Research Journal of Mathematical and Statistical Sciences ________________________________ISSN 2320–6047
Vol. **2(1),** 6-9, January **(2014)** *Res. J. Mathematical and Statistical Sci.*

A Pictorial Representation of Multivariate data and its limitation: An example from Anthropometric data

Seal Babulal[1] and Pal Baidyanath[2]
[1]Department of Statistics, Burdwan University, Burdwan, West Bengal, INDIA
[2]Biological Anthropological Unit, Indian Statistical Institute, Kolkata, West Bengal, INDIA

Available online at: www.isca.in, www.isca.me
Received 26th December 2013, revised 2nd January 2014, accepted 11th January 2014

Abstract

Higher dimensional observations cannot be plotted to check dependency among its components and thus cannot be investigated whether these come from mixture population. However, a method is obtained for these two purposes in a simpler manner. Whole observations are realized by its first two principle components and from the scatter plot of these, nature of mixing and dependency may be obtained. Even scanning the pixel value of the plot, it is found that the distribution retrieved from pixel values and original distributions are different. It is due to limitation of scanning. The whole work is based on a large dimensional anthropological data set. Although, the first two principal components should be uncorrelated if it is from a unimodal distribution, but from a large data set, an impression of dependency is seen. It is shown that, it is due to mixing of distributions. Thus it is a way of identifying mixture population also, with higher dimensional observations. Also the impact over dependency of components is shown in presence of mixing.

Keywords: Pixel, anthropometric measurement, mixture distribution.

Introduction

In practice we plot bivariate data and study the scatter diagram to understand the joint relationship of the variables. But the point is that: from pictorial representation we may not recover the exact joint distribution clearly, as many points overlap on others and as from shades, density or frequency cannot be recovered exactly. This is a starting point of study in this paper.

Actually we were trying to reduce the dimensions of large number of anthropometric data and for the same we had tools of Principle Component Analysis method mainly. After plotting these two component scores we find very irregular shapes of the joint pattern and still an impression of relation. In fact, this should become uncorrelated if data set is from a non-mixture distribution, but from the pictorial distribution, it appeared that population is mixture of two populations, as evidenced from two uncorrelated scatter pattern in two places i.e. with a trend also. In a following section it is shown that mixing of two uncorrelated sets lead to correlated pattern. To see whether it is so, we read the pixel values with appropriate assigned frequencies as obtained from pixel count. Then, we see difference between actual bivariate distributions of first two principle components and the bivariate distribution obtained from pixel count of the same plotted data. In order to compare, we have scaled both and then used Chi-Square test of goodness of fit whether one table was generated from the other. But the test method rejects the hypothesis. That is why it is reasonable to believe that pictorial diagram has some limitations. Also from the plotting of two components we have an impression of pattern i.e. a relationship, which indicates mixing. Thus, it becomes a way to check whether multidimensional observations are from a mixture distribution in a simpler way.

Now let us give the application of pca to some areas and especially in biological anthropology in reducing dimension, as starting from real life data from this area, we came to this work and some problems related to data from this area.

In general, anthropometry is the scientific measurements of human body from which the shape and size of the body in different postures are measured. It refers to collection of physical dimensions of human body. Understanding of human body shapes can be useful in many applications of human species. These vary with age, gender and other determinant, non determinant factors. In the field of anthropological research several data sets with enormous number of variables are obtained, from the human body segments. Apart from these other data are obtained from physiological and demographic fields related to body. In this case, analyzing the data, it is very difficult to draw conclusion due to multi-co linearity and to remove such problem, it requires reducing the number of involved variables. This is achieved by transforming to a new set of variables, using the Principle Components (PC). In this paper we describe how it's two dimensional distributions may be retrieved from its two dimensional pictorial presentation through the image, viz reading its pixel values and then check relationship between variables and pixel values. The technique of principal component analysis was first described by Karl Pearson[1]. The statistical properties of principal components were investigated in detail by Hotelling[2]. Principal component and related form of multivariate analysis have been used by a number of authors in attempt to describe complex growth patterns in terms of a minimum number of basic trends. Biometrical studies (Bailey)[3]; Kraus and Choi[4], Olson and Miller[5] of character-complexes have made increasingly clear that, whenever functional relationships is considered, joint consideration of all characters become important and multivariate statistical techniques are necessary. Conventional principal component analysis minimizes the total error variance, which may be inappropriate even in the non-Gaussian distribution systems. Geo et at.[6] proposed entropy in a more general case for such model, and then modified it with the optimization for the minimum error entropy through a genetic algorithm. In that technique the extra constraint is in the form of a bound on the sum of the absolute values of the loading in that component. This type bound used regression equation Tibshirani[7], where similar problems of interpretation occur. Using curvature, inference[8] and regression of nonlinear type is attempted.

