



DESIGN AND IMPLEMENTATION OF AN ALGORITHM FOR PCA BASED FACTOR ANALYSIS AND CLUSTERING APPLIED TO DERMATOLOGY DATA SET

Prof. Dr. P.K. Srimani¹, Manjula Sanjay Koti²

¹Director, R & D, ²Asst. Prof, Dept. of MCA, Research Scholar,

¹Bangalore University, ²Dayananda Sagar College of Engineering, Bharathiar University

^{1,2}Bangalore, ²Coimbatore.

India

Abstract: Medical data mining has immense potential for exploring the hidden patterns in the medical data and the patterns could be utilized for clinical diagnosis. Medical databases are of high dimension and the increase in the dimensionality of data gives rise to a number of new computational challenges not only due to the record size but also due to the increase in the number of attributes. This paper makes an attempt to analyze the dermatology data set by using PCA-based factor analysis (PBF) and clustering analysis. An algorithm for the PBF is proposed and implemented and the predicted results would help the physician to make right predictions at right time so that the patients could be comfortable.

Keywords: Clustering; Communalities; Factor analysis; Loadings; PBF; PCA

I. Introduction

Data mining has become a fundamental methodology for computing applications in medical informatics. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain and these patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. Relationships and patterns within this data could provide new medical knowledge. Progress in data mining applications and its implications are manifested in the areas of information management in healthcare organizations, health informatics, epidemiology, patient care and monitoring systems, assistive technology, large-scale image analysis to information extraction and automatic identification of unknown classes.

"Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors". Both factor analysis and cluster analysis are data reduction techniques and unsupervised learning methods in which the outcome is not determined by feedback from some target. The difference between them are that Factor analysis reduces the number of variables describing a model, by projecting the n-dimensional cloud on a k-dimensional subspace, taking the k axes as factors, while the Cluster analysis reduces the number of statistical units, by grouping them into classes, such that the variance within classes is minimal and the variance between classes is maximal [1]. Cluster analysis groups individuals or objects into clusters so that objects in the same cluster are homogeneous and there is heterogeneity across clusters [2].

Since similarity is fundamental to the definition of a cluster, a measure of the similarity between two patterns drawn from the same feature space is essential to most clustering procedures. Because of the variety of feature types and scales, the distance measure (or measures) must be chosen carefully. It is most common to calculate the *dissimilarity* between two patterns using a distance measure defined on the feature space. We will focus on the well-known distance measures used for patterns whose features are all continuous. The most popular metric for continuous features is the *Euclidean distance* is

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^d (x_{i,k} - x_{j,k})^2 \right)^{1/2}$$
$$= \|\mathbf{x}_i - \mathbf{x}_j\|_2,$$

The Exploratory factor analysis (EFA) enables us to provide a linear model to a set of variables by latent factors. This requires three additional model assumptions viz., (i) The factors have zero mean, unit variance and are

uncorrelated, (ii) The error terms have zero mean and are uncorrelated, and (iii) The factors and error terms are uncorrelated [3].

Factor analysis is related to principal component analysis but not identical. Principal component analysis performs a variance-maximizing rotation of the variable space and it takes into account all variability in the variables. On the other hand, factor analysis estimates how much of the variability is due to common factors (so called “communality”). The two methods become essentially equivalent if the error terms in the factor analysis model (the variability not explained by common factors, communality) can be assumed to have the same variance [1].

The following are the steps involved in the PC based Factor Analysis:

Step 1: PCA produces weights for computing the principal components PC from the z 's

Step 2: Factor analysis based on PC's uses these weights, PC scores to produce factor loadings L and factor scores FS to estimate factors, but only the first k are used

$$Z = L(1) \cdot FS(1) + L(2) \cdot FS(2) + \dots + L(k) \cdot FS(k) + u$$

Step 3: loadings are combined as entries in a matrix are called the factor (pattern) matrix such that one row is reserved for each standardized item z (each containing loadings on all k factors for that standardized item) and one column is reserved for each factor F (each containing loadings for all z 's on that factor).

Step 4: One row is reserved for each standardized item z such that each contains loadings on all k factors for that standardized item

Step 5: one column is reserved for each factor F such that each contains loadings for all z 's on that factor

Factor Analysis Properties

1. The loading L of z on F is the correlation between z and F
2. The square of the loading L is the portion of the variance of z explained by F
3. The sum of the square loadings over all factors is the portion of the variance of z explained by all the factors and this sum equals the communality of z
4. The sum of the squared loadings over all z is the portion of the total variance explained by F and this sum equals the eigen value EV for F
5. The correlation between any 2 z 's is the sum of the products of their loadings on each of the factors.

