ejpmr, 2016,3(7), 451-456.

SJIF Impact Factor 3.628

EUROPEAN JOURNAL OF PHARMACEUTICAL AND MEDICAL RESEARCH
www.ejpmr.com

Review Article
ISSN 2394-3211
EJPMR

DETECTING OUTLIERS IN MULTIVARIATE DATA

***[1]Dr. N. Senthil Vel Murugan, [2]Dr. V. Vallinayagam and [3]Dr. K. Senthamarai Kannan**

[1]Department of Mathematics, Rohini College of Engineering and Technology, Kanyakumari, 9444544739.
[2]Department of Mathematics, St. Joseph's College of Engineering, Chennai, 9444489146.
[3]Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli. 9443436364.

***Corresponding Author: Dr. N. Senthil Vel Murugan**
Department of Mathematics, Rohini College of Engineering and Technology, Kanyakumari, 9444544739.

Article Received on 20/05/2016 **Article Revised on 10/06/2016** **Article Accepted on 30/06/2016**

ABSTRACT

In this paper a method for constructing Principal Component Analysis and the data taken for study is DNA sequence of samples affected by liver cancer. It can be inferred from the analysis that increases or decrease in protein level, hormone level contributes to liver cancer. The aim of this paper is to analyze the affected liver cancer DNA sequence data and extract the possible factors using Principal Component Analysis and it also deals with the outlier detection by distance measures. The reasonable results verify the validity of our method.

KEYWORDS: DNA; Data Mining; Principal Component Analysis (PCA), Outliers.

1. INTRODUCTION

Data mining generally deals with non-trivial process of discovering hidden and interesting useful knowledge from various types of data. In data mining, one of the data reduction methods is principal component analysis (PCA) used to dimensionality reduction. Interpreting the gene expression data is the role of statisticians. A rigorous approach to gene expression analysis must involve an up-front characterization of the structure of the data. In addition to a broader utility in analysis methods, singular value decomposition and principal component analysis can be valuable tools in obtaining such a characterization, Andreas (2003).

Over the recent years data mining has been establishing itself as one of the major disciplines in Computer Science and Statistics with growing industrial impact. Data mining is finding interesting structure in databases, Fayyad (1996). Data Mining is the interface of Computer Science and Statistics, utilizing advantages in both disciplines to make progress in extracting information from large data base and it is an emerging field that has attracted much attention in a very short period of time.

Biological data set is a data or measurements collected from biological sources, which is stored or exchanged in a digital form. It is regularly stored in files or databases. Deoxyribonucleic acid (DNA) sequences are basic carriers of life, Johanna Hardin (2005). The DNA code is made up of four chemical bases: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), the combination of base, sugar and phosphate is called a nucleotide, which is arranged in two long strands that form a twisted spiral famously known as the double helix, Durbin R (1998).

Principal Component Analysis exploits the redundancy in multivariate data, enabling us to pick out patterns in the variables; and reduce the dimensionality of a data set without a significant loss of information. It involves a mathematical procedure that transforms a set of correlated response variables into a set of independent variables called the principal components. It help find out one sample is different from another, which variables contribute most to this difference, and whether the variables contribute in the same way or independent of one another. It can also detect sample patterns and quantify the amount of useful information.

2. Proposed Procedures

2.1. Principal Component Analysis

In biological sequence analysis many DNA and RNA sequences discovered in laboratory experiments are not properly identified. Several investigators have successfully used principal component analysis (PCA) in interpreting Biological data. It is widely used statistical technique for unsupervised dimension reduction. It is a traditional multivariate statistical method commonly used to reduce the number of predictive variables and solve the multi-co linearity problem, Bair et al. (2006). Here the focus is on using Principal Component Analysis, Singular Value Decomposition and Cluster Analysis to provide a structure to the data.

The data for the present study were collected from the Gen Bank. The data has twenty factors (proteins), say, STARD13, CD276, AGER, SIRT6, DOK2, TP53AIP1, CDKN3, Hdgf, DYNLRB2, FNTB, Gh, PSMG2, RCOR3, Itih4, IL22RA1, CCNG1, Dedd, RAB4B, REG3A and MAT2B. It can be inferred from the analysis that increases or decrease in protein level, hormone level contributes to Liver cancer.

