega_k}} \sum_{u=2}^{n} X_{t,u-1} Y_{t,u-1} - \left( \frac{\rho_k \phi_{12k}}{\sigma_{\omega_k}} - \frac{\phi_{22k}}{\sigma_{\omega_k}} \right) \sum_{u=2}^{n} Y_{t,u-1} ^2 \right) = 0 \quad \text{for} \quad k = 1, 2, \ldots, M
\]
(4.3.10)

\[
\sum_{i=1}^{N} P(\omega_k | Z_i, \Theta(r)) \left[ (n-1) \rho_k \sum_{u=2}^{n} X_{t,u} - \frac{X_{t,u}}{\sigma_{\omega_k}} + \left( \frac{\phi_{11k} + \phi_{21k}}{\sigma_{\omega_k}} \right) X_{t,u-1} Y_{t,u-1} + \left( \frac{\phi_{21k} + \phi_{22k}}{\sigma_{\omega_k}} \right) Y_{t,u-1} ^2 \right]
\]
\[
- \left( \frac{\phi_{11k} X_{t,u} X_{t,u-1}}{\sigma_{\omega_k} ^2} + \frac{\phi_{12k} X_{t,u} Y_{t,u-1}}{\sigma_{\omega_k} ^2} - \left( \frac{\phi_{11k} \phi_{12k} + \phi_{21k} \phi_{22k}}{\sigma_{\omega_k} ^2} \right) X_{t,u-1} Y_{t,u-1} + \left( \frac{\phi_{21k} Y_{t,u} X_{t,u-1}}{\sigma_{\omega_k} ^2} + \frac{\phi_{22k} Y_{t,u} Y_{t,u-1}}{\sigma_{\omega_k} ^2} \right) Y_{t,u-1} ^2 \right)
\]

\[
+ \left( \frac{1+\rho_k ^2}{1-\rho_k ^2} \right) \sum_{u=2}^{n} X_{t,u} Y_{t,u} - \phi_{21k} X_{t,u} X_{t,u-1} - \phi_{22k} X_{t,u} Y_{t,u-1} - \phi_{11k} X_{t,u-1} Y_{t,u} + \phi_{12k} X_{t,u-1} ^2 \left[ Y_{t,u-1} \right] \right) = 0
\]
(4.3.11)

Solving the equations (4.3.4), (4.3.5), (4.3.6), (4.3.7), (4.3.8), (4.3.9), (4.3.10) and (4.3.11) simultaneously and iteratively the refined estimates of the model parameters

\[ P(\omega_k), \sigma_{\omega_k} ^2, \phi_{11k}, \phi_{12k}, \phi_{21k}, \phi_{22k}, \text{and } \rho_k \]

can be obtained.

**Expectation Maximization Algorithm**

**Step 1:** Find the initial parameters using equations (4.4.1) to (4.4.3) given in section (4.4).

**Step 2:** Obtain revised estimates of the parameters \[ P(\omega_k), \sigma_{\omega_k} ^2, \phi_{11k}, \phi_{12k}, \phi_{21k}, \phi_{22k}, \text{and } \rho_k \]

\[ \phi_{21k} \text{ and } \rho_k \] using equations (4.3.4), (4.3.5), (4.3.6), (4.3.7), (4.3.8),
Step 3: Repeat the process until the parameters do not change or the difference in successive computations is within the given threshold value.

Step 4: Write the final estimates of parameters $P(o_k), \sigma^2_{e_k}, \sigma^2_{\eta_k}, \phi_{11k}, \phi_{12k}, \phi_{21k}, \phi_{22k}$ and $\rho_k$.

4.4 INITIALISATION OF PARAMETERS

To initialize the parameters $\sigma^2_{e_k}, \sigma^2_{\eta_k}$ and $\rho_k$ using K-means algorithm one can obtain the number of classes (clusters) of the training data set. After classifying the time series data into $M$ clusters, the residual variance of each attribute for each class, the correlation coefficient between the attributes in each class and the model parameters within the class are obtained using the clustered time series of that class.

