Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.12785/jsap/020205

Time Series Classification

M. M. Gabr, L. M. Fatehy

Department of Mathematics, Faculty of Science, Alexandria University, Egypt

Email Address: *mahgabr@yahoo.com, Lm_fatehy@yahoo.com*

Received: 9 Sept. 2012; Revised:6 Dec. 2012 ; Accepted: 17 Dec. 2012

Published online: 1 Jul. 2013

Abstract: Time series is an important class of temporal data objects and it can be easily obtained from scientific and financial applications. The nature time series data includes: large in data size, high dimensionality and necessary to update continuously. The increasing use of time series data has initiated a great deal of research and developing attempts in the field of data mining. Classification of time series data has a wide range of applications and has attracted researches from a wide range of discipline. In this paper the classical discriminant analysis is modified using the principal component analysis (PCA) to overcome the large dimensionality. The PCA modification can reduce the size of the data and improve the efficiency and accuracy. The new method is investigated using a simulation study to classify the linear AR(2) model and the bilinear BL(1,0,1,1) model. The results of our investigation show that the designed algorithm has a significant rate of correct classification especially if it is compared with the other methods. The PCA modification method is also applied to a real set of time series data and gave a superior rate of correct classification.

Keywords: Discriminant Analysis; time series classification; PCA; KNN

1 Introduction

Classification of time series has attracted much interest from the data mining community. The high dimensionality, high feature correlation, and typically high levels of noise found in time series provide an interesting research problem, (see Keogh and Kasetty (2002), Ye and Keogh (2009) and Zhang, Cheng, Li, Bian, and Tao (2012)). A time series often produces a pattern or features that may form a basic for discriminating between different classes. The problem of time series classification arises in many real-world fields. For example, in medicine, to distinguish the difference of ECG (electrocardiograms) signal between a normal person and a patient. In geophysical applications time series classifications methods have been used to discriminate between the earthquakes and nuclear explosions. In signal processing, detecting a radar signal, the time series classification has been used for discriminating between a pattern generated by a signal plus noise and a pattern generated by noise alone.

There are several classification methods used in time series classification such as classification trees, nearest neighbors, discriminant analysis, iterative classification, etc. (see Fu (2011) and Keogh and Kasetty (2002)).

The general problem of time series classification is to classify (or allocate) an observed time series **{X(t), t = 1, 2, …, N}** to one of k populations (or categories) $\pi_1, \pi_2, \ldots, \pi_k$ with small rate of error.

2 Discriminant Analysis

Discriminant analysis **(DA)** is a multivariate statistical technique, one of the data mining techniques concerned with separating distinct sets of objects (or observations) and with allocating new

objects to one of the pre-defined groups based on the knowledge of the multi-attributes. When the distribution within each group is multivariate normal, a parametric method can be used to develop a discriminate function using a generalized squared distance measure. The classification criterion is derived based on either the individual within-group covariance matrices or the pooled covariance matrix that also takes into account the prior probabilities of the classes. Non-parametric discriminant methods are based on non-parametric group-specific probability densities. Either a kernel or the k-nearest-neighbor method can be used to generate a non-parametric density estimate in each group and to produce a classification criterion.

The performance of a discriminant criterion could be evaluated by estimating the probability of miss-classification or the probability of correctly-classification, which known as "Accuracy", of new observations in the validation data. Accuracy can be calculated by the following formula:

$$\textbf{Accuracy} = \frac{\textbf{The number of correctly classified observations}}{\textbf{The total number of observations}}.$$

3 Separation and Classification for Two Populations

We assume that there are only two (k = 2) populations, π_1 and π_2 of interest to which the observed time series $\underline{X} = [X(1)\ \ X(2)\ \ \ldots\ X(N)]'$ can belong. Let us define $\mathbf{f_i(\underline{x})}$ as the probability density of $\underline{X}$ being in class π_i , $i = 1, 2$. Calculation of the overall total probability of error depends on the prior probability $\mathbf{p_i}$ of an observation belonging to the i^{th} class, i= 1, 2 where ($\mathbf{p_1 + p_2 = 1}$). Then the overall probability of classification error is minimized, (see Johnson and Wichern (2007)) by allocating or classifying $\underline{X}$ into π_1 if

$$\frac{\mathbf{f_1(\underline{x})}}{\mathbf{f_2(\underline{x})}} > \frac{\mathbf{p_2}}{\mathbf{p_1}}. \qquad \textbf{(1)}$$

4 Classification with Two Multivariate Normal Populations

Assuming that $\underline{X}$ is stationary normally distributed in each class and that the means and covariance matrices are $\underline{\mu}_1, \underline{\mu}_2, \mathbf{R_1}$ and $\mathbf{R_2}$ respectively, so that the joint densities of $\underline{X} = [X(1)\ \ X(2)\ \cdots X(N)]'$ for π_1 and π_2 are given by

$$\mathbf{f_i(\underline{x}) = (2\pi)^{\frac{-N}{2}} |R_i|^{-1/2} e^{-1/2(\underline{x}-\underline{\mu}_i)' R_i^{-1} (\underline{x}-\underline{\mu}_i)}} \quad , \quad \mathbf{i = 1, 2.} \qquad \textbf{(2)}$$