The Principal Component Analysis (PCA) is performed to extract a set of independent variables. The extracted variables viz., Principal components (PCs) is the linear combinations of the study variables. It involves a mathematical procedure that transforms a set of correlated response variables into a set of independent variables called principal components. Suppose that the vector $(x_1, x_2, ..., x_p)$ of p random variable has a distribution with mean vector μ and covariance $\sum$. The elements of μ and $\sum$ are assumed to be finite. The rank of $\sum$ is $r \leq p$; and the q largest roots $\lambda_1, \lambda_2, ..., \lambda_q$ of $\sum$ are all distinct. The problem of extract in principal component is related to finding the characteristic roots and characteristic vectors for the covariance matrix $\sum$. The elements in the characteristics vectors are the coefficients of the study variables in the PCs. In addition to these, the contribution of each PC in expanding the total variation in the data is also given in terms of percentages. With reference to these, one can compute cumulative % contribution of the PCs in explaining the total variation in the data, which is done according to the decreasing order of the characteristic roots.

2.2 Outlier Detection Approaches

An outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism, Hawkins (1980). Manoj and Kannan (2013) are identifying outliers in univariate case and compared with various formal outlier detection methods. Rousseeuw (1990), Hadi (1992) and Gao (2005), proposed various distance measures to compare with Mahalanobis distance. In this section outlier detection using distance measure and also compared with performance of outlier identification.

2.2.1 Mahalanobis Distance

The standardized Mahalanobis distance depends on estimates of the mean, standard deviation and correlation for the data. A classical Approach for outlier detection is to compute the Mahalanobis distance (MD_i) for each observation x_i is,

$$MD_i = \sqrt{(\mathrm{x}_i - \mu)^T \Sigma^{-1} (\mathrm{x}_i - \mu)} \qquad i = 1, 2, ..., n$$

Where μ and Σ are mean vector, covariance matrix respectively.

2.2.2 Jackknife Distances

Jackknife method was introduced by Quenouille (1949) to estimate the bias of an estimator. The Jackknife distance for each observation is calculated with estimates of the mean, standard deviation, and correlation matrix that do not include the observation itself. The jack-knifed distances are useful when there is an outlier.

2.2.3 T^2 Statistic

T^2 was proposed by Harold Hotelling (1951) and it is given as

$$T_i^2 = \left(X_i - \bar{X}\right)^T S^{-1} \left(X_i - \bar{X}\right) .... (1).$$

From equation (1) derived the upper control limit (UCL) of T^2 statistic as,

$$UCL = ((n-1)^2 / n)\beta(\alpha, p/2, (n-p-1/2),$$

where n is the sample size, p is the number of variables, α is the level of significance.

3. Numerical Illustration

The principal component analysis is performed to extract a set of independent variables (principal components). The PCA is performed individually to the data individually pertaining to the DNA sequences. The collected data is analyzed in view of the objective using appropriate statistical techniques. The analysis carried out using SPSS 16.0. Former to mining of the factors, several tests should be used to judge the appropriateness of the respondent data for factor analysis. These tests include Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Barlett's Test of Sphericity.

Table 3.1 KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.513
Bartlett's Test of Sphericity	Approx. Chi-Square	229.739
	df	190
	Sig.	.026

From table 3.1, he KMO value is 0.503 then it is inferred that patterns of correlations are relatively dense and factor analysis should yield distinct and reliable factors and Barlett's test is significant, i.e. p <0.05, so the factor analysis is appropriate. The result of the analyses for DNA sequences is presented in table 3.2.