Thus the initial estimates of $\sigma^2_{e_k}$ and $\sigma^2_{\eta_k}$ can be taken as

$$
\hat{\sigma}^2_{e_k} = \frac{1}{(n-1)G_k} \sum_{s=1}^{G_k} \sum_{u=2}^{n} e^{2g}_{s,u} \quad \text{and} \quad \hat{\sigma}^2_{\eta_k} = \frac{1}{(n-1)G_k} \sum_{s=1}^{G_k} \sum_{u=2}^{n} \bar{e}^{2g}_{s,u}, \quad \text{for} \quad k=1,2,\ldots,M
$$

(4.4.1)

where $G_k$ is the number of time series in the $k^{th}$ cluster.

The initial estimate of $\rho_k$, the correlation coefficient between $X_t$ and $Y_t$ and is given by

$$
\hat{\rho}_k = \frac{1}{G_k} \sum_{s=1}^{G_k} \frac{\text{Cov}(X_t, Y_t)}{\sqrt{\text{Var}(X_t)\text{Var}(Y_t)}}, \quad k=1,2,\ldots,M
$$

(4.4.2)

where $G_k$ is the number of time series in the $k^{th}$ cluster.

For initializing parameters $\phi_{11k}, \phi_{12k}, \phi_{21k}$ and $\phi_{22k}$, ordinary least squares estimate of the parameters of the bivariate autoregressive process of order 1 are
obtained for each time series in each class and averaged over the number of time series in that class. The sample realization of bivariate autoregressive process of order 1 can be written in the following format

\[
\begin{bmatrix}
X_{t,2} & Y_{t,2} \\
X_{t,3} & Y_{t,3} \\
\vdots & \vdots \\
X_{t,n} & Y_{t,n}
\end{bmatrix}
= \begin{bmatrix}
X_{t,1} & Y_{t,1} \\
X_{t,2} & Y_{t,2} \\
\vdots & \vdots \\
X_{t,n-1} & Y_{t,n-1}
\end{bmatrix}
\begin{bmatrix}
\phi_{1t} & \phi_{2t} \\
\phi_{12t} & \phi_{22t}
\end{bmatrix}
+ \begin{bmatrix}
e_{X_{t,2}} & e_{Y_{t,2}} \\
e_{X_{t,3}} & e_{Y_{t,3}} \\
\vdots & \vdots \\
e_{X_{t,n-1}} & e_{Y_{t,n-1}}
\end{bmatrix}
\]

which can be represented as

\[
Y_t = X_t \Phi_t + \xi_t
\]

where, \(Y_t = \begin{bmatrix}
X_{t,2} & Y_{t,2} \\
X_{t,3} & Y_{t,3} \\
\vdots & \vdots \\
X_{t,n} & Y_{t,n}
\end{bmatrix} \), \(X_t = \begin{bmatrix}
X_{t,1} & Y_{t,1} \\
X_{t,2} & Y_{t,2} \\
\vdots & \vdots \\
X_{t,n-1} & Y_{t,n-1}
\end{bmatrix} \), \(\Phi_t = \begin{bmatrix}
\phi_{1t} & \phi_{2t} \\
\phi_{12t} & \phi_{22t}
\end{bmatrix} \) and \(\xi_t = \begin{bmatrix}
e_{X_{t,2}} & e_{Y_{t,2}} \\
e_{X_{t,3}} & e_{Y_{t,3}} \\
\vdots & \vdots \\
e_{X_{t,n-1}} & e_{Y_{t,n-1}}
\end{bmatrix} \)

Then the ordinary least squares estimate of \(\Phi_t\) is

\[
\hat{\Phi}_t = \left(X_t^T X_t\right)^{-1} X_t^T Y_t \quad \text{for } t=1,2,\ldots,N \quad (3.4.3)
\]

4.5 UNSUPERVISED LEARNING ALGORITHM

Here, the unsupervised learning algorithm is presented for identifying the new time series data with one of the available clusters. The steps involved in this algorithm are as follows:

Step 1: Draw scatter surface diagram for training data set in order to obtain the initial number of clusters for using the K-means algorithm.
Step 2: Using K means algorithm obtain the refined number of classes and elements in each cluster.

