Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 02, February-2020

Identification of Factors Causing Breast Cancer using Factor Analysis

Dr. Josephine Prem Kumar,
Professor (CSE), Cambridge Institute of Technology, Bangalore.

***Abstract*—Factor analysis is a data reduction technique that is used to reduce a number of observed variables from large data to a lesser number of latent variables. These latent variables are called factors. These factors are hypothetical variables used to explain the correlation between the variables. Factor analysis is found to be a useful tool to relationships among the variables and to classify them. In breast cancer diagnosis, pathologists examine FNA (Fine Needle Aspirate) tissue samples. In these samples, certain characteristics are considered to determine whether the tissue is malignant or benign. So here the Factor analysis is carried out for breast cancer dataset to find out the factors which is more prominent in detecting the breast cancer. Maximum likelihood estimation is used in extracting the factor and Varimax rotation is done to rotate the obtained factors. It is used to simplify the description of a specific sub-space in terms of a few major factors. Finally the factors used in the detection of breast cancer has been reduced to just two – appearance and nature of the cell. Factor analysis has been carried out using R tool.**

Keywords—Unobserved variables, factors, factor analysis, factor structure

I. INTRODUCTION

The diagnosis of a disease is based upon the results of various tests performed upon the patient. When many tests are involved, it is very difficult to obtain the ultimate diagnosis, even for a medical expert. Hence, over the past few decades, many computerized diagnostic tools have been developed to help the physician to make the diagnosis, making sense out of all the available results.

Breast cancer is the second most common cancer occurring among women, the first one being skin cancer. Its incidence increases with age. If the disease can be detected at an early stage, then many precious lives can be saved. Fine needle aspiration (FNA) of breast masses is a cost-effective, non-traumatic, and mostly non-invasive diagnostic test that obtains information that are needed to evaluate malignancy.

When pathologists examine FNA (Fine Needle Aspirate) tissue samples in breast cancer diagnosis, they consider nine characteristics, namely clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. Each of these characteristics is assigned a number from 1 to 10 by the pathologist. The larger the number the greater the likelihood of malignancy [11].

Clump thickness: Cancer cells will be grouped into multiple layers, whereas the normal cells will be in monolayers.

Marginal adhesion: Normal cells usually sticks together but cancer cells loses this ability.

Uniformity of cell size/shape Cancer cells will vary in shape and size.

Single epithelial cell size: In some cases, Epithelial cells which are enlarged may be a infectious cell.

Bare nuclei: In cancer cells nuclei will not be surrounded by cytoplasm.

Bland Chromatin: In cancer cells the chromatin tend to be more harsh.

Normal nucleoli: Usually the nucleolus will be very small in normal cells, but in cancer cells it becomes more protuberant.

There are a number of variables that can be used to explain or describe the complex variety and interconnections in real time systems. Among them only a few are basic variables that are central to the system; these are to be determined. In the factor analysis model, the measured variables depend on a small number of unobserved variables or latent factors. These are also known as "common factors" because each factor may affect many of these variables. Factor analysis helps to investigate the possible underlying structure in a set of interrelated variables; and extract these factors. Here the factors that are more prominent in determining the breast cancer are identified.

II. FACTOR ANALYSIS

The broad purpose of factor analysis is to summarize the available data so that relationships and patterns among them can be easily interpreted and understood. It is usually used to regroup the different variables into a limited set of clusters based on shared variance. Hence, it helps to understand constructs and concepts. Researchers can interpret concepts that cannot be measured easily by reducing a large number of variables into a lesser number of factors.

Each factor captures some amount of the overall variance of the measured variables. A new factor is a linear combination of the already existing measured variables.

Large datasets contain several observed variables. These can be reduced by grouping related variables to a set of factors. These factors are few in number. They categorize the variables. It is easy to focus on these few factors than encompass the large number of variables than to consider all the measured variables (which may be insignificant). This is the major goal of factor analysis – reducing the large number of observed variables into a lesser number of significant factors that make it easy to understand and interpret the data. Factor analysis is also useful into grouping similar variables into categories.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 02, February-2020

The main aim of factor analysis is to reveal the few factors that underlie the large number of observed variables. It is used to explore the structure underlying the set of variables without having any preconceived idea about the structure [5]. This involves factor extraction and factor rotation. The data given as input may be raw data or the covariance matrix. If raw data is used, then during the process of factor analysis the covariance matrix is computed automatically.