In section 2 we give the fact that mixing of two independent populations gives rise to dependent bivariate observation, though the individual components are uncorrelated. This suggests that our data is not from a homogeneous population but from mixing population and therefore the diagram of two pc's which carry maximum representation of the data says that original is mixture also.

Section 3 gives scatter diagram of two pc's[9, 10]. In section 4 methods of scaling the bivariate data and bivariate data after reading the pixel values are given and also the formulation of comparing true distribution and graphical representation is given. Tables are given after computations in section 5. In section 6, the analysis and conclusions are given and it is found that graphical representation differs from true distribution, perhaps it is due to limitation of reading grey values.

Impact over correlation due to mixing of two bivariate populations

For this, suppose (X, Y) ~ $pf_1(x,y) + (1\text{-}p)\ pf_2(x,y)$, i.e. observations are form mixture distribution. Now let us find out expression for linear relationship.
We have $EX = pE_1X + (1\text{-}p)\ E_2X$ and $EY = pE_1Y + (1\text{-}p)\ E_2Y$
Therefore, $EXEY = p^2E_1XE_1Y+ (1\text{-}p)^2\, E_2X\, E_2Y+p\,(1\text{-}p)(E_1XE_2Y+E_2X\, E_1Y)$
If X and Y are independent in each population separately, then, $EXY = pE_1XY+ (1\text{-}p)\ E_2XY$
$= pE_1XE_1Y+(1\text{-}p)E_2XE_2Y$

Now,
$EXY\text{-}\ EXEY = pE_1XE_1Y+(1\text{-}p)E_2XE_2Y\text{-}\ p^2E_1XE_1Y\text{-}(1\text{-}p)^2E_2X\, E_2Y\text{-}\ p(1\text{-}p)(E_1XE_2Y+E_2X\, E_1Y)$
$= p(1\text{-}p)E_1XE_1Y+p(1\text{-}p)E_2XE_2Y\text{-}\ p(1\text{-}p)(E_1XE_2Y+ E_2XE_1Y)$
$= p(1\text{-}p)[E_1XE_1Y+E_2XE_2Y\text{-}E_1X\, E_2Y\text{-}E_2XE_1Y)]$
$=p(1\text{-}p)[E_1X(E_1Y\text{-}E_2Y)\text{-}E_2X(E_1Y\text{-}E_2Y)]$
$= p(1\text{-}p)(E_1Y\text{-}E_2Y)(E_1X\text{-}\ E_2X)$. Hence,

$V(X) = EX^2 – (EX)^2\ = p\ E_1X^2+(1\text{-}p)E_2X^2–(pE_1X+(1\text{-}p)E_2X)^2$
$= p[V_1(X) + (E_1X)^2]+(1\text{-}p)[V_2(X)+(E_2X)^2]–p^2(E_1X)^2– (1\text{-}p)^2(E_2X)^2–2p(1\text{-}p)E_1XE\ {}_2X$
$= pV_1(X)+(1\text{-}p)V_2(X)+p(E_1X)^2+(1\text{-}p)(E_2X)^2\text{-}p^2(E_1X)^2– (1\text{-}p)^2(E_2X)^2–2p(1\text{-}p)E_1XE\ {}_2X$
$= pV_1(X)+(1\text{-}p)V_2(X)+p(1\text{-}p)(E_1X\text{-}E_2X)^2$
From above we see that, though individual bivariate distributions are independent, the mixture of these is dependent.