By substituting (2), with $\mathbf{R_1 \neq R_2}$, into (1) and taking the natural logarithm we obtain the following ***quadratic classifier***

$$\begin{aligned} \mathbf{d_Q(\underline{x})} &= \mathbf{\ln \frac{f_1(\underline{x})}{f_2(\underline{x})}} \\ &= -\frac{1}{2}\ln\frac{|\mathbf{R}_1|}{|\mathbf{R}_2|} - \frac{1}{2}\underline{\mathbf{x}}'(\mathbf{R}_1^{-1} - \mathbf{R}_2^{-1})\underline{\mathbf{x}} + (\underline{\mu}_1'\mathbf{R}_1^{-1} - \underline{\mu}_2'\mathbf{R}_2^{-1})\underline{\mathbf{x}} - \frac{1}{2}\underline{\mu}_1'\mathbf{R}_1^{-1}\underline{\mu}_1 \qquad \textbf{(3)} \\ &\quad + \frac{1}{2}\underline{\mu}_2'\,\mathbf{R}_2^{-1}\underline{\mu}_2 + \ln(\frac{\mathbf{p}_1}{\mathbf{p}_2}) \end{aligned}$$

The decision rule becomes allocate $\underline{\mathbf{x}}$ into π_1 if $\mathbf{d_Q(\underline{x}) \geq 0}$ and into π_2 otherwise. A difficulty in this case is that the classifier $\mathbf{d_Q(\underline{x})}$ is a quadratic function of $\underline{\mathbf{x}}$ and therefore its distribution is intractable under either π_1 or π_2 and so the theoretical error rates, in this case, is a very difficult to obtain. Moreover the computation required for $\mathbf{d_Q(\underline{x})}$ is excessive and $\mathbf{R_1}$ and $\mathbf{R_2}$ may not known, even if they are, their inversion may pass computational problems especially for long series.

In many applications the populations are differ only in means and are common in covariance matrix **R**, say. Using a common covariance matrix assumption the two populations have the same covariance matrix $\mathbf{R_1 = R_2 = R}$ and differ only by their means, then (3) is simplified to the new classifier given by

$$\mathbf{d_L(\underline{x}) = (\underline{\mu}_1 - \underline{\mu}_2)'R^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)'R^{-1}(\underline{\mu}_1 + \underline{\mu}_2) + \ln(\frac{p_1}{p_2}),} \qquad \textbf{(4)}$$

and $\mathbf{d_L(\underline{x})}$ is clearly ***a linear function*** of $\underline{\mathbf{X}}$, we classify $\underline{\mathbf{X}}$ into π_1 or π_2 according to whether $\mathbf{d_L(\underline{x}) \geq 0}$ or $\mathbf{d_L(\underline{x}) < 0}$. The linear version of the discriminate analysis (LDA) thus has drawn a hyper-plane in the space $\mathbf{R^N}$ where N is the dimension of $\underline{\mathbf{x}}$, this hyper-plane being the decision border between class π_1 and class π_2.

One should note that what we just did is nothing more than using the log-likelihood criterion with jointly normally distributed random variables. For the case $\mathbf{p_1 = p_2}$, $\mathbf{d_L(\underline{x})}$ is a normal random variable with means $\frac{1}{2}\mathbf{D_N^2}$ under π_1 and $-\frac{1}{2}\mathbf{D_N^2}$ under π_2 and variance $\mathbf{D_N^2}$ under both hypotheses, where

$$\mathbf{D_N^2 = (\underline{\mu}_1 - \underline{\mu}_2)'R^{-1}(\underline{\mu}_1 - \underline{\mu}_2)} \qquad \textbf{(5)}$$

is the generalized Mahalanobis distance between the mean vectors $\underline{\mu}_1$ and $\underline{\mu}_2$, for more details see Johnson and Wichern (2007). The error probability to either π_1 or to π_2 is $\Phi(-\frac{1}{2}\mathbf{D_N})$ and the probability of correct classification to either class is $\Phi(\frac{1}{2}\mathbf{D_N})$.

The above discriminant function contains some unknown values of the means $\underline{\mu}_1$, $\underline{\mu}_2$ and the covariance matrix **R**. Note that, since $\underline{\mathbf{X}} = [\mathbf{X(1)\ \ X(2)} \cdots \mathbf{X(N)}]'$ is stationary, then the covariance matrix **R** is given by

$$\mathbf{R = Cov(\underline{X}) = E\left[\left(\underline{X} - E(\underline{X})\right)\left(\underline{X} - E(\underline{X})\right)'\right]}$$

$$= \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(N-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(N-2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma(N-1) & \gamma(N-2) & \cdots & \gamma(0) \end{bmatrix}$$

where $\gamma(k) = Cov(X(t), X(t+k)) = E[(X(t) - \mu)(X(t+k) - \mu)]$.

