



Multivariate Analysis on various samples for water pollution

Iqbal Kaur

Faculty of Mathematics, Govt P.G. College,
Karnal, Haryana-india

Abstract: It is common pollution studies of Hisar City of the Haryana State to have multiple informant reports for the pollution control. The key challenge in analyzing multiple informant data concerns how they should best be represented in statistical models [1]. Here we propose Principal Component Analysis, Cluster Analysis and Correlation Analysis as the preferred methods when the multiple informant outcome data are continuous. This approach permits the informant-specific information about pollution outcomes to be included in a single regression analysis, at the same time adjusting for the correlation between informant responses. The advantages of using a multivariate model[1] include the ability to: (1) test for informant differences in outcome and assess if the effect of a risk factor on the outcome varies by informant; (2) estimate separate effects for each informant where necessary, or common effects where appropriate; (3) estimate the correlation between informant reports; (4) appropriately handle missing data by including data from all subjects with at least one informant report. To illustrate the application of this approach an example from the Pollution Study is presented, examining risk factors for "Internalizing" behavior using pollution control board informants.

Key words: Principal Component Analysis, multivariate analysis, risk factors, pollution control.

I. Introduction

The central issue in analyzing multiple informant data is how to represent multiple outcomes for a subject from different informants in a statistical model [1]. There are several traditional approaches for analyzing multiple informant data in psychiatric research that are not completely satisfactory. One approach is to conduct separate regression analyses for each of the informants and to report the results separately [1]. This approach has a number of distinct drawbacks[1]: (1) separate analyses yield multiple and often differing sets of results for the different informants; (2) separate analyses provide no formal statistical means of evaluating how similar or different the results are across the various informants; (3) in cases where the separate analyses yield results that are sufficiently similar, this approach provides no formal means of summarizing effects in a single set of results; (4) separate analyses may be based on different subsets of the data if some subjects are missing one informant report and others are missing another.

The other main alternative is to use some sort of "pooling" strategy[1]. There are three common pooling strategies which produce a summary measure from informant ratings to be used as the single outcome variable. Major disadvantages of these pooling strategies include: (1) the optimal algorithm for combining outcomes depends upon the type of measurement error present; although the measurement error in informant reports of psychopathology is believed to be substantial and also complex, there have been few studies that would support the use of one pooling strategy over another; (2) pooling informant reports does not permit the assessment of potential differences in risk factor effects across the various informants; (3) many of the pooling algorithms are not clearly defined in the presence of missing informant reports.

For the special case where the outcome variable is dichotomous, Fitzmaurice et al.[2] suggest handling multiple informant data by using multivariate logistic regression. This approach permits the informant-specific information about case status to be included in a single multivariate regression analysis, at the same time adjusting for the correlation between informant responses. The advantages of such an approach include the ability to: (1) test for informant differences in outcome and assess whether the effect of a risk factor on the outcome varies by informant; (2) estimate a common risk factor effect if it does not; (3) obtain measures of prevalence based on each informant or based on combined data from all informants (where appropriate); (4) handle missing data appropriately by including data from all subjects with at least one informant report; (5) assess informant agreement. The current article extends the approach described in Fitzmaurice et al.[1] to consider the setting where the outcome variable is continuous rather than dichotomous. Multivariate linear regression is proposed in place of multivariate logistic regression.

II. Principal Component Analysis

In principal component analysis [3], we seek to maximize the variance of a linear combination of the variables [3]. For example, we might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking[3]. All the linear combinations that we have considered previously were related to other variables or to the data structure. In regression, we have linear combinations of the independent variables that best predict the dependent variable(s); in canonical correlation, we have linear combinations of a subset of variables that maximally correlate with linear combinations of another subset of variables; and discriminant analysis involves linear combinations that maximally separate groups of observations. Principal components, on the other hand, are concerned only with the core structure of a single sample of observations on p variables. None of the variables is designated as dependent, and no grouping of observations is assumed [4].

