Principal Component Analysis of Infertility Data

Nazera Khalil Dakhil *Hind Restom Mohammed* *Hayder Abood*
College of Mathematics and Computer Sciences
University of Kufa
E-mail: n_dakhil@hotmail.com

Abstract

This paper applied PCA on infertility set of data, that was collected from Al-Nasiriya province. Infertility of women that have been unable to conceive a child after one year of their marriage without birth control. Since infertility is very common among Iraqi women. Hence our research mainly concern with this disease in order to identify the causes that contribute to this problem.

Introduction

Principal component analysis (PCA) is the best known of the techniques of multivariate data. The basis idea of PCA is reduce the dimensionality of a data set consisting of a many correlated variables, and maintain highest possible of the variation existing in the data set.

Principal components analysis is amongst the oldest and most widely used of multivariate techniques. Originally introduced by Pearson (1901) and independently by Hotelling (1933), the basic idea of the method is two describe the variation of a set of multivariate data in terms of a set of uncorrelated variables each of which is a particular linear combination of original variables. The new variables are derived in decreasing order of importance so that, for example, the first principal component accounts for as much as possible of the variation in the original data. The usual objective of this type of analysis is to see whether the first few components accounts for most of the variation in the original data. It is argued that if they do, then they can be used to summaries the data with littlie information, thus providing a reducing in the dimensionally of the data, which might be useful in simplifying later analyses.

Previous studies examined factors influencing fertility and contraceptive behavior within the

socio-economic and cultural frameworks in order to assist the Government of Pakistan in formulating such population policies as will promote the family planning programme and reproductive health and cause a fertility decline in the Pakistan. Gronbach's alpha and Principal Component Analysis procedure were also used to identify the questions (statements) which measure the same idea [5].

Other studies investigate five year age group ASFRs (Age Specific Fertility Rates 1999-2003),from 88 countries for which the basic data were available for the years considered in the present analysis, that the first Three components accounted for more than 95 percent of the total variance [3].
principal components

Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, X_3, \dots, X_P$.Geometrically, These linear combinations represent The selection of a new coordinate system obtained by rotation the original system with $X_1, X_2, X_3, \dots, X_P$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure. As we shall see, principal components depend solely on the covariance matrix Σ (or the correlation matrix)

of $X_1, X_2, X_3, \dots, X_P$. Their development does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be made from the sample components when the population is multivariate normal.

Let the random vector $X^T = X_1, X_2, X_3, \dots, X_P$ have been covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq, \dots, \geq \lambda_p \geq 0$.

Consider the linear combination

$$Y_1 = a_1^T X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{12}X_P$$

$$Y_2 = a_2^T X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2P}X_P$$

$$\vdots \quad \vdots \qquad (1)$$

$$Y_P = a_P^T X = a_{P1}X_1 + a_{P2}X_2 + \cdots + a_{P2}X_P$$

$$Var(Y_1) = a_i^T \Sigma a_i \qquad i = 1,2,3, \dots, p \qquad (2)$$

$$Cov(Y_1, Y_k) = a_i^T \Sigma a_k \qquad i,k = 1,2,3, \dots, p \qquad (3)$$

principal components from the correlation matrix

It is quite common to calculate the principal components of a set of variables after they have been standardized to have unit variance. This means that one is effectively finding the principal components from the correlation matrix P rather than from the covariance matrix, Σ. The mathematical derivation is the same, so that the components turn out to be the eigenvectors of P. However, it is important to realize that the eigenvectors of P with generally not be the same as these of Σ. Choosing to analyse P rather than Σ Involves a definite but arbitrary decision to make the correlation or the variables 'equally important'.

f or the correlation matrix, diagonal terms are all unity. Thus the some of the diagonal teams (or the some of the variance of the standardized variables)will be equal to p. thus the sum of the eigenvalues of P will also be equal to P so that the proportion of the total variation accounted for by the *j*th component is simple λ_j/P [2].

Results and Discussion

Data were collected from Bent Al-Huda in Al-Nasiriya province, which were available for 779 women. Variables were classified into categories as follow: Type of infertility was categorized to primary, and secondary; duration of infertility and duration of treatment were measured in days; age of wife or husband were measured in years.

The means and standard deviations (SDs) for all variables were presented in Table one. Average age of husband and wife were very similar. While average of duration of infertility was 4.6 years and average of duration of treatment was 1.7 year. The maximum (max)and the minimum (min) of the values of the variables were also given in the table for showing the variations.