Let $\underline{\mathbf{x}}^{(11)}, \underline{\mathbf{x}}^{(12)}, \ldots, \underline{\mathbf{x}}^{(1n_1)}$ and $\underline{\mathbf{x}}^{(21)}, \underline{\mathbf{x}}^{(22)}, \ldots, \underline{\mathbf{x}}^{(2n_2)}$ be two independent samples of dimension N×1 from the populations π_1 and π_2 respectively. Then

$$\hat{\underline{\mu}}_i = \bar{\underline{x}}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} \underline{x}^{(ij)}(t) \qquad , \quad i = 1, 2 \qquad \textbf{(6)}$$

and

$$\hat{\mathbf{R}} = \mathbf{S} = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(N-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(N-2) \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\gamma}(N-1) & \hat{\gamma}(N-2) & \cdots & \hat{\gamma}(0) \end{bmatrix}$$

where

$$\hat{\gamma}(k) = \frac{1}{(n_1+n_2)N} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \sum_{t=1}^{N-k} (x^{(ij)}(t) - \bar{\underline{x}}^{(i)})(x^{(ij)}(t+k) - \bar{\underline{x}}^{(i)})$$

$$\hat{\mathbf{R}} = \mathbf{S} = \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\underline{x}^{(ij)} - \bar{\underline{x}}^{(i)})(\underline{x}^{(ij)} - \bar{\underline{x}}^{(i)})' \qquad (7)$$

Then, the linear discriminant function (4) can be proceeded as

$$\hat{d}_L(\underline{x}) = (\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)})'\hat{\mathbf{R}}^{-1}\underline{x} - \frac{1}{2}(\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)})'\hat{\mathbf{R}}^{-1}(\bar{\underline{x}}^{(1)} + \bar{\underline{x}}^{(2)}) + \ln\left(\frac{p_1}{p_2}\right) \qquad (8)$$

and

$$\hat{D}_N^2 = (\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)})'\hat{\mathbf{R}}^{-1}(\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)}) \qquad (9)$$

5 PCA Modification

As we mentioned above the computations of either the linear or quadratic discriminant function given by equation (3) and (4) are rather cumbersome matrix calculations specially if the time series is long (N-the dimension of $\underline{\mathbf{x}}$ is large), therefore the construction of the PCA coefficients, (see Al-Kandar and Jolliffe (2001) and Vines (2000)), simplifying greatly the classifier since they are uncorrelated and reducing the matrix multiplication in simple sums (instead of double).

5.1 Linear discriminant function

Now, suppose the data $\underline{x}^{(11)}, \underline{x}^{(12)}, ..., \underline{x}^{(1n_1)}$ and $\underline{x}^{(21)}, \underline{x}^{(22)}, ..., \underline{x}^{(2n_2)}$ represent two independent samples of dimension N×1 from the populations π_1 and π_2 respectively. These data yield the sample mean vectors $\bar{\underline{x}}^{(1)}$, $\bar{\underline{x}}^{(2)}$ and the sample covariance matrix $\hat{\mathbf{R}} = \mathbf{S}$ defined by (6) and (7) respectively. Let $(\hat{\lambda}_1, \hat{\underline{e}}_1), (\hat{\lambda}_2, \hat{\underline{e}}_2), ..., (\hat{\lambda}_N, \hat{\underline{e}}_N)$ be the Eigen value and Eigen vector pairs of the sample covariance matrix $\hat{\mathbf{R}}$, the i[th] sample principal component is given by

$$\hat{Y}_i = \hat{\underline{e}}_i' \underline{X} = \hat{e}_{i1} X_1 + \cdots + \hat{e}_{iN} X_N \quad , \qquad i=1, 2, \ldots, N \qquad (10)$$

where $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_N$ are the roots of the equation $|\hat{\mathbf{R}} - \hat{\lambda} \mathbf{I}_N| = 0$ and $\underline{\mathbf{X}}$ is any observed time series of the two samples. Also

$$\text{Sample variance } (\hat{Y}_i) = \hat{\lambda}_i \quad , \qquad i=1, 2, \ldots, N$$

$$\text{Sample covariance } (\hat{Y}_i, \hat{Y}_j) = 0 \quad , \qquad i \neq j.$$

In addition,

$$\textbf{Total time series variance} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_N$$

Let $\mathbf{P} = [\underline{\hat{e}}_1\ \underline{\hat{e}}_2 \cdots \underline{\hat{e}}_2]'$, then by virtue of the orthonormality of the eigen vectors, the principal components (10) can be written in matrix form as

$$\underline{Y} = \mathbf{P}'\underline{X} \quad \Leftrightarrow \quad \underline{X} = \mathbf{P}\underline{Y} \tag{11}$$

Clearly we have $\mathbf{PP}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$ which implies that $\mathbf{P}^{-1} = \mathbf{P}'$ and $\mathbf{R} = \mathbf{P}\Lambda\mathbf{P}' \Leftrightarrow \Lambda = \mathbf{PRP}' \Leftrightarrow \mathbf{R}^{-1} = \mathbf{P}\Lambda^{-1}\mathbf{P}'$ where Λ is the diagonal matrix of the Eigen values

