JKAU: Sci., **vol. 12,** pp. 33-39 (1420 A.H. / 2000 A.D.)

An Application of Principal Component Analysis to Data Supplied by Some Saudi Arabian Male Diabetics

B.M. ASSAS
Department of Statistics, King Abdulaziz University,
Jeddah, Saudi Arabia

ABSTRACT. An application of Principal Component Analysis (PCA) to a data set. The 10 factors included are likely to have an effect on blood sugar level of diabetics. It appears that none of the factors are redundant. Principal Component Analysis was applied, using Saudi Arabian male diabetic patients attending King Abdulaziz University Hospital [400 questionnaires were distributed to patients: 300 replied]. The information considered here is age, bodymass, family history, composition of children, marital status and food habits.

Introduction

A data set relating to some Saudi Arabian male diabetic patients was used as a case study in Principal Component Analysis. The patients, who had already been diagnosed as diabetic, attended King Abdulaziz University Hospital. 400 questionnaires were distributed during 1997: 300 patients replied. Fasting blood sugar levels were available for each patient[1].

Factors, Labels and Coding of the Variables

The following provides a list of labels and coding, together with their explanations where needed.

Age	:	X_1 (in years)
Bodymass	:	X_2 (weight in kilograms) / (height in metres)^2
Family History	:	X_3 = 1 if either father or mother of patient or both were also diabetic, = 0 otherwise.

Composition of Children : $(X_4, X_5) = (0, 0)$ if there are no children, $= (1, 0)$ if there is at least one male child, $= (0, 1)$ if there is no male child but at least one female child.

In Saudi society, as in various other societies, parents suffer some stress if no children are born. Also, absence of a male child sometimes causes concern. We can recognize three broad types exposed to different kinds of stress. The above coding (in preference to simply counting male and female children) is done to differentiate the 3 classes.

Marital Status : $(X_6, X_7) = (0, 0)$ if the person is married at the time interviewed, $= (1, 0)$ if the person is divorced, $= (0, 1)$ if the person is a widow/widower.

Food Habits : $(X_8) = 1$ if the food is high in carbohydrate, $= 0$ if the food is low in carbohydrate.

$(X_9) = 1$ if the food is high in protein, $= 0$ if the food is low in protein.

: $(X_{10}) = 1$ if the food is high in fat, $= 0$ if the food is low in fat.

Correlation Matrix between the X's

The table below gives the lower triangle of the correlation matrix for the 10 variables $X_1, X_2, ..., X_{10}$.

TABLE 1. Correlation matrix of the 10 variables: only the upper triangle is given.

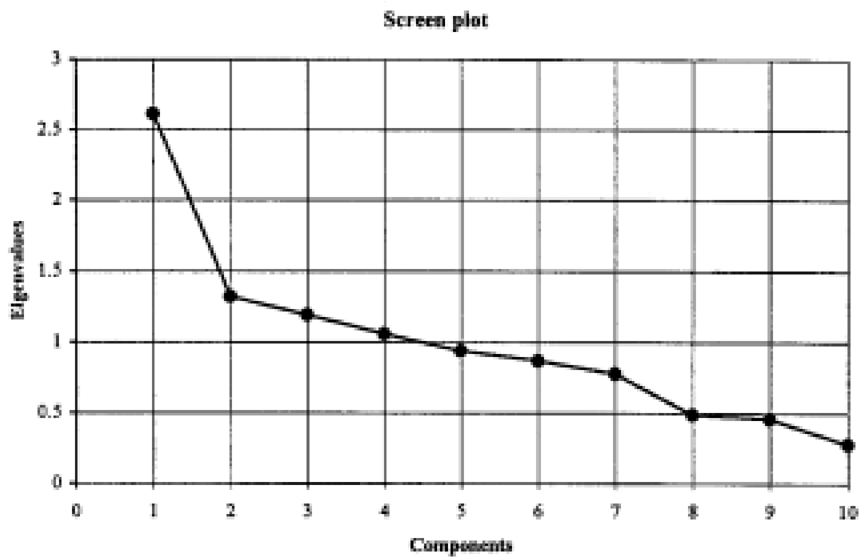
X_i \ X_j	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	–	– 0.101	0.454	0.199	– 0.144	– 0.159	– 0.104	– 0.042	– 0.092	– 0.035
X_2	–	–	– 0.247	– 0.330	0.197	0.350	0.208	0.071	0.034	0.089
X_3	–	–	–	0.390	– 0.308	– 0.204	– 0.256	0.002	– 0.100	– 0.025
X_4	–	–	–	–	– 0.703	– 0.204	– 0.204	– 0.056	– 0.11	– 0.046
X_5	–	–	–	–	–	0.107	0.257	0.100	0.052	0.005
X_6	–	–	–	–	–	–	– 0.190	0.042	0.042	0.095
X_7	–	–	–	–	–	–	–	0.019	0.052	0.017
X_8	–	–	–	–	–	–	–	–	0.203	0.072
X_9	–	–	–	–	–	–	–	–	–	0.094
X_{10}	–	–	–	–	–	–	–	–	–	–