Table 3.2 Total Variance Explained

Component Number	Eigen value	Percent of Variance	Cumulative Percentage	Component Number	Eigen value	Percent of Variance	Cumulative Percentage
1	1.42656	7.133	7.133	11	0.992431	4.962	63.808
2	1.29474	6.474	13.607	12	0.941323	4.707	68.514
3	1.25576	6.279	19.885	13	0.909651	4.548	73.063
4	1.21912	6.096	25.981	14	0.877729	4.389	77.451
5	1.15565	5.778	31.759	15	0.862468	4.312	81.764
6	1.14778	5.739	37.498	16	0.826569	4.133	85.896
7	1.11166	5.558	43.056	17	0.781976	3.910	89.806
8	1.07919	5.396	48.452	18	0.731688	3.658	93.465
9	1.05107	5.255	53.708	19	0.711006	3.555	97.020
10	1.02757	5.138	58.846	20	0.596055	2.980	100.000

Table 3.2 is to obtain a small number of linear combinations of the 20 variables which account for most of the variability in the data. In this case, 10 components have been extracted, since 10 components had eigenvalues greater than or equal to 1.0. Together they account for 58.8455% of the variability in the original data. The scree plot is a useful visual aid for determining an appropriate number of principal components. The scree plot graph draws the eigen value against the component number. The component number is taken to be the point at which the remaining eigen values are relatively small and all about the same size. The coefficient of the twenty variables in the selected principal components is given in table 3.3 and shown in the fig. 3.1.

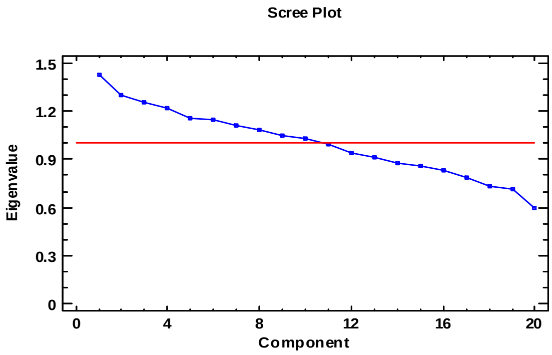

Fig. 3.1: Scree plot indicating that the data have ten factors

Table 3.3 Component Matrix

	Component Number									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
STARD13	-0.15622	0.04874	0.173369	0.04541	0.12400	0.34524	-0.4075	0.332111	-0.10074	0.204927
CD276	0.367658	-0.1623	0.142152	0.04798	0.37691	0.13113	-0.3275	-0.27797	-0.09558	0.133943
AGER	-0.31741	0.0372	-0.15555	-0.3475	-0.0870	-0.3994	-0.0918	-0.1619	-0.10895	0.0296184
SIRT6	-0.02768	-0.3295	-0.44270	0.24186	0.1552	-0.0854	-0.0830	-0.00821	0.15506	-0.221518
DOK2	-0.41337	0.1762	-0.05044	0.3079	0.0014	-0.1001	-0.3155	0.310439	0.07513	0.19967
TP53AIP1	-0.09724	0.1820	0.358774	0.0220	0.3075	-0.2164	0.26862	0.176195	0.118503	0.369262
CDKN3	0.08951	-0.2236	0.00672	0.07882	0.32153	-0.402	0.0956	0.378761	-0.11642	-0.158221
Hdgf	0.006690	-0.2116	0.204839	-0.1410	-0.3282	-0.5036	-0.2669	0.001745	-0.13014	0.0267336
DYNLRB2	0.267133	-0.2638	-0.21177	-0.0952	-0.1743	0.1174	0.1953	0.355637	-0.09266	0.19813
FNTB	0.12787	0.4700	-0.00574	0.0536	-0.1848	-0.0853	-0.2263	0.07970	0.376652	-0.270646
Gh	0.05042	0.1044	0.01028	0.4032	-0.4831	0.05484	0.1832	0.107073	-0.19606	0.0447576
PSMG2	0.258124	-0.0029	-0.31248	-0.1238	-0.0674	-0.0631	-0.1741	0.419623	-0.16272	0.227461
RCOR3	0.315643	0.2258	-0.10829	0.1923	-0.2075	-0.0858	-0.1828	-0.25660	-0.28324	0.192799
Itih4	0.407526	0.1150	0.112528	0.1363	0.1373	-0.3437	-0.2451	0.00932	0.170127	-0.020692
IL22RA1	0.259777	0.0934	0.430967	-0.0272	-0.0720	-0.0222	0.2719	0.162394	-0.11999	-0.155947
CCNG1	-0.07145	0.21538	-0.18845	0.31917	0.26759	-0.0906	0.1052	-0.20647	-0.43227	0.144743
Dedd	0.08522	0.0336	-0.18110	-0.1004	-0.0446	-0.1299	0.2189	-0.17867	0.385151	0.626942
RAB4B	-0.05255	-0.2545	0.07011	0.5735	-0.0699	-0.1506	0.15418	-0.05955	0.220878	-0.004097
REG3A	-0.19970	-0.3398	0.322645	0.0641	-0.1673	-0.0277	-0.1264	-0.15165	-0.22538	0.14011
MAT2B	0.07362	-0.3162	0.148074	0.0609	-0.1704	0.14573	-0.1988	-0.03850	0.347694	0.16225