Step 3: Obtain initial estimates of model parameters using bivariate autoregressive process of order 1 and the least squares method of estimation given in section 4.4.

Step 4: Obtain the refined estimates of model parameters using updated equations of EM algorithm for bivariate AR process of order 1 given in section 4.3.

Step 5: For a new bivariate time series data set, compute the conditional likelihood with model parameters of \(i^{th}\) class derived from step 4 over \(1 \text{ to } K\) and assign it to a class for which the component conditional likelihood is maximum.

4.6 EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

In this section, the utility of the developed algorithm for clustering regions is demonstrated. After discussion with the people in pollution control board at Visakhapatnam, it is understood that 2 attributes \(\text{NO}_2\) and \(\text{SO}_2\) concentration in air are most important for taking control measures of pollutant concentration of a region. To impose the statutory norms rigorously in the regions of this area, it is required to segment the total area into various regions based on the bivariate \(\begin{bmatrix} \text{NO}_2 \\ \text{SO}_2 \end{bmatrix}\) concentration in air measured in microgram/m\(^3\). The number of regions in an area is not known as a priori and requires unsupervised learning methods to identify a sample time series with the region and analysis. Hence a study is carried out by
collecting the pollution concentration level of the bivariate \(\left[ \begin{array}{c} \text{NO}_2 \\ \text{SO}_2 \end{array} \right] \) in air over a period of 100 time points with a sample taken at \(\frac{1}{2}\) hour duration collected over 100 locations covering the whole area.

Using the K-means algorithm, the number of regions according to bivariate pollutant concentration in that area is determined. For implementing K-means algorithm, the initial number of regions is required. Hence, using training data the bivariate time series are plotted in scatter responses through a 3-dimensional graph shown in figure 4.1.

**Figure 4.1**

Scatter plot of \(\text{NO}_2\) and \(\text{SO}_2\) against time

From figure 4.1, it is observed that the time series are stationary after obtaining deviation from mean values and it is also observed that there are 3 surfaces which represent 3 regions having low pollution level, moderate pollution level and high pollution level respectively. Taking the initial number of clusters as 3, the K-
means algorithm is implemented on the training dataset of 100 time series found that the number of regions in that area also represent 3 levels of pollution concentration level air. The whole training data is classified into 3 segments representing 3 regions of pollution concentration level. Assuming that these time series are generated from bivariate AR(1) process of order 1, the model is trained. For determining the lag of AR process for each variant of every cluster, the autocorrelation and partial autocorrelations for each variant for a sample of time series are computed and correlograms are presented in figure 4.2 to figure 4.13.

**Figure 4.2**

Autocorrelation Function for NO₂
Low Pollution Region

**Figure 4.3**

Partial Autocorrelation Function for NO₂
Low Pollution Region

**Figure 4.4**

Autocorrelation Function for SO₂
Low Pollution Region

**Figure 4.5**

Partial Autocorrelation Function for SO₂
Low Pollution Region
Figure 4.6
Autocorrelation Function for NO\textsubscript{2} 
Moderate Pollution Region

Figure 4.7
Partial Autocorrelation Function for NO\textsubscript{2} 
Moderate Pollution Region

Figure 4.8
Autocorrelation Function for SO\textsubscript{2} 
Moderate Pollution Region

Figure 4.9
Partial Autocorrelation Function for SO\textsubscript{2} 
Moderate Pollution Region

Figure 4.10
Autocorrelation Function for NO\textsubscript{2} 
High Pollution Region

Figure 4.11
Partial Autocorrelation Function for NO\textsubscript{2} 
High Pollution Region
From figure 4.2 to figure 4.13, autocorrelogram tails off and partial autocorrelogram cut off at lag. Therefore the assumption that the sample time series of each class follows a bivariate autoregressive process of order 1 is verified.