A key feature in most clustering algorithms is that it treats the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined over the entire clustering solution [8].

II. Related Work

Some of the recent works in this direction include [5][6]. In [7], the authors have made a focused study on outlier detection techniques developed in machine learning and statistical domains. In [8], the authors have compared three data mining methods: linear statistical methods, neural network method, and non-linear multivariate methods. An overview of application of data mining techniques such as decision tree, neural network, association rules, factor analysis and etc., in stock markets is presented in [9]. In [10], the authors have compared ridge regression, pair wise correlation method, forward selection, best subset selection, on quantitative structure retention relationship study based on multiple linear regression on predicting the retention indices for aliphatic alcohols. In [11], the authors have discussed about the outlier mining in medical databases by using statistical methods. In [12], the authors have compared least square regression, ridge and partial least square in the context of the varying calibration data size using only squared prediction errors as the only model comparison criteria. In [13], the authors utilized cluster analysis composed of 2 methods. In [14], the authors have done clustering and after letting the algorithm create its own clusters, added an additional step.

The present work concentrates on clusters and PCA based factor analysis. The structure of the paper is as follows: Section I deals with an introduction to the subject. In section II an overview of the literature pertaining to the work is presented. Section III deals with data set description and Section IV deals with methodology. Section V deals with experiments and results and finally, the conclusions are presented in the last section.

III. Data Set description

The data set used in this study contains the records of 366 patients. This database contains 34 attributes, 33 of which are linear valued and one of them is nominal [15]. The differential diagnosis of erythema-squamous disease is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Patients are first evaluated clinically with 12 features. Afterwards, skin samples are taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope. In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and

histopathological) is given a degree in the range of 0 to 3. Here, 0 indicates that the feature is not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

The following are some of the important observations to be borne in mind while using the data in factor analysis since the very purpose of using factor analysis is to identifying the dimensions underlying the response (outcome) variables y : Observed values for the variables y are called manifest variables since they are available (ii) standardized variables z for the y are typically used (iii) the correlation matrix R for the z is modelled (iv) dimensions corresponding to variables F are called factors (v) observed values for the variables F are called latent variables since they are not available and (vi) most types of manifest variables can be used provided they have more than a few distinct values and an approximate bell-shaped distribution.

Factor analysis is used in many different application areas in particular in health sciences. It is usually applied to survey instrument data.

IV. Methodology

In this study, a detailed analysis pertaining to PCA based factor analysis (PBF) and clustering analysis is presented. The proposed algorithm for PBF is presented here along with the code. Further, the clustering analysis is performed by using the k-means clustering algorithm.

Input : Dermatology dataset,

Output : Rotation sum of squared loadings, Scatter-plot :Components vs eigen values.

function (mat, factors=NULL, correlation=TRUE)

```
{
  princ ← princomp ( covmat = mat , correlation = correlation )
  eigen ← princ$sdev^2
  if (is.null(factors)) factors ← sum ( eigen >= 1 )
  loadings ← princ$loadings [ , 1:factors ]
  coefficients ← loadings [ , 1:factors ] %*% diag ( princ$sdev[1:factors] )
  rotatedmat ← varimax ( coefficients ) $ loadings
  fct.s1 ← apply( rotated, 2 , function (x) sum (x^2) )
  pct.s1 ← fct.s1 / sum (eigen)
  cum.s1 ← cumsum ( pct.ss )
  s1 ← t ( cbind ( fct.s1 , pct.s1 , cum.s1 ) )
  return ( coefficients , rotated , s1 )
}
```

V. Experiments and Results

The algorithm is implemented on the dermatology data set and experimental results are tabulated.

Table I Component Matrix								Table II Rotated Component Matrix							
Component								Component							
Attributes	1	2	3	4	5	6	7	Attributes	1	2	3	4	5	6	7
Erythema	.09	-	-	.23	-	.20	-	Erythema	.10	-.04	-.04	.10	.82	.17	-.20
Scaling	.20	-	-	-	-	-	.21	Scaling	.13	.05	.03	.72	.06	-	-.06
Definite borders	.38	-	-	.45	.25	.20	.10	Definite borders	.30	-.02	.03	.67	-.08	.02	-.08
Itching	.26	-	-	.06	-	.02	.12	Itching	.51	.48	-.06	.22	.10	.04	-.19
Koebner	-	-	-	.10	.33	-	-	Koebner	-.21	.39	-.21	-.03	-.15	.14	.53
Polygonal	.37	.28	.24	.10	.33	.31	.28	Polygonal	.10	.49	-.18	-.04	.05	-	.29
Follicular	.18	.51	.01	.01	.33	.35	.25	Follicular	-.18	.93	-.06	.02	.03	.02	.01
Scalp	.58	.73	.10	.14	.02	-.06	.07	Scalp	.78	-.13	.31	.09	-.03	-	-.08
Family	.76	-	-	.17	.04	.03	.09	Family	.80	-.13	.00	.19	-	-	.10
Melanin	.78	-	-	.04	.08	.00	-.09	Melanin	.36	-.10	.38	.18	.07	-.05	.16
Eosinophils	.38	-.01	-.38	.11	.06	-	.20	Eosinophils	-.18	.94	-.07	.03	.03	.02	-.03
	-.59	.73	.10		.00	.07	.10								