From the previous table, the first principal component increase with increasing CD276 protein, the second principal component increase with only one of the protein IL22RA1, the third principal component is correlated with two of the original variables increase with increasing RAB4B and Gh proteins, the fourth component increase with only one of the protein Itih4, the fifth component increase with increasing PSMG2 and DYNLRB2 proteins, the sixth component increase with increasing DOK2 and STARD13 proteins, the seventh component increase with increasing CCNG1 protein, the eighth component increase with increasing Hdfg and AGER proteins, the ninth component increase with increasing CDKN protein, the tenth principal component increase with increasing Dedd protein.

The factor matrix shows you the factor loadings earlier to rotation are the following table 3.4. The factor loadings show that out factors are moderately attractive with at least one variable per factors that are above 0.2. It is inferred from the study that these characteristics of factor 1, 2…, 10 can be analyzed by considering one individual biological sequence included in the combination instead of analyzing each biological sequence separately. Also it is inferred from the above that conducting analysis of one single factor from this cluster, we can predict the characteristics of other factors.

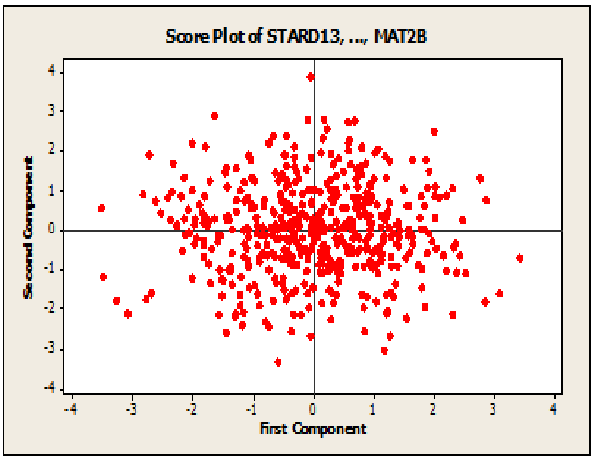

Fig 3.2(a) Score plot

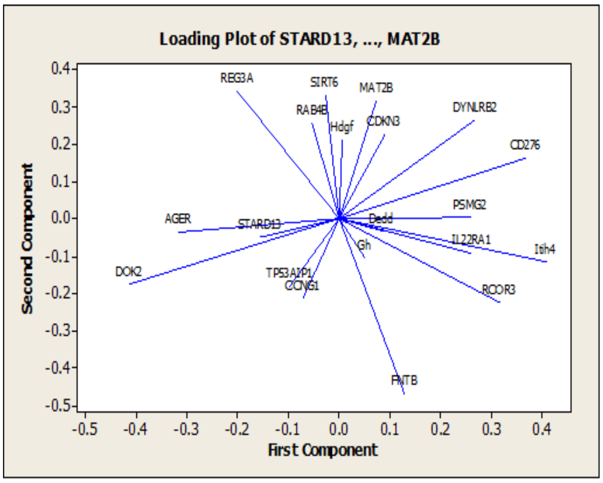

Fig. 3.2(b) Loading plot

The score plot is a projection of data onto subspace. It is used for interpreting relations among observations. It is useful to display the values of 2 or 3 selected factors for each of the *n* cases, after rotation and examine any points far away from the others. Its shown fig. 3.2 (a) and Loading Plot is a relationship between original variables and subspace dimensions. It is used for interpreting relationships among variables. It is shown in fig.3.2 (b).