Table 1. Descriptive Statistics

Variable (N = 779)	Min.	Max.	Mean	SD	Variance
Type of Infertility	1	2	1.39	.487	.237
Age of Wife	14	52	27.49	6.707	44.978
Number of Children	0	7	.48	.931	.867
Age of Husband	12	70	32.04	8.017	64.277
Duration of Infertility (years)	.1	24.0	4.572	3.5770	12.795
Duration of Treatment (years)	.08	15.00	1.7421	2.25269	5.075

Table 2. Correlation Matrix for all variables

Variable	Type of Infertility	Age of Wife	Number of Children	Age of Husband	Duration of Infertility (years)	Duration of Treatment (years)
Type of Infertility	1.000					
Age of Wife	.217	1.000				
Number of Children	.655	.242	1.000			
Age of Husband	.199	.716	.230	1.000		
Duration of Infertility (yrs)	.128	.381	.150	.319	1.000	
Duration of Treatment (yrs)	.094	.208	.041	.226	.570	1.000

Since our data consisted of mixed variables categorical and continuous and One way to avoid the scaling problem , we performed principal component on the correlation matrix rather than the covariance matrix. Correlation between number of children and type of infertility was nearly 0.7, which was very high. This may be because women who suffered from primary infertility had no children compared to women with secondary infertility who usually had children. It also interesting that there was very good relationship between duration of infertility and duration of treatment.

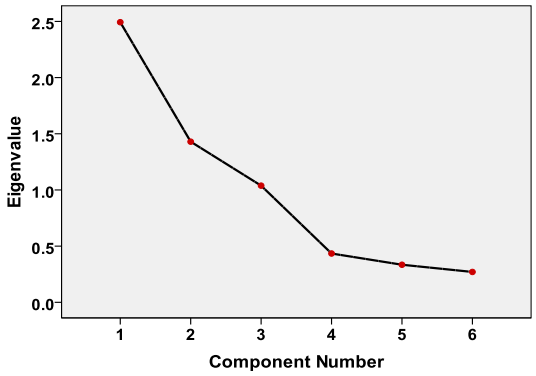

Figure 1. A scree diagram for data

Kaiser-Mayer-Olkin was a measure of whether the sample distribution of values is adequate for conducting principal component analysis. A measure > 0.9 is generally thought of as excellent, > 0.8 as good, > 0.7 as acceptable, > 0.6 as marginal, > 0.5 as poor. Our analysis showed that the Kaiser-Mayer-Olkin is equal to 0.611 which is acceptable.

Bartlett test of sphericity is essentially a measure of the multivariate normality of our set of distributions. A significance value of 0.00 < 0.01 with 15 degree of freedom, suggested that these data do not differ from significantly from normal.

A scree diagram of the components was given in figure 1. This is clearly indicated that there were three components in the data, a fact which is further emphasized by examining the eignvalues, the first three of which had values greater than one.

The first three principal component explained nearly 83% of the total variance. The first component alone accounted 42% of the variation, that is half of the total variation of the first three components.

Therefore sample variation was summarized very well by 3 principal components and a reduction in the data from 779 observations on 6 variables to 779 observations on 3 principal components was reasonable.

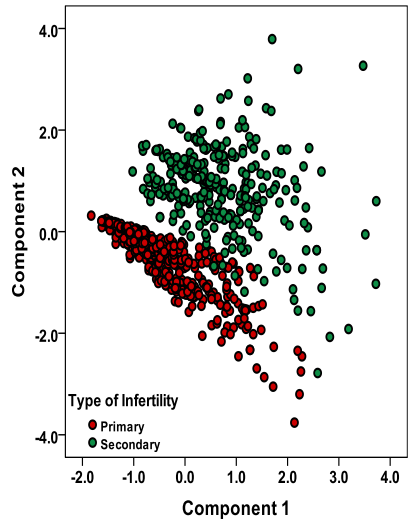

Figure 2.

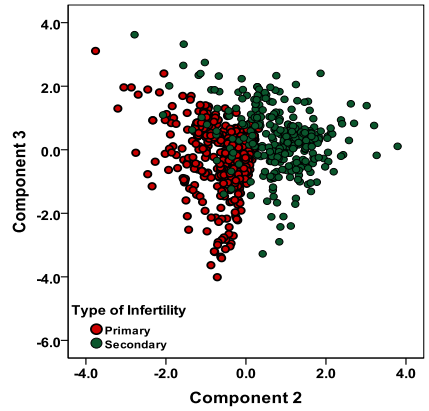

Figure 3.

The first scatter plot, contain nearly 42% of the variation in the data., it be seen that much of this variation is because the data separate into two groups along the first principal component.

In summary, the first three principal component explained nearly 83% of the total variance. The first component alone accounted 42% of the variation, that is half of the total variation of the first three components

Reference

1. Brian S.E. & Grham D.(1991*).Applied multivariate data analysis*. (3er ed.) . New York : John Wiley & Sons.

2. Chatfield C. & Collins A.J.(1980).*Introduction to multivariate analysis*. Chapan & Hall.

3. Mathada Sivamurthy and Chetna M. Sivamurthy. 2009. *Principal Components Analysis of ASFR (Application to the Recent Fertility Schedules around the World).*

4. Richard A.J. & Dean W.W. (2007).*Applied multivariate statistical analysis.*(6th ed.), Prentice Hall.

5. ZAFAR M. I. (1996) . *Husband-wife Roles as a Correlate of Contraceptive and Fertility Behaviour.* The Pakistan Development Review 35 : 2 (Summer 1996) pp. 145—170