PNL	-	.04	.07	-	-	-	-	PNL	-.33	-.09	-.16	.52	.02	.17	.29
Fibrosis	.50	.02	.01	-	.17	-	.23	Fibrosis	.31	-.31	-.12	.39	-.52	-	-.18
Exocytosis	-	.40	.51	.52	-	-	.04	Exocytosis	-.20	-.23	-.14	-.57	.05	.56	.21
Acanthosis	-	.73	.01	.20	.29	.07	-.18	Acanthosis	-.71	.29	.05	.21	.11	-	.01
Hyperkeratosis	.14	-	.25	.12	.34	-	-.01	Hyperkeratosis	.21	.19	-.02	.11	.16	.75	.04
Parakeratosis	.27	.06	-	.03	.27	.42	-	Parakeratosis	.28	-.12	.18	-.10	-.06	-	.61
Clubbing	.53	-	.35	.15	.02	.12	-.06	Clubbing	.59	.11	.10	.24	-.22	.17	.05
Elongation	.89	-	.30	.05	.08	.02	.00	Elongation	.92	-.15	-.10	.11	.00	.01	-.01
Thinning	.76	-	.03	.35	.31	-	.04	Thinning	.72	-.31	-.19	-.01	.42	.42	.08
Spongiform	.86	-	.30	.06	.06	.03	-.04	Spongiform	.89	-.15	-.11	.12	.01	-	.03
Munro	.58	-	.17	.03	-	.14	-.35	Munro	.52	-.12	-.10	.19	-.44	.07	-.23
Focal	.67	-	.26	.08	.12	.22	-.29	Focal	.74	-.10	-.11	.03	.07	-	.31
Vacuolization	.57	-	.30	.03	.06	-	-.01	Vacuolization	.64	-.04	-.15	.08	.33	-	-.22
Spongiosis	-	.59	.72	.09	.14	.04	-.11	Spongiosis	-.18	.94	-.06	-.01	.01	.02	-.01
Perifollicular	-	.02	.32	-	.32	.00	-.02	Perifollicular	-.05	-.05	.92	-.01	-.09	.01	-.03
Inflammatory	-	.01	.34	-	.82	.30	-.02	Inflammatory	-.06	-.07	.93	-.01	.07	-	-.04
Age	-	.60	.71	.09	.14	.07	-.09	Age	-.19	.93	-.07	-.03	.01	-	.01
Class	.04	-	.29	.46	-	.11	-.05	Class	.11	.09	-.56	.02	.15	.01	-.05

From Table I, it is evident that the matrix of loadings consists of 29 rows (1 for each item *z*) and 7 columns (1 for each factor *F*). The loadings are distributed as follows: 50% on factor 1 and 25% on factor 4.

From Table II, it is clear that the highest load bearing factors are different for different attributes and is distributed in a random way.

Table III Communalities

Attributes	Initial	Extraction	Attributes	Initial	Extraction
Erythema	1.00	.76	Hyperkeratosis	1.00	.67
Scaling	1.00	.55	Parakeratosis	1.00	.50
Definite borders	1.00	.55	Clubbing	1.00	.51
Itching	1.00	.59	Elongation	1.00	.89
Koebner	1.00	.57	Thinning	1.00	.85
Polygonal	1.00	.58	Spongiform	1.00	.85
Follicular	1.00	.91	Munro	1.00	.59
Scalp	1.00	.74	Focal	1.00	.71
Family	1.00	.70	Vacuolization	1.00	.61
Melanin	1.00	.35	Spongiosis	1.00	.92
Eosinophils	1.00	.92	Perifollicular	1.00	.86
PNL	1.00	.53	Inflammatory	1.00	.88
Fibrosis	1.00	.66	Age	1.00	.90
Exocytosis	1.00	.79	Class	1.00	.36
Acanthosis	1.00	.70			

From Table III, it is found that the Communalities started out as all 1's since the PC method was used to extract factors. The new values are less than 1 as they should be when the number of factors is less than the number of items.