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

Thus,

$$\underline{\mu}(\underline{Y}) = \begin{cases} \mathbf{P}'\underline{\mu}_1(\underline{X}) & \text{in} \quad \pi_1, \\ \mathbf{P}'\underline{\mu}_2(\underline{X}) & \text{in} \quad \pi_2. \end{cases} \tag{12}$$

Substituting in the linear discriminant function (4) we obtain

$$\begin{aligned} d_L(\underline{Y}) &= (\underline{\mu}_1(\underline{Y}) - \underline{\mu}_2(\underline{Y}))'\mathbf{P}'(\mathbf{P}\Lambda^{-1}\mathbf{P}')(\mathbf{P}\underline{Y}) - \frac{1}{2}(\underline{\mu}_1(\underline{Y}) - \underline{\mu}_2(\underline{Y}))'\mathbf{P}'(\mathbf{P}\Lambda^{-1}\mathbf{P}')\mathbf{P}(\underline{\mu}_1(\underline{Y}) \\ &\quad + \underline{\mu}_2(\underline{Y})) + \ln\left(\frac{p_1}{p_2}\right) \\ &= \sum_{j=1}^{N} \frac{\mu_{1j}(\underline{Y}) - \mu_{2j}(\underline{Y})}{\lambda_j} Y_j - \frac{1}{2}\sum_{j=1}^{N} \frac{\mu_{1j}^2(\underline{Y}) - \mu_{2j}^2(\underline{Y})}{\lambda_j} + \ln\left(\frac{p_1}{p_2}\right) \end{aligned} \tag{13}$$

Note that:

$$E[d_L(\underline{Y})/\pi_1] = \sum_{j=1}^{N} \frac{(\mu_{1j}(\underline{Y}) - \mu_{2j}(\underline{Y}))^2}{2\lambda_j}, \tag{14}$$

$$E[d_L(\underline{Y})/\pi_2] = -\sum_{j=1}^{N} \frac{(\mu_{1j}(\underline{Y}) - \mu_{2j}(\underline{Y}))^2}{2\lambda_j}, \tag{15}$$

and,

$$\mathbf{Var}[d_L(\underline{Y})] = \sum_{j=1}^{N} \frac{(\mu_{1j}(\underline{Y}) - \mu_{2j}(\underline{Y}))^2}{\lambda_j} \tag{16}$$

5.2 Quadratic discriminant function

Let $(\hat{\lambda}_{11}, \underline{\hat{e}}_{11}), (\hat{\lambda}_{12}, \underline{\hat{e}}_{12}), \ldots, (\hat{\lambda}_{1N}, \underline{\hat{e}}_{1N})$ and $(\hat{\lambda}_{21}, \underline{\hat{e}}_{21}), (\hat{\lambda}_{22}, \underline{\hat{e}}_{22}), \ldots, (\hat{\lambda}_{2N}, \underline{\hat{e}}_{2N})$ are the Eigen values and Eigen vector pairs of the sample covariance matrices

$$\hat{\mathbf{R}}_i = \begin{bmatrix} \hat{\gamma}_i(0) & \hat{\gamma}_i(1) & \ldots & \hat{\gamma}_i(N-1) \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \hat{\gamma}_i(N-1) & \hat{\gamma}_i(N-2) & \ldots & \hat{\gamma}_i(0) \end{bmatrix} \quad i = 1, 2$$

respectively, where

$$\hat{\lambda}_{i1} \geq \hat{\lambda}_{i2} \geq \ldots \geq \hat{\lambda}_{iN} \geq 0 \quad \text{and} \quad \hat{\mathbf{R}}_i \hat{\underline{e}}_{ij} = \hat{\lambda}_{ij} \hat{\underline{e}}_{ij}, \quad i = 1,2 \; ; \; j = 1,2,\ldots,N,$$

Let

$$\mathbf{P}_i = [\hat{\underline{e}}_{i1} \; \hat{\underline{e}}_{i2} \ldots \hat{\underline{e}}_{iN}]' \text{ then } \mathbf{P}_i \mathbf{P}_i' = \mathbf{P}_i' \mathbf{P}_i = \mathbf{I} \Leftrightarrow \mathbf{P}_i^{-1} = \mathbf{P}_i', \text{ i=1, 2.}$$