The Loading Plot fig. 3.2. (b) reveals the relationships between variables in the space of the first two components. The loading plot, shows that proteins CD276 and DYNLRB2 have similar heavy loadings for principal component 1. The variables are far away to each other and they correlate positively. Proteins REG3A and AGER, however, have similar heavy loadings for principal component 2. They are far away to each other and AGER is near to center and REG3A is far away the center and they correlate positively.

4. Outlier plots

Mahalanobis Distance

The Mahalanobis Outlier Distance plot shows the Mahalanobis distance of each point from the multivariate mean (centroid). The distance is plotted for each observation number. Extreme multivariate outliers can be identified by highlighting the points with the largest distance values.

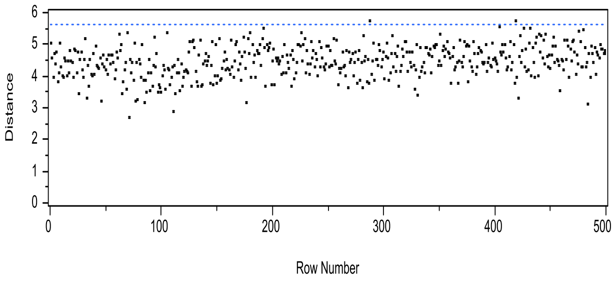

Fig 4 (a) Mahalanobis distance

Jackknife Distance

The Jackknife Distances plot shows distances that are calculated using a jackknife technique.

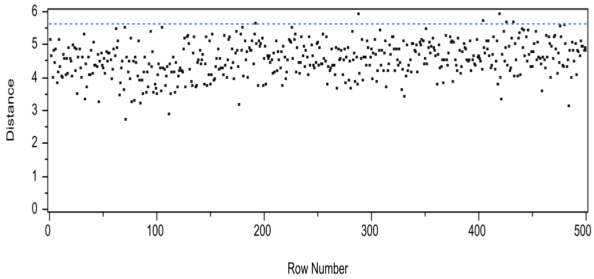

Fig 4 (b) Jackknife distance

T^2- Statistic

The T^2 plot shows distances that are the square of the Mahalanobis distance. This plot is preferred for multivariate control charts. The plot includes the value of the calculated T^2 statistic, as well as its upper control limit. Values that fall outside this limit might be outliers.

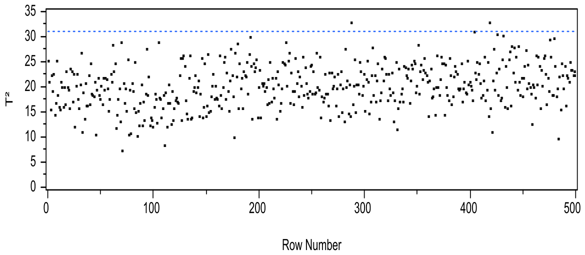

Fig 4 (c) Hotelling's T^2 distance

From the above figures, Mahalanobis distance, Jackknife and T^2- statistic are ploted seperately. The plots shows some of the outlier values fall outside for the control limits. Detecting the outlier points are tabulated below:

Table 3.4 Number of Outliers detected from Distances

Distances Measures	Number of Outliers Detected
Mahalanobis distance	02
Jackknife distance	07
Hotelling's T^2 distance	03