Table IV Total Variance Explained

Components	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	7.59	26.19	26.19	7.59	26.19	26.19	6.57	22.64	22.64
2	4.12	14.21	40.40	4.12	14.21	40.40	4.65	16.03	38.67
3	2.54	8.77	49.17	2.54	8.77	49.17	2.57	8.85	47.52
4	1.93	6.66	55.83	1.93	6.66	55.83	2.06	7.11	54.63
5	1.50	5.17	61.00	1.50	5.17	61.00	1.43	4.93	59.56
6	1.30	4.48	65.48	1.30	4.48	65.48	1.49	5.14	64.70
7	1.02	3.50	68.98	1.02	3.50	68.98	1.24	4.28	68.98
8	.92	3.16	72.15	-	-	-	-	-	-
9	.87	3.00	75.15	-	-	-	-	-	-
10	.80	2.76	77.90	-	-	-	-	-	-
11	.74	2.54	80.44	-	-	-	-	-	-
12	.66	2.27	82.71	-	-	-	-	-	-
13	.63	2.17	84.88	-	-	-	-	-	-
14	.59	2.02	86.90	-	-	-	-	-	-
15	.55	1.89	88.79	-	-	-	-	-	-
16	.50	1.71	90.50	-	-	-	-	-	-
17	.45	1.56	92.06	-	-	-	-	-	-
18	.43	1.50	93.56	-	-	-	-	-	-
19	.32	1.09	94.64	-	-	-	-	-	-
20	.30	1.02	95.67	-	-	-	-	-	-
21	.27	.93	96.60	-	-	-	-	-	-
22	.25	.87	97.47	-	-	-	-	-	-
23	.25	.86	98.33	-	-	-	-	-	-
24	.10	.36	98.69	-	-	-	-	-	-
25	.10	.34	99.02	-	-	-	-	-	-
26	.09	.33	99.35	-	-	-	-	-	-
27	.07	.26	99.61	-	-	-	-	-	-
28	.06	.22	99.83	-	-	-	-	-	-
29	.05	.17	100.00	-	-	-	-	-	-

Each factor F has an associated eigen value EV which is called a characteristic root since it is a solution to the characteristic equation for the correlation matrix R. The sum of the eigen values over all factors equals the total variance. Hence, an eigen value measures how much of the total variance of the z's is accounted for by its associated factor (or PC). In fact, factors with larger eigen values contribute more towards explaining the total variance of the z's. The eigen values at the start have the more important factors (or PC's) and are generated in decreasing order – $EV(1) \geq EV(2) \geq EV(3) \geq \dots$.

From Table IV, it is found that EV-one extracts 7 factors and these factors explain about 69% of the total variance.

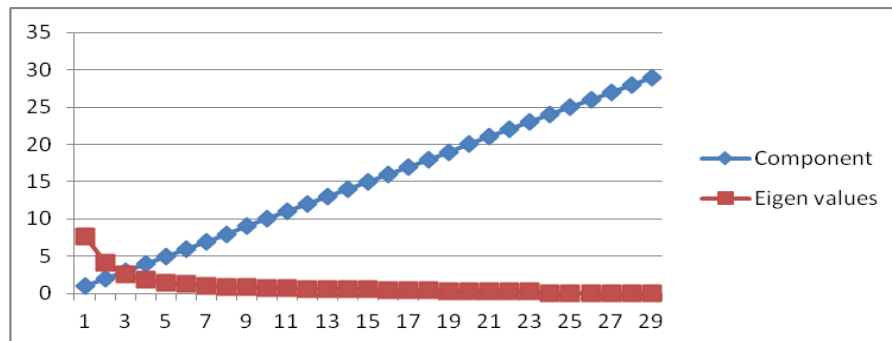


Figure 1 Scree Plot

Here by scree plot(Fig.1)we mean that it is the debris at the bottom of a cliff (i.e., a biggest change in the slope). Here it is between 4 & 5.

Table V Results of Clustering Analysis

Cluster No.	Instances
0	72 (20%)

1	111 (30%)
2	183 (50%)

SSE is computed as:

$$SSE = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where C_k is the set of instances in cluster k ; μ_k is the vector mean of cluster k . From Table V, it is found that the Sum of squared errors happens to be 556.43.