It is also possible that factor analysis would enable one to test theories involving variables which are hard to measure directly. Finally, at a more basic level, factor analysis can help to combine certain sets of questionnaire items (observed variables) into a single more reliable measure of that factor by establishing the fact that they are all measuring the same underlying factor (perhaps with varying reliability).

III. COMPONENTS OF FACTOR ANALYSIS

Factor Analysis consists of the following steps:

1. Factor Extraction
2. Factor Rotation
3. Interpretation of Factor Loadings
4. Naming the Factors, Writing the Results

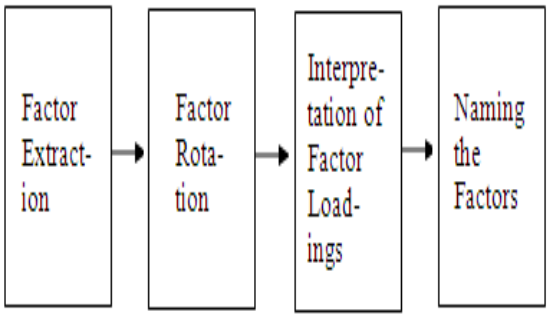

Figure 1. Steps in Factor Analysis

A. Factor Extraction

Factor analysis is based on the 'common factor model'. This is a theoretical model that is useful for studying relationships among variables. It is the general way to estimate more than one factor from the data. This model postulates that observed measures are affected by underlying common factors and unique factors, and the correlation patterns need to be determined. In the factor analysis model, each variable is assumed to depend on a linear combination of factors. These coefficients are called loadings. It also includes a component called specific variance. This is due the variable's independent random variability. It is called specific variance because it specific to each variable. During factor extraction it is the variable's shared variance that is partitioned from its unique or specific variance. The shared variance contributes to the determination of factors [2]. There are a number of extraction methods available. In this paper the Maximum Likelihood Estimation method is used. Maximum Likelihood attempts to analyze the maximum likelihood of sampling the observed correlation matrix [10]. When applied to a data set, maximum-likelihood estimation provides estimates for the model's parameters. The parameters are chosen such that they maximize the likelihood (or probability) of the samples.

Maximum Likelihood Estimation basically uses a mathematical expression called Likelihood function. This function contains the unknown model parameters. The parameter's values are estimated such that it maximizes the sample likelihood.

B. Factor Rotation

The purpose of factor rotation is to obtain a simple and clear structure of the underlying factors. The estimated loadings from a factor analysis model may give a large weight on several factors for some of the measured variables. In this case it's difficult to explain what those factors represent. To get a clearer idea on what these factors mean, they are rotated. Rotation does not change the basic aspects of the analysis. Rotation helps to make the results more understandable by giving a simpler structure [9]. Rotation allows the items to group into factors more distinctly. It reduces the ambiguities that occur in the preliminary analysis. Ultimately, the simple structure attempts to have each factor define a distinct cluster of interrelated variables so that interpretation is easier [4].

Suppose the coordinates of a point in the loadings matrix represent each row of the loadings matrix i.e. each factor corresponds to a coordinate axis. Factor rotation involves rotating these axes. Rotating these axes leads to the computation of new loading in the rotated coordinate system. The actual coordinate system is unchanged.

Factor rotation attempts to give results such that each variable contributes to a lesser number of factors i.e. each variable has large loadings for only a few factors. It is preferred that each variable has large loadings for only a single factor. There are a number of ways to perform factor rotation.

Two main types of factor rotation are used: orthogonal and oblique. Orthogonal rotation methods assume that the factors in the analysis are uncorrelated. [6] lists four different orthogonal methods: equamax, orthomax, quartimax, and varimax. In contrast, oblique rotation methods assume that the factors are correlated. [6] lists 15 different oblique methods - direct oblimin, promax, etc. Here varimax rotation and promax rotation are used to rotate the factors.