Using the initialization of parameters discussed in section 4.4, the initial estimates of parameters \( P(\omega_k), \sigma_{e_k}, \sigma_{\epsilon_k^2}, \phi_{11k}, \phi_{12k}, \phi_{21k}, \phi_{22k}, \rho_k \) are obtained for three regions which corresponds to low pollution level region as cluster 1 with \( \text{NO}_2, \text{SO}_2 \) levels range from 0-40 microgram/m\(^3\), moderate pollution level region as cluster 2 with \( \text{NO}_2, \text{SO}_2 \) levels range from 40-80 microgram/m\(^3\) and high pollution level region as cluster 3 with \( \text{NO}_2, \text{SO}_2 \) levels range from 80-120 microgram/m\(^3\). The computed initial estimates of the model parameters are presented in Table 4.1.
Table 4.1
Initial Estimates of the Model Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Cluster 1 (Low Pollution Region)</th>
<th>Cluster 2 (Moderate Pollution Region)</th>
<th>Cluster 3 (High Pollution Region)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\omega_k)$</td>
<td>0.3333</td>
<td>0.3333</td>
<td>0.3333</td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon_k}$</td>
<td>4.6023</td>
<td>8.3990</td>
<td>11.1890</td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon_y}$</td>
<td>7.5955</td>
<td>13.5402</td>
<td>8.4042</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.0361</td>
<td>0.3843</td>
<td>0.329</td>
</tr>
<tr>
<td>$\phi_{11}$</td>
<td>-0.3793</td>
<td>-0.0202</td>
<td>0.1931</td>
</tr>
<tr>
<td>$\phi_{12}$</td>
<td>0.0456</td>
<td>0.0937</td>
<td>-0.2830</td>
</tr>
<tr>
<td>$\phi_{21}$</td>
<td>0.0358</td>
<td>0.0694</td>
<td>0.0966</td>
</tr>
<tr>
<td>$\phi_{22}$</td>
<td>-0.3097</td>
<td>0.0977</td>
<td>-0.3329</td>
</tr>
</tbody>
</table>

Using these initial estimates and EM algorithm, the refined estimates of parameters for each cluster are obtained and presented in Table 4.2.

Table 4.2
Final Estimates of the Model Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Cluster 1 (Low Pollution Region)</th>
<th>Cluster 2 (Moderate Pollution Region)</th>
<th>Cluster 3 (High Pollution Region)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\omega_k)$</td>
<td>0.268</td>
<td>0.523</td>
<td>0.209</td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon_k}$</td>
<td>4.206</td>
<td>9.605</td>
<td>10.340</td>
</tr>
<tr>
<td>$\sigma^2_{\epsilon_y}$</td>
<td>7.271</td>
<td>15.186</td>
<td>9.043</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.021</td>
<td>0.382</td>
<td>0.330</td>
</tr>
<tr>
<td>$\phi_{11}$</td>
<td>-0.394</td>
<td>-0.023</td>
<td>0.054</td>
</tr>
<tr>
<td>$\phi_{12}$</td>
<td>-0.023</td>
<td>0.074</td>
<td>-0.015</td>
</tr>
<tr>
<td>$\phi_{21}$</td>
<td>0.087</td>
<td>0.099</td>
<td>-0.132</td>
</tr>
<tr>
<td>$\phi_{22}$</td>
<td>-0.309</td>
<td>0.070</td>
<td>-0.283</td>
</tr>
</tbody>
</table>
With these final estimates, the model characterizing three clusters of regions are estimated as

**Cluster 1 (Low Pollution Region)**

\[
\begin{pmatrix}
X_t \\
Y_t
\end{pmatrix} = \begin{pmatrix}
-0.394 & -0.023 \\
0.087 & -0.309
\end{pmatrix}
\begin{pmatrix}
X_{t-1} \\
Y_{t-1}
\end{pmatrix} + \begin{pmatrix}
e_x \\
e_y
\end{pmatrix}
\]