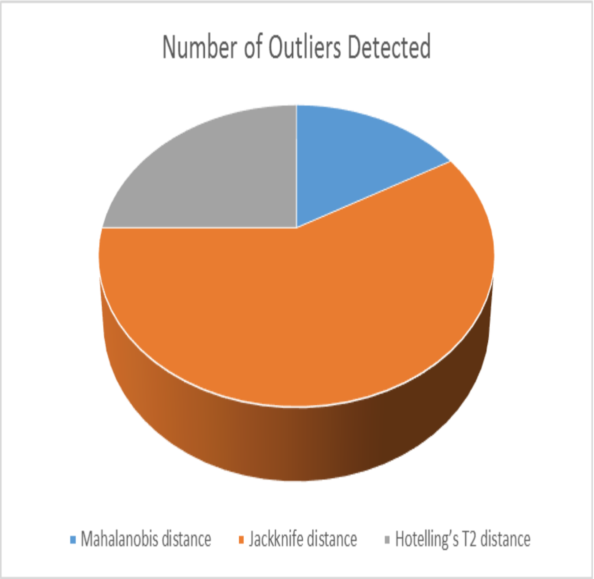

Fig. 5 Number of Outliers detected from distances

5. CONCLUSION

Factor analysis is a composite multivariate statistical approach involving many linear and sequential steps. It is commonly used to reduce variables into a smaller set to save time and easier interpretations. The interpretation of factor analysis based on rotated factor loadings, rotated Eigen values and scree plot. In authenticity, researchers often used more than one extraction and rotation technique based on pragmatic reasoning rather than theoretical reasoning.

Data were subjected to factor analysis using principal component analysis. The KMO value is greater than 0.5 then the data were sufficient. The Barlett's test of sphericity is 229.739 and the p value is less than 0.05, showed that there were decorative relationships between them. Using the Eigen value cut-off of 1.0, there were 10 factors that explain a cumulative variance of 59%. The scree plot confirmed the findings of retaining 10 factors. Finally, we have to identify the outliers using three distance measures in the data set. The Mahalanobis distance and T^2 statistic are identified minimum number of same outlier vallues. Jackknife distance is identifying maximum number of outliers. After we conclude that the jackknife method gives is better result for finding outlier detection.

6. REFERENCES

1. Alfred Ultsch et.al. (2004), Knowledge Discovery in DNA Microarray data of cancer patients with emergent self-organizing maps, ESANN' 2004 proceedings, pg. 501-506.
2. Andrias et al (2003), Data mining, Pearson education (6th ed.), New Delhi, India.
3. An Gie et.al. (2013), A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis, Tutorials in Quantitative Methods for Psychology, 9(2): 79-94.
4. Brett Williams et.al. (2010), Exploratory factor analysis: A five-step guide for novices, Journal of Emergency Primary Health Care (JEPHC), 8(3): 01-13.
5. Buldyrev et al. (1998), Analysis of DNA sequences using methods of statistical physics, Physica A, 249: 430–438.
6. Durbin, R., Eddy, S., Krogh, A., et al. (1998), Biological Sequence Analysis, Cambridge University Press, Cambridge, UK.
7. Fayad, U.M et al. (1996): Advances in knowledge discovery and data mining, AAAI/MIT press.
8. Gulumbe et.al. (2012), Analysis of crime data using principal component analysis: A case study of Katsina State, CBN Journal of Applied Statistics, 3(2): 39-49.
9. Hardin. J. (2005), Microarray data from a statistician's point of view in STATS 42, winter 2005.
10. Hawkins, D. M (1980), Identification of Outliers, Chapman and Hall, New York.
11. Jin Woo Kim et al. (2003), Gene expression profiling of preneoplastic liver disease and liver cancer: a new era for improved early detection and treatment of these deadly diseases? Carcinogenesis, 24(3): 363–369.
12. Manoj K, Senthamarai Kannan K (2013). Comparison of Methods for detecting Outliers, International Journal of Scientific and Engineering Research, 4:9, 709-714.
13. Samuel Karlin et al. (1992), "Chance and Statistical Significance in Protein and DNA sequcence analysis", Science, 257.
14. Wall Michael E. et.al (2003), Singular value decomposition and Principal Component Analysis, in a practical approach to Microarray data analysis, pp. 91-109.

15. Warren. J. Ewens et. al (2004), Statistical methods in Bioinformatics, Springer Publication, New Delhi.
16. Ying Guo et al (2008), A new method to analyze the similarity of the DNA sequences, Journal of Molecular structure: THEOCHEM, 853: 62 – 67.
17. Zakaria Suliman Zubi, Marim Aboajela Emsaed (2010), Using sequence DNA chips data to Mining and Diagnosing Cancer Patients, International Journal of Computers, 4(4): 201-214.