With these choices, we have

$$\hat{\mathbf{R}}_i = \mathbf{P}_i \hat{\Lambda}_i \mathbf{P}_i' \Leftrightarrow \hat{\Lambda}_i = \mathbf{P}_i' \hat{\mathbf{R}}_i \mathbf{P}_i = \begin{bmatrix} \hat{\lambda}_{i1} & 0 & \ldots & 0 \\ 0 & \hat{\lambda}_{i2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \hat{\lambda}_{iN} \end{bmatrix} \Leftrightarrow \hat{\mathbf{R}}_i^{-1} = \mathbf{P}_i \hat{\Lambda}_i^{-1} \mathbf{P}_i', \; i = 1,2$$

and

$$\text{In } \pi_1 : \; \underline{Y} = \mathbf{P}_1' \underline{X} \quad \Rightarrow \quad \underline{X} = \mathbf{P}_1 \underline{Y} \Rightarrow \quad \underline{\mu}(\underline{Y}) = \mathbf{P}_1' \hat{\underline{\mu}}_1(\underline{x}) \quad , \text{cov}(\underline{Y}) = \Lambda_1.$$

$$\text{In } \pi_2 : \underline{Y} = \mathbf{P}_2' \underline{X} \quad \Rightarrow \quad \underline{X} = \mathbf{P}_2 \underline{Y} \Rightarrow \quad \underline{\mu}(\underline{Y}) = \mathbf{P}_2' \hat{\underline{\mu}}_2(\underline{x}) \quad , \text{cov}(\underline{Y}) = \Lambda_2.$$

Substituting in the quadratic discriminant function (3) and assuming $\mathbf{p}_1 = \mathbf{p}_2$ so that $\ln(\frac{\mathbf{p}_1}{\mathbf{p}_2}) = 0$, we obtain:

$$\mathbf{d}_Q(\underline{y}) = -\frac{1}{2} \ln \frac{|\mathbf{p}_1 \Lambda_1 \mathbf{p}_1'|}{|\mathbf{p}_2 \Lambda_2 \mathbf{p}_2'|} - \frac{1}{2} \underline{x}' (\mathbf{P}_1 \Lambda_1^{-1} \mathbf{P}_1' - \mathbf{P}_2 \Lambda_2^{-1} \mathbf{P}_2') \underline{x}$$

$$+ (\underline{\mu}_1' \mathbf{p}_1 \Lambda_1^{-1} \mathbf{p}_1' - \underline{\mu}_2' \mathbf{p}_2 \Lambda_2^{-1} \mathbf{p}_2')' \underline{x} - \frac{1}{2} \underline{\mu}_1' \mathbf{p}_1 \Lambda_1^{-1} \mathbf{p}_1' \underline{\mu}_1 + \frac{1}{2} \underline{\mu}_2' \mathbf{p}_2 \Lambda_2^{-1} \mathbf{p}_2' \underline{\mu}_2 \quad (17)$$

In π_1:

$$\underline{x}' \mathbf{P}_1 \Lambda_1^{-1} \mathbf{P}_1' \underline{x} = \underline{y}' \mathbf{P}_1' \mathbf{P}_1 \Lambda_1^{-1} \mathbf{P}_1' \mathbf{P}_1 \underline{y} = \underline{y}' \Lambda_1^{-1} \underline{y} = \sum_{i=1}^{N} \frac{y_i^2}{\lambda_{1i}},$$

$$\underline{\mu}'_1(\underline{y}) \mathbf{P}_1 \Lambda_1^{-1} \mathbf{P}_1' \underline{x} = \underline{\mu}_1 \mathbf{P}_1' \mathbf{P}_1 \Lambda_1^{-1} \mathbf{P}_1' \mathbf{P}_1 \underline{y} = \underline{\mu}_1 \Lambda_1^{-1} \underline{y} = \sum_{i=1}^{N} \frac{\mu_{1i} y_i}{\lambda_{1i}},$$

$$\underline{\mu}'_1(\underline{y}) \mathbf{P}_1 \Lambda_1^{-1} \mathbf{P}_1' \underline{\mu}_1(\underline{y}) = \underline{\mu}'_1 \mathbf{P}_1' \mathbf{P}_1 \Lambda_1^{-1} \mathbf{P}_1' \mathbf{P}_1 \underline{\mu}_1 = \underline{\mu}'_1 \Lambda_1^{-1} \underline{\mu}_1 = \sum_{i=1}^{N} \frac{\mu_{1i}^2}{\lambda_{1i}}. \quad (18)$$

Similarly, in π_2:

$$\underline{x}' \mathbf{P}_2 \Lambda_2^{-1} \mathbf{P}_2' \underline{x} = \mathbf{y}' \mathbf{P}_2' \mathbf{P}_2 \Lambda_2^{-1} \mathbf{P}_2' \mathbf{P}_2 \underline{y} = \mathbf{y}' \Lambda_2^{-1} \underline{y} = \sum_{i=1}^{N} \frac{y_i^2}{\lambda_{2i}},$$

$$\underline{\mu}'_2(\underline{y}) \mathbf{P}_2 \Lambda_2^{-1} \mathbf{P}_2' \underline{x} = \underline{\mu}_2 \mathbf{P}_2' \mathbf{P}_2 \Lambda_2^{-1} \mathbf{P}_2' \mathbf{P}_2 \underline{y} = \underline{\mu}_2 \Lambda_2^{-1} \underline{y} = \sum_{i=1}^{N} \frac{\mu_{2i} y_i}{\lambda_{2i}}, \quad (19)$$

$$\underline{\mu}'_2(\underline{y}) \mathbf{P}_2 \Lambda_2^{-1} \mathbf{P}_2' \underline{\mu}_2(\underline{y}) = \underline{\mu}'_2 \mathbf{P}_2' \mathbf{P}_2 \Lambda_2^{-1} \mathbf{P}_2' \mathbf{P}_2 \underline{\mu}_2 = \underline{\mu}'_2 \Lambda_2^{-1} \underline{\mu}_2 = \sum_{i=1}^{N} \frac{\mu_{2i}^2}{\lambda_{2i}}.$$