Varimax provides a structure such that each variable has a small number of large loadings and large number of smaller loadings. This allows each variable to be associated with just one (or a lesser number of) factors making it easy to interpret the factors. Varimax maximizes the unique variance of each variable. Varimax rotation simplifies the expression of a particular sub-space in terms of just a few major items each. Varimax minimizes the number of variables that have high loadings on each factor and works to make small loadings even smaller.

Promax rotation is fast and simple. It simplifies orthogonal rotation by making small loadings even closer to zero. In effect, promax rotates the factor axes, allowing them to have an oblique angle between them, i.e. it performs a non-rigid rotation of the axes. This makes an interpretation of the rotated factors more precise. Also, if

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 02, February-2020

the factors are in fact uncorrelated, then promax will tell you that. The structure may appear more simple in oblique, but correlation of factors can be difficult to reconcile.

C. Interpretations of Factor Loadings

The factor loadings of the variables are used to determine the factors. They indicate the strength of the relationship between the variables and the factors. The variables that have the highest values of loadings for each factor contribute to that factor. To confirm the identification of factors, it is necessary to examine the zero and low loadings.

In order to determine the reliability of the factor, the relationship between the individual loading and magnitude of the absolute sample size is considered. The larger the sample size, the smaller are the loadings allowed for a factor to be considered significant. The cut-off size depends on the ease of interpretation and how complex variables are being handled.

D. Number of Factors to Retain, Naming the Factors, Writing the Results

When extracting factors neither should too many factors be extracted nor should very few factors be extracted. When many factors are extracted there is an undesirable error variance while extracting very few factors leaves out common variance. This makes it very important to select which criterion is most suitable when deciding on the number of factors to extract. To determine the number of factors to retain, you will need to pick the solution that provides the most desirable rotated factor structure. Factors that have less than three variables, many complex variables and item loadings that are less than 0.32 are generally viewed as undesirable.

Naming of factors is more of an 'art' since there are no rules for naming factors, except to give names that best represent the variables within the factors. Depending on your research questions, you may want to extend your findings. Finally, you could perform reliability testing if you are using factor analysis to validate or construct a questionnaire.

IV. EXPERIMENTAL RESULTS

A. Tool Used

R tool is a very powerful tool to analyze data, that is gaining in popularity due to its costs (its free) and flexibility (its open-source). This is specifically useful for factor analysis work, as one can examine the properties of data, check any model assumptions, conduct exploratory factor analysis, and then follow up with a confirmatory factor analysis.

The packages that are available with R tool are used in the implementation of factor analysis in this paper. The psych package, a part of R tool, has a function, fa(), that does factor extraction using either principal axis/factor, maximum likelihood according to the user's specifications [8].

B. Wisconsin breast cancer diagnosis dataset

The data available at the WDBC web site [3] has been used in the work presented in this paper. This is a processed form of the FNA data. The original dataset was obtained by Wolberg and Mangasarian [11]. This dataset contains real data taken from needle aspirates from human breast tissue. It consists of 699 instances with 9 attributes; viz clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. These are represented as integer numbers in the range [1,10].

Table 1: WBCD data summary

Domain	1	2	3	4	5	6	7	8	9	10
Clump Thickness	139	50	104	79	128	33	23	44	14	69
Uniformity of Cell Size	373	45	52	38	30	25	19	28	6	67
Uniformity of Cell Shape	346	58	53	43	32	29	30	27	7	58
Marginal Adhesion	393	58	58	33	23	21	13	25	4	55
Single Epithelial Cell Size	44	376	71	48	39	40	11	21	2	31
Bare Nuclei	402	30	28	19	30	4	8	21	9	132
Bland Chromatin	150	160	161	39	34	9	71	28	11	20
Normal Nucleoli	432	36	42	18	19	22	16	23	15	60
Mitoses	563	35	33	12	6	3	9	8	0	14

Table 1 shows data summary statistics [7]. Factor Analysis is used to extract two factors from the dataset. The system takes as input, the FNA data and outputs the two factors.

The data set is first converted to a matrix form. This is given as input to the fa() procedure.

Table 2: Factor Loadings for all the variables

	ML1	ML2
V1	0.68	-0.06
V2	0.95	-0.08
V3	0.95	-0.19
V4	0.75	0.18
V5	0.79	0.09
V6	0.47	-0.02
V7	0.80	0.10
V8	0.78	0.12
V9	0.50	0.17

Table 2 gives the factor loadings for all the variables corresponding to the two factors.