**Cluster 2 (Moderate Pollution Region)**

\[
\begin{pmatrix}
X_t \\
Y_t
\end{pmatrix} = \begin{pmatrix}
-0.023 & 0.074 \\
0.099 & 0.07
\end{pmatrix}
\begin{pmatrix}
X_{t-1} \\
Y_{t-1}
\end{pmatrix} + \begin{pmatrix}
e_x \\
e_y
\end{pmatrix}
\]

**Cluster 3 (High Pollution Region)**

\[
\begin{pmatrix}
X_t \\
Y_t
\end{pmatrix} = \begin{pmatrix}
0.054 & -0.015 \\
-0.132 & -0.283
\end{pmatrix}
\begin{pmatrix}
X_{t-1} \\
Y_{t-1}
\end{pmatrix} + \begin{pmatrix}
e_x \\
e_y
\end{pmatrix}
\]

Here, \(X_t\) is Nitrogen dioxide (NO\(_2\)) concentration in a region at time \(t\) and \(Y_t\) is Sulphur dioxide (SO\(_2\)) concentration at time \(t\).

Therefore the model that characterizes whole data set is a three component mixture of bivariate autoregressive process of order 1 (BAR(1)) with component weights as \(P(\omega_1) = 0.268\), \(P(\omega_2) = 0.523\) and \(P(\omega_3) = 0.209\) respectively.

For evaluating the developed algorithm discussed in section 4.5, the test data consisting of 100 time series collected over 100 locations data points is considered. The developed unsupervised learning algorithm with bivariate AR(1) model identified 27 regions as low level pollution regions, 52 regions as moderate pollution level regions and 17 regions as high pollution level regions. For evaluating the performance
of the proposed algorithm, true positive rate (TPR), false positive rate (FPR), false discovery rate (FDR) and F-measure are used. For the proposed supervised learning algorithm of bivariate autoregressive process of order 1, the performance measures for each class are computed and presented in Table 4.3.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>True Positive Rate (TPR)</th>
<th>False Positive Rate (FPR)</th>
<th>False Discovery Rate (FDR)</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>0.9642</td>
<td>0.0138</td>
<td>0.0357</td>
<td>0.9641</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.9629</td>
<td>0.0217</td>
<td>0.0188</td>
<td>0.9719</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.8947</td>
<td>0.0123</td>
<td>0.0555</td>
<td>0.9788</td>
</tr>
</tbody>
</table>

To compare the efficiency of the developed bivariate AR(2) classifier with the earlier AR(2) classifiers, the true positive rate (TPR), false positive rate (FPR), false discover rate (FDR) and F-Measure are computed and presented in Table 4.4 and Table 4.5.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>True Positive Rate (TPR)</th>
<th>False Positive Rate (FPR)</th>
<th>False Discovery Rate (FDR)</th>
<th>F Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>0.9333</td>
<td>0.0142</td>
<td>0.03448</td>
<td>0.9491</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.9200</td>
<td>0.0200</td>
<td>0.0212</td>
<td>0.9484</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.8400</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.9130</td>
</tr>
</tbody>
</table>
From Table 4.3, Table 4.4 and Table 4.5, it is observed that the F value for all categories using the proposed classifier are more compared to that of the classifier with univariate AR(1) models for both the variables. This indicates that the proposed classifier with bivariate AR(1) model is much better in classifying time series data than the other two classifiers.

To compare the efficiency of developed unsupervised bivariate learning algorithm with existing univariate unsupervised learning algorithms with AR(1) model for both the attributes NO₂ and SO₂, the same test data have been considered and the misclassification rates are computed. Table 4.3 presents the misclassification rates of BAR(1), univariate AR(1) process of attribute NO₂ and univariate AR(1) process of attribute SO₂.
From Table 4, it is observed that the misclassification rate for bivariate AR(1) process is less compared to misclassification rate of univariate AR(1) process of the individual attributes NO$_2$ and SO$_2$. Therefore the developed unsupervised learning algorithm with bivariate autoregressive process of order 1 outperforms the existing unsupervised learning algorithm with univariate autoregressive process of order 1. This algorithm is useful in other application domains like medical, biological, financial, etc.