Then,

$$\mathbf{d}_Q(\underline{y}) = -\frac{1}{2} \ln \left(\frac{\prod_{i=1}^{N} \lambda_{1i}}{\prod_{i=1}^{N} \lambda_{2i}} \right) - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{1}{\lambda_{1i}} - \frac{1}{\lambda_{2i}} \right) y_i^2 + \frac{1}{2} \sum_{j=1}^{N} \left(\frac{\mu_{1i}}{\lambda_{1i}} - \frac{\mu_{2i}}{\lambda_{2i}} \right) y_i - \frac{1}{2} \sum_{i=1}^{N} \left(\frac{\mu_{1i}^2}{\lambda_{1i}} - \frac{\mu_{2i}^2}{\lambda_{2i}} \right) \quad (20)$$

The summations and products in (13) and (20) are reduced to only k components instead of N (<< N) if the first k principal components for the estimated covariance matrix $\hat{\mathbf{R}}$ contribute at least 90% of the total variation. Thus k is chosen so that

$$\sum_{i=1}^{k}\hat{\lambda}_i \Big/ \sum_{i=1}^{N}\hat{\lambda}_i \geq 0.9$$

6 Applications

In this section the new PCA-modified discriminant analysis method is applied to both simulated and real datasets. This new method is also compared with the other known time series classification methods.

6.1 Simulated series examples

Two sets of independent normally distributed with mean zero and variance one white noises $\{a_1(t)\}$ and $\{a_2(t)\}$ are generated by using the statistical packages MINITAB14, using these $\{a_1(t)\}$ and $\{a_2(t)\}$, 40 time series of $\{x(t); t = 1, 2, \ldots, 128\}$ from Autoregressive model **AR(2)** and 30 time series of $\{y(t); t = 1, 2, \ldots, 128\}$ from Bilinear Autoregressive model **BL(1,0,1,1)** are generated with length 128 (dimension = N = 128) as follows:

$$X(t) = 4.5 + 0.2\,X(t-1) - 0.9\,X(t-2) + a_1(t),$$

and,

$$Y(t) = 1.5 + 0.6\,Y(t-1) + 0.4\,Y(t-1)\,a_2(t-1) + a_2(t).$$

The coefficients in the two models $\{X(t)\}$ and $\{Y(t)\}$ are chosen so that the two series are overlapping and satisfy the stationarity conditions. Figure 1. shows a plot of a series from the AR model and a series from the BL model. The overlapping between the two series is clear; moreover the means of the two series, theoretical and sampling are close to each other.

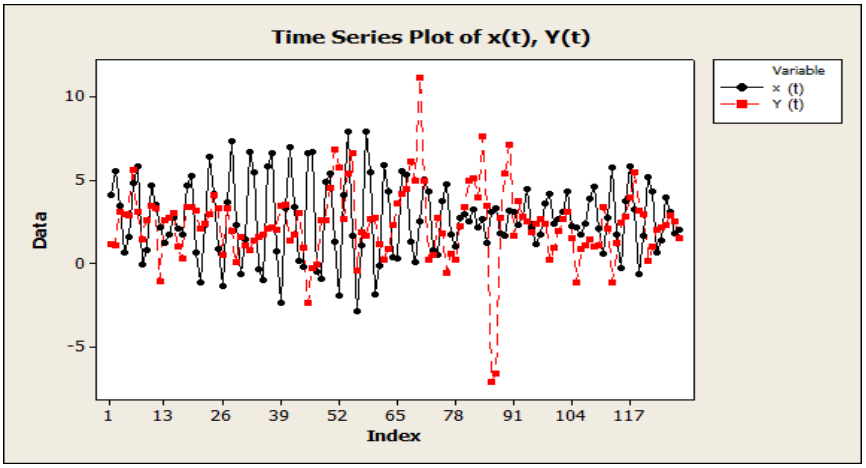

Figure 1.

Now, we are going to discriminate between the above 70 generated time series described above as two different populations (groups) in which 40 time series belong to one population and the others 30 time series belong to another population by using different known methods and our PCA modified method.

(A) Discriminant Analysis (DA) method

Using the statistical package Minintab14, and by applying the "multivariate discriminant analysis", the discrimination cannot be done and we got an error.

Be care, we can discriminate between the two groups only for series with fewer dimensions, for example, it can be done by taking $N \leq 66$ in linear case and $N \leq 26$ in quadratic case, these values of N according to our several trails in our practical work.