VI. Conclusion

In order to facilitate clinicians to make better diagnosis for treatment purposes, data mining tools have been developed. Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved, uncorrelated variables called factors. Principal Components are weighed linear combination of correlated constructs which explain a maximal amount of variance of the variables which are uninterpretable. Factor analysis is related to principal component analysis (PCA), but the two are not identical. Latent variable models, including factor analysis, use regression modeling techniques to test hypotheses producing error terms, while PCA is a descriptive statistical technique. It is observed that factors are underlying constructs with the following properties: (i) they are not directly measured (ii) they influence responses on measured variables, and (iii) they are influenced by the measurement error(unique factors). The difference between them are that Factor analysis reduces the number of variables describing a model, by projecting the n-dimensional cloud on a k-dimensional subspace, taking the k axes as factors, while the Cluster analysis reduces the number of statistical units, by grouping them into classes, such that the variance within classes is minimal and the variance between classes is maximal. This paper concentrates on the performance of data mining technique on dermatology data through PCA based factor and clustering analyses. An algorithm for PCA based factor analysis is proposed and is implemented on the data set. The results are found to be interesting and it is concluded that the results of the present investigation could be considered as an effective diagnostic tool for the dermatitis disease.

Acknowledgments

Mrs. Manjula Sanjay Koti is grateful to Bharathiar University, Tamil Nadu and Dayananda Sagar College of Engineering, Bangalore for providing the facilities to carry out the research work.

References

- [1] Sigbert Klink, Andrija Mihoci, and Wolfgang Härdle, "Exploratory factor analysis in MPLUS, R and SPSS", ICOTS8, 2010.
- [2] M. Matteucci, "A Tutorial on Clustering Algorithms", http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/, 2008
- [3] Klink, S., & Wagner, C., "Visualizing exploratory factor models", *COMPSTAT 2008 - Proceedings in Computational Statistics - 18th Symposium held in Porto (Portugal)* by P. Brito (ed.). Heidelberg: Physika Verlag (CDROM), Software: [http://stirner.wiwi.huberlin.de/mediawiki/mmstat/de/index.php/Spezielle Themen - R-Visualisierung von Faktormodellen](http://stirner.wiwi.huberlin.de/mediawiki/mmstat/de/index.php/Spezielle%20Themen%20-%20R-Visualisierung%20von%20Faktormodellen), 2008.
- [4] Srimani, P.K., Manjula Sanjay Koti, "Application Of Data Mining Techniques For Outlier Mining In Medical Databases", *IJCR*, Vol. 33, Issue, 6, 402-407, 2011.
- [5] Srimani, P. K., Manjula Sanjay Koti, "A comparison of different learning models used in data mining for medical data", *AIP Conf. Proc.* 1414, 51-55 ; doi: 10.1063/1.3669930, 2011.
- [6] Srimani, P. K., Manjula Sanjay Koti, "The impact of rough set approach on medical diagnosis for cost effective Feature selection", *IJCR* Vol. 3, Issue, 12, pp 175-178, 2011.
- [7] Hodge, V. and Austin, J., "A survey of outlier detection methodologies. Artificial Intelligence Review", 22, 2, 85-126, 2004.
- [8] Munoz, Jesus, and Angel M. Felicísimo, "Comparison of Statistical Methods Commonly used in Predictive Modeling," *Journal for Vegetations Science* 15, pp. 285-292, 2004..
- [9] Maitra Saikat and Yan Jun, "Principle Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression" *Casualty Actuarial Society, Discussion Paper Program*, pp. 79-90, 2008.
- [10] Fakas, Orsolya and Heberger Karoly, "Comparision of ridge Regression, PLS, Pairwise Correlation, Forward and Best subset selection methods for prediction of Retention indices for Aliphatic Alcohols," *Journal of Information and Modeling*, 45:2, pp. 339-346, 2005.
- [11] Srimani, P. K., Manjula Sanjay Koti, "Outlier Mining in medical databases using statistical Techniques", *IJEST*, Vol 4, ,pp 239-246, 2011.
- [12] Huang, J. et al., "A Comparison of Calibration Methods Based on Calibration Data size and Robustness," *Journal of Chemometrics and Intelligent Lab. Systems*, 62:1, pp. 25-35, 2002.
- [13] Y. Pen, G. Kou, Y. Shi, and Z. Chen, "Improving Clustering Analysis for Credit Card Accounts Classification," *LNCS* 3516, pp. 548- 553, 2005.
- [14] A. Kalton, K. Wagstaff, and J. Yoo, "Generalized Clustering, Supervised Learning, and Data Assignment," *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, ACM Press, 2001.
- [15] <http://archive.ics.uci.edu/ml/datasets.html>