Looking at Table 1 we notice that there are relatively few highly significant correlations. The highest is that between X_4 and X_5, which are indicator variables relating to the composition of children belonging to the patient. Other high correlations are between (X_1, X_3), (X_3, X_4), (X_2, X_6), (X_2, X_4), (X_3, X_5), (X_5, X_7) and (X_3, X_7). Only these seven correlations are significant at the 1% level. A Principal Component Analysis (PCA) of this correlation matrix is therefore unlikely to reduce the dimensions considerably.

The Eigenvalues of this matrix, the proportion of the total variability accounted for, and the cumulative proportions, are given in Table 2.

TABLE 2. Eigenvalues, proportion of variability and the cumulative proportion of the variability for the correlation matrix in Table 1.

Eigenvalues	2.6082	1.3236	1.1893	1.0607	0.9408	0.8668	0.7806	0.4922	0.4618	0.2761
Proportion	0.261	0.132	0.119	0.106	0.094	0.087	0.078	0.049	0.046	0.028
Cumulative proportion	0.261	0.393	0.512	0.618	0.712	0.799	0.877	0.926	0.972	1.000



As is readily seen in the screen plot, the last three components are very small, so there are at most 7 important components. However, if one considers eigenvalues less than 1 to be unimportant, then there are really 4 important ones. So the number of dimensions has been reduced from 10 to 7, or even 4. [There are no hard and fast rules for the choice, except a subjective feeling[2,3]].

Eigenvectors

We now give the list of all the 10 principal components together with their associated Eigenvectors (i.e. the multiplier coefficients for X_1, X_2,...,X_{10}) which will produce these components.

TABLE 3. The eigenvectors associated with each principal components.

Principal component	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
PC1	0.303	0.344	0.429	0.499	– 0.441	– 0.230	– 0.273	– 0.102	– 0.141	– 0.066
PC2	0.084	– 0.257	– 0.030	– 0.151	0.257	– 0.578	0.482	– 0.249	– 0.252	– 0.381
PC3	0.260	– 0.056	0.237	– 0.043	0.104	– 0.341	0.219	0.606	0.539	0.208
PC4	– 0.654	– 0.297	– 0.385	0.278	– 0.310	– 0.230	0.017	0.058	0.315	– 0.040
PC5	– 0.040	– 0.171	0.042	– 0.083	0.184	0.183	– 0.367	0.357	0.109	– 0.787
PC6	– 0.157	– 0.608	0.009	– 0.236	0.360	– 0.072	– 0.397	0.084	– 0.230	0.429
PC7	0.165	– 0.149	0.036	– 0.178	0.106	0.020	– 0.193	– 0.637	0.680	– 0.031
PC8	– 0.561	0.368	0.641	– 0.103	0.105	– 0.284	– 0.160	– 0.102	0.005	– 0.019
PC9	0.193	0.394	– 0.440	– 0.039	– 0.015	– 0.568	– 0.535	0.036	– 0.046	0.009
PC10	– 0.004	0.112	– 0.055	0.736	0.657	0.037	0.004	– 0.068	0.058	0.033

We first look at the multiplier coefficients for the three redundant components PC8, PC9 and PC10. PC10 as we might have guessed from our inspection of the correlation matrix, is a linear function of X_4 and X_5. (These 2 indicator variables are highly negatively correlated).

PC9 is essentially a linear combination of X_2, X_3, X_6 and X_7, some of which were noticed as being highly correlated. PC8 is a linear combination of X_1, X_3 and less importantly, X_2 and X_6. Again we see that X_1 and X_3 are highly correlated.