Plot of the bivariate data realized through PCs

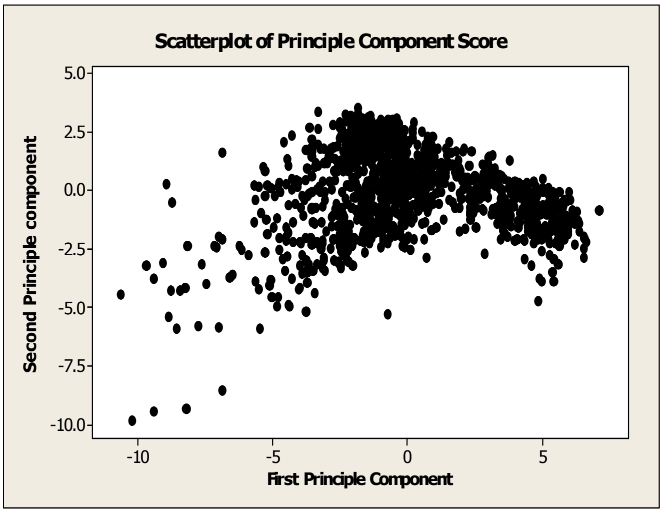

Figure-1
Two isolated scatters but having an overall pattern, perhaps due to mixing population

Method of comparing graphical representation with original tabular form data: Multivariate anthropometric data from the fishing community of Digha, West Bengal and Udaypur, Odissa are the point of study. As the dimension is huge, at first its principal component is chosen. Apart from studying anthropometric information, the main point is to study the relationship between the principal components. In order to do that we have plotted these and have taken also its bivariate distribution. Then in order to check the closeness of the pictorial presentation and data representation in terms of frequency, we have traced backward i.e. from pictorial presentation, we have scanned its pixel values which may be assumed to be proportion to which may be assumed to be proportion to its frequency of that portion if the picture truly reads the data. So from this we make a bivariate distribution and initially we have also the bivariate distribution of the true data. Then in order to compare these two, we make proper scaling. After that we use chi-square goodness of fit to test whether these two distributions are equal. From this test we see that the hypothesis of equality is rejected.

Principal component analysis was performed on 19 anthropometric variables. We consider only first two significant principle component scores. On a bivariate plot, the abscissa X-axis represents the first PC score and the ordinate Y-axis represents second score. Each point of the plot corresponds to the X and Y scores from a 19 variate observation as realized by pcs. The images are converted to grey-scale. The plot may be regarded as a graphical display of matrix. The matrix consists in m rows and n columns. The total resolution area (m x n) is divided into 10 x10 pixels which generates so many small square areas namely, boxes of size 100 sq. pixel. Each box may be blank or partial or complete covered by the solid circle. Maximum pixel number of each circle is 36 and consequently counts the pixel number in each box. The range of pixel in each box is 1 to 100 while eliminating the zero pixel boxes. It re-expresses the pc’s score with a new set of axes through image coordinate ($I_{1,,}I_2$) transformation of the midpoint of each box. Finally image coordinates are converted into principle component coordinate (P, Q) which are less than original principle component.

The following transformations are used to change the original X-score and Y score by applying on it.

$$\frac{\mathrm{I1-Min(I1)}}{\mathrm{Max(I1)-Min(I1)}} = \frac{\mathrm{P-Min(PC1)}}{\mathrm{Max(PC1)-Min(PC2)}} \qquad (1)$$

and

$$\frac{\mathrm{I2-Min(I2)}}{\mathrm{Max(I2)-Min(I2)}} = \frac{\mathrm{Q-Min(PC2)}}{\mathrm{Max(PC2)-Min(PC2)}} \qquad (2)$$

By solving the equations (1) and (2) we get P and Q respectively corresponding to each I_1 and I_2 .By using the value of ($PC_{1,}\ PC_2$) and (P,Q) we prepare the two bivariate tables with same scale range. We want to test the hypothesis that the samples collected by using the pixel point (P, Q) and the PCA values ($PC_{1,}$ PC2) belong to the same distribution. To reach the above objective, Chi-square test for the equality or homogeneity of two distributions is adopted.