(B) Mahalanobis distance using R 2.15.2

Applying the Mahalanobis distance measure by using "R 2.15.2 Package" by writing the suitable code, the classification cannot be done and we have got the error message "system is computationally singular". Also**,** according to our several trails in our practical work, we can classify between the two groups only for series with fewer dimensions, which can be done by taking $N \leq 30$.

(C) PCA modification method

Now, we are going to apply our PCA- modification method that described in section 5. To solve the problem of classification between the two groups that failed in the previous methods, our modification concerning on the linear and quadratic discriminate functions (13) and (20) respectively, then calculate the classification accuracy in each case to differentiate between them.

I. PCA-modification with linear discriminate function

We have two populations π_1 and π_2 in which $n_1=40$ belong to π_1 from the AR model, $n_2=30$ belong to π_2 from the BL model with N=128 for each. Writing a suitable code using "Mathematica 8" to calculate the estimated variance covariance matrix of $\hat{\mathbf{R}}$ (128 x 128) that defined in (7). Calculate the k principal components for the above estimated covariance matrix $\hat{\mathbf{R}}$ which contribute at least 90% of the total variation by writing the suitable Matlab code. We found that there are k = 65 principals components contributes 90% of all variation,

so $(\hat{\lambda}_1, \hat{\underline{e}}_1), (\hat{\lambda}_2, \hat{\underline{e}}_2), \ldots, (\hat{\lambda}_{65}, \hat{\underline{e}}_{65})$ are the Eigen values and Eigen vectors pairs of the covariance matrix.

Now, we have 65 principal components that will be represented as a linear combination of the Eigen-vectors using Equation (13).

Calculate the two population means $\underline{\mu}_1(\underline{Y})$ and $\underline{\mu}_2(\underline{Y})$ by using Equations (6) and (12), and then substituting in the modified linear discriminant function (13) to classify between the two populations according to its value, if it is ≥ 0 then we classify the series as it belongs to the first population π_1, (the Autoregressive model AR(2)), otherwise classify it as the second population π_2 (the Bilinear model BL(1,0,1,1)).

II. Applying the PCA-modification in quadratic discriminant function

In the PCA quadratic discriminate method we need to calculate the principal components for the estimated covariance matrices $\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_2$ which contribute 90 % of the total variation and complete in the same way as in the linear discriminate modified method.

(D) Applying the K-Nearest Neighbor (KNN) method

Appling the known KNN method using **weka 3.6** and also by using **R 2.15.2** packages by writing the suitable code for implementation, then comparing between our PCA modification in linear and quadratic discriminant function and the KNN method, remembering the failing in the usual discriminate analysis and the Mahalanobis distance measure methods, we have got the following results described in Table 1.:

6.1.1 Conclusions

In the two predefined models the average of all 40 series of {X(t)}'s = 2.64 while the average of all 30 series of {Y(t)}'s = 2.45, which are too close to each other.

Applied Method		Number of misclassified observations	Overall Accuracy
DA		Failed	___
Mahalanobis Distance		Failed	___
Linear PCA-modification		6	0.91
Quadratic PCA-modification		0	1
KNN	***K = 3***	22	0.687
	K= 5	24	0.657
	K= 4, 7	27	0.614
	K=10	29	0.586

Table 1.

Hence with this small difference between the two averages the KNN method becomes difficult to classify between the two groups, and when applying the PCA modification in quadratic discriminate function, the number of misclassified observations are less than the number of misclassified observations when applying the linear discriminate function. This result clears that our PCA modification method produces a road map way for classification with great accuracy rather than the other methods.

6.2 Real–World Case Study - ECG Data.

In this section, we apply our PCA-modification method to a real-world time series ECD data sets from UCR time Series data sets (Keogh, Xi, Wei, and Ratanamahatana (2006)). The electrocardiogram (ECG) is a recording of body surface potentials generated by the electrical activity of the heart. It is a

two-class disease time series classification problem with length = 96, a "normal" class (the majority/negative class) with many examples = 69, and very few examples of the "abnormal" (the minority/positive class) with number of examples = 31. ECG dataset can be plotted using "Minitab 14" by taking one example from the "normal" class as "pop1" and other example from the "abnormal" class as "pop2" and plotting them as given in Figure 2. and Figure 3.

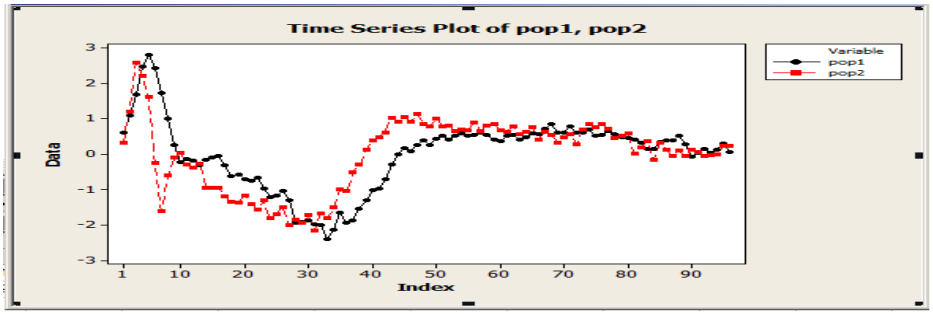

Figure 2.