To have a better understanding and to better interpret the results, factor rotation is done. Maximum likelihood estimation is used in extracting the factors. Two factors are extracted. Varimax rotation is done to rotate the obtained factors. Table 3 gives the results.

Table 3: Results of Factor Rotation using Varimax rotation

	ML1	ML2
V1	0.55	0.41
V2	0.76	0.57
V3	0.83	0.49
V4	0.45	0.64
V5	0.53	0.60
V6	0.37	0.29
V7	0.53	0.61
V8	0.51	0.61
V9	0.26	0.46

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 02, February-2020

From the above results, it is seen that the variables corresponding to clump thickness, uniformity of cell size, uniformity of cell shape and bare nuclei can be interpreted as a single factor, namely "appearance of the cell" and the variables corresponding marginal adhesion, single epithelial cell size, bland chromatin, normal nucleoli and mitoses to can be interpreted as another factor, namely the "nature of the cell". Thus the factors used in the detection of breast cancer has been reduced to just two – appearance and nature of the cell.

Table 4: Results of Factor Rotation using Promax rotation

	ML2	ML1
V1	0.33	0.39
V2	0.36	0.54
V3	0.30	0.70
V4	0.77	0.02
V5	0.64	0.18
V6	0.25	0.24
V7	0.67	0.17
V8	0.68	0.13
V9	0.59	-0.07

Once again, the interpretation is quite similar i.e. from the results shown in Table 4, it is seen that the variables corresponding to clump thickness, uniformity of cell size, uniformity of cell shape and bare nuclei can be interpreted as a single factor and the variables corresponding marginal adhesion, single epithelial cell size, bland chromatin, normal nucleoli and mitoses to can be interpreted as another factor.

The study is done to compare the two commonly used methods of rotation, Varimax and Promax, in terms of their ability to correctly link items to factors and to identify the presence of simple structure. The results have shown that the two approaches are equally able to recover the underlying factor structure, regardless of the correlations among the factors; however, the Promax method is better able to identify the presence of a "simple structure." The results also show that for identifying which variables are associated with which factors, either approach is effective.

V. CONCLUSION

Factor analysis is used to identify latent constructs or factors. It is commonly used to reduce variables into a smaller set to save time and facilitate easier interpretations. The interpretation of factor analysis is based on rotated factor loadings. In reality, researchers often use more than one extraction and rotation technique based on pragmatic reasoning rather than theoretical reasoning. It has been shown that both the Varimax and Promax rotation are equally effective, but for identifying "simple structure" when it is present, the Promax method is preferable.

Breast cancer can be detected using nine characteristics namely, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. Factor analysis has been done on Wisconsin breast cancer diagnosis dataset using R tool. Two factors have been extracted from these nine characteristics. They are appearance of the cell and nature of the cell. Therefore only these two factors need to be examined for the detection of breast cancer. Reduction in the number of variables saves time and facilitates easier interpretations.

REFERENCES

[1] A. Alexander Beaujean, Factor Analysis using R. Baylor University. Volume 18, Number 4, February 2013.

[2] Anna B. Costello and Jason W. Osborne, "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis", North Carolina State University. Volume 10 Number 7, July 2005.

[3] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, http://www.ics.uci.edu/»mlearn/mlrepository.html. University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[4] Cattell, R.B. (1973). Factor analysis. Westport, CT: Greenwood Press.

[5] Diana Suhr, "Exploratory Factor Analysis with the World Values Survey". Ph.D. University of Northern Colorado, Paper 331-2012.

[6] Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

[7] Jung-Ying Wang, "Data Mining Analysis (breast-cancer data)", Register number: D9115007, May, 2003

[8] Revelle, W. (2012). psych: Procedures for psychological, psychometric, and personality research (Version 1.2.4) [Computer software]. Evanston, IL: Northwestern University.

[9] Rummel R J. Applied factor analysis. L~Evanston, IL: Northwestern University Press, 1970. 617 p.

[10] Tabachnick and Fidell, 'Using Multivariate Statistics (5th ed.)', 2007, pp. 633.

[11] W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, National Academy of Sciences, Washington, DC, 87:9193–9196, 1990.