There appear to be no obvious contenders for redundant variables among these components.

Looking now at the important components, we see that PC1, which accounts for only 26.1% of the total variability in the X's, has relatively high coefficients for all the X's except X_8, X_9 and X_{10}. It describes a linear comparison between X_1 and X_4, and between X_5 and X_7. There is no obvious practical interpretation of this component.

For the other components, one could say that PC4 is accounting mainly for X_1 = Age, PC5 is accounting mainly for X_{10} = Fat content of diet and PC6 is accounting mainly for X_2 = bodymass, i.e. PC4, PC5 and PC6 each relate primarily to a single variable. This is a consequence of the relatively few highly significant correlations between the original variables.

PC7 relates to a contrast between X_8 and X_9. PC2 is a linear combination of most variables, except X_1, X_3 and X4 and PC3 is a linear combination of most variables except X_2, X_4 and X_5. PC3 has large coefficients for X_5 and X_9. Neither of these has an obvious interpretation.

We have identified at least 3 redundant components, but overall it appears that all 10 of the original variables are contributing to the overall variation in the original data.

Components and Blood Sugar

It is of some interest to see how important components are related to the blood sugar levels.

TABLE 4. Correlation coefficient of principal components with fasting blood sugar levels.

Principal components	PC1	PC2	PC3	PC4	PC5	PC6	PV7	PC8	PC9	PC10
Correlation coefficient	– 0.768	– 0.090	– 0.125	0.089	– 0.127	– 0.096	– 0.046	– 0.128	0.019	– 0.179

Table 4 gives the correlation coefficients for the component scores and fasting blood sugar level. Looking at Table 4 we can see that the first component has by far the highest correlation. None of the others is significant.

The multiple regression of blood sugar level on the 10 components for the 300 patients is:

$$\begin{aligned} Y \quad = \quad & -49.0\ \mathrm{PCI} - 8.08\ \mathrm{PC2} - 11.8\ \mathrm{PC3} \\ & -8.95\ \mathrm{PC4} - 13.5\ \mathrm{PC5} - 10.7\ \mathrm{PC6} \\ & -5.36\ \mathrm{PC7} - 18.8\ \mathrm{PC8} - 2.89\ \mathrm{PC9} \\ & -35.2\ \mathrm{PC10} \end{aligned}$$

leading to R^2 = 69.9%, which checks with the result of R^2 from multiple regression of Y on X_1, X_2,..., X_{10}[1]. In the multiple regression analysis, the variables X_1, X_8 and X_9 appeared to be superfluous.

However, the PCA tells us that none of the 10 variables is really redundant. So the implication is that all 10 variables are worth recording in future studies.

References

[1] **Assas, B.M., Samiuddin, M., Samra, A.A.** and **Meccawi, A.,** *Statistical Studies of Sample Data of Diabetes Patients in the Western Region of Saudi Arabia,* Project No. 082/414, King Abdulaziz University, Jeddah, Saudi Arabia (1996).

[2] **Chatfield, C.** and **Collins, A.J.,** *Introduction to Multivariate Analysis*. Chapman and Hall, London (1980).

[3] **Flury, B.** and **Riedwyl, H.,** *Multivariate Statistics,* Chapman and Hall, London (1988).

استخدام أسلوب المركبات الرئيسية في تحليل بيانات السكر في الدم لدى مجموعة من المرضى السعوديين الذكور

بكري معتوق عساس

قسم الإحصاء - كلية العلوم - جامعة الملك عبدالعزيز

جـــــدة - المملكة العربية السعودية

المستخلص . يعد مرض السكر في الدم من أخطر أمراض العصر الذي أصبحت نسبة الإصابة به عالية في العالم . وقد تم جمع بيانات عن مجموعة من المتغيرات الملائمة لدراسة هذا المرض من عينة من السعوديين الذكور المصابين . والمتغيرات التي تم جمع البيانات منها : العمر ، السمنة ، التاريخ الأسري ، عدد الأطفال ، الحالة الاجتماعية ، العادات الغذائية وأخيراً مستوى السكر في الدم . ولقد نوقشت المشكلة المألوفة لتقليل عدد المتغيرات المستقلة من خلال تحليل الانحدار وطريقة المركبات الرئيسية ، وقد وجد أهمية مساهمة جميع المتغيرات المستقلة في البيانات المتوفرة .