Tables of original values and image values under same range

Table-1
Original values and image values under same range

Sl. No.	pc1	pc2	I1	I2	Pixel	Mid_X-_axis	Mid_Y-_axis	Converted to_pc1 scale(a)	Converted to _pc2 scale(b)	Class x froma col.	Class y from b col.
1	-6.880	-2.025	0	170	9	5	165	-7.120	0.739	-2.00	2.00
2	0.642	1.790	0	190	9	5	185	-7.120	1.249	-2.00	2.00
3	-3.921	-0.166	0	160	13	5	155	-7.120	0.484	-2.00	2.00
4	-0.109	0.326	0	210	13	5	205	-7.120	1.760	-2.00	2.00
5	-3.723	-2.934	0	200	43	5	195	-7.120	1.505	-2.00	2.00
6	-2.163	0.898	10	170	6	15	165	-6.673	0.739	-2.00	2.00
7	-3.844	-1.052	10	160	15	15	155	-6.673	0.484	-2.00	2.00
8	-5.246	-2.588	10	180	29	15	175	-6.673	0.994	-2.00	2.00
9	-4.737	-1.985	10	250	32	15	245	-6.673	2.781	-2.00	2.00
10	-3.519	0.004	10	240	56	15	235	-6.673	2.526	-2.00	2.00
....											
	3.254	0.263	230	30	8	235	25	3.153	-2.834	2.00	-2.00
	4.445	0.349	230	60	29	235	55	3.153	-2.069	2.00	-2.00
	-0.928	2.027	240	60	8	245	55	3.600	-2.069	2.00	-2.00
	-2.174	-2.365	240	10	29	245	5	3.600	-3.345	2.00	-2.00
	-0.512	1.012	240	30	29	245	25	3.600	-2.834	2.00	-2.00
	-0.036	-0.036	0.2								
	-1.723	-1.723	1.6								
	0.536	0.536	1.1								
	6.134	6.134	-0.9								

This will be used to build up following tables, leading to test procedure.

Testing equality of the graphical representation and Tabular form: PC scores (Expected) and pixel count (Observed) are given in the following table.

Table-2
Table2 of PC scores i.e. assumed expected frequency

Expected (x/y)	**-2**	**0**	**2**	**Total**
-2	103	65	58	226
0	15	124	299	438
2	39	249	204	492
Total				1156

Table 2 is to be compared with table 4.

Table-3
Table 3 of Observed pixel counts

Observed(x/y)	**-2**	**0**	**2**	**Total**
-2	107	1219	15751	17077
0	235	458	1020	1713
2	202	124	83	409
Total				19199

Table 3 is necessary for table 4.

Table-4
Table 4- Transformed Table 3 with respect to total frequency 1156

Observed(x/y)	**-2**	**0**	**2**	**Total**
-2	6.442627	73.39778	948.3909	1028.231
0	14.1497	27.57685	61.4157	103.1422
2	12.16272	7.466222	4.997552	24.62649
Total				1156

Table 2 and table 4 are used to calculate the following table to test whether distribution obtained from tabular form and from pixel counts are equal.

From above table it is to be checked, whether table-2 and transformed table-4 are same or not i.e. goodness of fit test. In the following table calculation for Chi-square is given.

Table-5
Calculation for Chi-square

(x/y)	**-2**	**0**	**2**	**Total**
-2	28.5291	11.93894	822.8458	915.5029
0	12.22037	5.73587	63.77514	371.9161
2	10.24494	133.1837	11913.27	1306.514
Total				**2593.933**

Conclusion

Frequency table obtained after reading pixel values from pictorial representation is different from the original frequency table.

Reference

1. Pearson Karl., On lines and planes of closest fit to systems of points in space, Philos. Mag., Ser., **6(2),** 559-572 **(1901)**
2. Hotelling H., Analysis of a complex of statistical variates into principal components, Jour. Educ. Psych., **24,** 417-441, 498-520 **(1933)**
3. Bailey D.W., A comparison of genetic and environmental principal component of morphogenesis in mice, Growth, **20,** 63-74 **(1956)**
4. Kraus B.S. and Choi SA.C., A factorial analysis of the prenatal growth of the human skeleton, Growth, 22, 231-242 **(1958)**
5. Olson E.C. and Miller R.L., Morphological Integration. Xv + 317. University of Chicago Press, Chicago, **(1958)**
6. Guo Z., Yeu H. and Wang H., A modified PCA based on the minimum error entrophy. Proceeding of the 2004 Americal Control Conference, Boston, Massachusetts June 30- July 2, 2004 **(2004)**
7. Tibshirani R., A comparison of some error estimates for neural network models" Neural Computation, **8,** 152-163 **(1996)**
8. Seal B. and Sadhu S., Inference in a curved poisson distribution, *Research journal of Mathematical and statistical sciences*, ISSN 2320-6047, **1(5),** 6-16 **(2013)**
9. Rao C.R., Linear Statistical Inference, Wiley, New York, **(1973)**
10. Johnson R.A. and Wicheren D.W., Applied multivariate statistical Analysis. Prentice Hall of India, New Delhi, **(2001)**