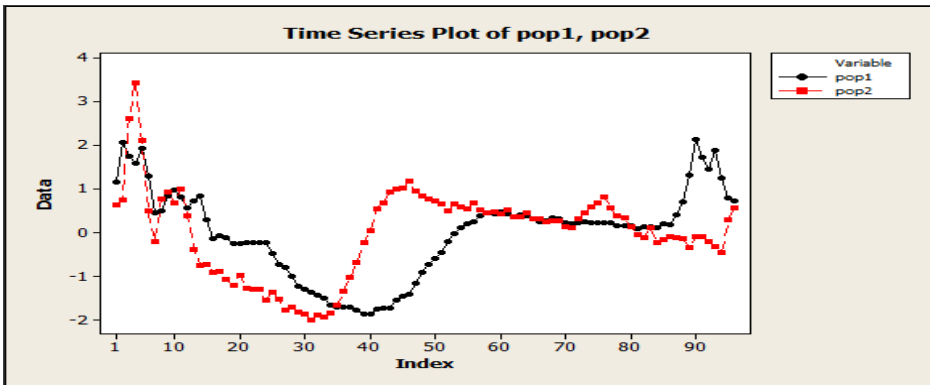

Figure 3.

Applying our PCA-modification method for the ECG data set and using the same criteria for determining the value of k principal components, it was found to be 25 components are required to represent 90% of the total variation, then we completed in the same way as in simulated series to classify the groups. And referring to Koknar-Tezel and Latecki (2009 and 2010) for using the SVM and SVM-GP classification methods, our results can be showed in Table 2.

Applied Method	No. of misclassified observations	Overall Accuracy
LDA	Failed	___
Mahalanobis measure	Failed	___
Linear PCA-modification	14	0.86
Quadratic PCA-modification	13	0.87
SVM	20	0.800
SVM-GP	21	0.79
KNN (with K= 10)	22	0.77

Table 2.

6.2.1 Conclusions

Table 2., showed that the "Overall Accuracy" when applying the PCA-modification in the Quadratic case is 0.87 with 13 misclassified observations, while in applying the PCA-modification in the Linear case it was 0.86 with 14 misclassified observations, in each case the result was better than applying the SVM and SVM-GP methods. In summary, our PCA-modification method for the ECG data set, has achieved a better performance than the traditional SVM method and the improved SVM method (SVM-GP).

References

[1] Al-Kandari, N.M. and Jolliffe, I.T. (2001) "Variable selection and interpretation of covariance principal components". Commun. Statist.—Simul. Computat., **30**, 339-354.

[2] Bishop, C. M. (2006) "*Pattern Recognition and Machince Learning"*. Springer, Cambridge.

[3] Cichocki, A. and Amari, S. (2002) "*Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*". Wiley.

[4] Fu, T. C. (2011) "A review on time series data mining. Engineering Applications of Artificial Intelligence", **24**(1), 164-181.

[5] Johnson, R. A. and Wichern, D. W. (2007) "*Applied Multivariate Statistical Analysis (6th Edition)* ". Prentice Hall.

[6] Jolliffe, T. (2002) "*Principal Component Analysis (2nd edition)*". Springer-Verlag.

[7] Keogh, E. and Kasetty, S. (2002) "On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration". *In proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* July 23 - 26, 2002. Edmonton, Alberta, Canada. pp 102-111.

[8] Keogh, E. , Xi, X. , Wei, L. and Ratanamahatana, C. A. (2006) "The UCR time series classification and clustering home page: http://www.cs.ucr.edu/~eamonn/ Time series data/".

[9] Keogh, E. and Kasetty, S. (2002) "On the need for time series data mining benchmarks: A survey and empirical demonstration". In proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 102-111.

[10] K¨oknar-Tezel, S. and Latecki, L. J. (2009) "Improving SVM Classification on Imbalanced Data Sets in Distance Spaces". IEEE Int. Conf. on Data Mining (ICDM, December 2009), Miami, Florida, USA.

[11] K¨oknar-Tezel, S. and Latecki, L. J. (2010) "Improving SVM Classification on Imbalanced Time Series Data Sets with Ghost Points". Knowledge and Information Systems.

[12] Li Wei , Eamonn Keogh (2006) "Semi-supervised time series classification", Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, August 20-23, 2006, Philadelphia, PA, USA.

[13] Shumway, R. H. (1988) "*Applied Statistical Time Series Analysis*". *Englewood Cliffs, New Jersey: Prentice Hall*.

[14] Vines, S. (2000) "Simple principal components," Applied Statistics, **49**, 441–451.

[15] Ye, L. and Keogh, E. (2009) "Time Series Shapelets: A New Primitive for Data Mining". In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris, France, p. 947–956.

[16] Zhang, Z. , Cheng, J. , Li, J. , Bian, W. and Tao, D. (2012) "Segment-Based Features for Time Series Classification"- The Computer Journal, 2012 **55**(9): 1088-1102.