The first principal component is the linear combination with maximal variance; we are essentially searching for a dimension along which the observations are maximally separated or spread out. The second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component, and so on. In general, the principal components define different dimensions from those defined by discriminant functions or canonical variates. In some applications, the principal components are an end in themselves and may be amenable to interpretation. More often they are obtained for use as input to another analysis[3]. For example, two situations in regression where principal components may be useful are (1) if the number of independent variables is large relative to the number of observations, a test may be ineffective or even impossible, and (2) if the independent variables are highly correlated, the estimates of regression coefficients may be unstable. In such cases, the independent variables can be reduced to a smaller number of principal components that will yield a better test or more stable estimates of the regression coefficients. For details of this application[3,4].

Principal component analysis deals with a single sample of n observation vectors $y_1, y_2, \dots, y_n$ that form a swarm of points in a p -dimensional space. Principal component analysis can be applied to any distribution of y , but it will be easier to visualize geometrically if the swarm of points is ellipsoidal.

If the variables $y_1, y_2, \dots, y_p$ in y are correlated, the ellipsoidal swarm of points is not oriented parallel to any of the axes represented by $y_1, y_2, \dots, y_p$. We wish to find the natural axes of the swarm of points (the axes of the ellipsoid) with origin at $\bar{y}$, the mean vector of $y_1, y_2, \dots, y_n$. This is done by translating the origin to $\bar{y}$ and then rotating the axes. After rotation so that the axes become the natural axes of the ellipsoid, the new variables (principal components) will be uncorrelated.

We could indicate the translation of the origin to $\bar{y}$ by writing $y_i - \bar{y}$, but we will not usually do so for economy of notation. We will write $y_i - \bar{y}$, when there is an explicit need; otherwise we assume that y_i has been centered.

The axes can be rotated by multiplying each y_i by an orthogonal matrix A :

$$z_i = Ay_i \quad \dots\dots\dots 1$$

Since A is orthogonal, $A'A = I$, and the distance to the origin is unchanged:

$$Z_i' Z_i = (Ay_i)'(Ay_i) = y_i' A' A y_i = y_i' y_i$$

Thus an orthogonal matrix transforms y_i to a point z_i that is the same distance from the origin, and the axes are effectively rotated. Finding the axes of the ellipsoid is equivalent to finding the orthogonal matrix A that rotates the axes to line up with the natural extensions of the swarm of points so that the new variables (principal components) $z_1, z_2, \dots, z_p$ in $z = Ay$ are uncorrelated.

Thus we want the sample covariance matrix of z , $S_z = A S A'$ [see (3.64)], to be diagonal:

$$S_z = A S A' = \begin{pmatrix} S_{z1}^2 & 0 & \dots & 0 \\ 0 & S_{z2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_{zp}^2 \end{pmatrix} \quad \dots\dots\dots 2$$

where S is the sample covariance matrix of $y_1, y_2, \dots, y_n$. $C'SC = D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, where the λ_i 's are eigenvalues of S and C is an orthogonal matrix whose columns are normalized eigenvectors of S . Thus the orthogonal matrix A that diagonalizes S is the transpose of the matrix C :

$$A = C' = \begin{pmatrix} a'_1 \\ a'_2 \\ a'_3 \\ \vdots \\ a'_p \end{pmatrix} \quad \dots\dots\dots 3$$

where a_i is the i th normalized ($a_i' a_i = 1$) eigenvector of S . The *principal components* are the transformed variables $z_1 = a'_1 y, z_2 = a'_2 y, \dots, z_p = a'_p y$ in $z = Ay$. For example, $z_1 = a_{11} y_1 + a_{12} y_2 + \dots + a_{1p} y_p$.

The diagonal elements of ASA' on the right side of (2) are eigenvalues of S. Hence the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$ of S are the (sample) variances of the principal components $z_i = a'_i y$:

$$s_{z_i}^2 = \lambda_i. \quad \dots\dots 4$$

Since the rotation lines up with the natural extensions of the swarm of points, $z_1 = a'_1 y$ has the largest (sample) variance and $z_p = a'_p y$ has the smallest variance. This also follows from 4, because the variance of z_1 is λ_1 , the largest eigenvalue, and the variance of z_p is λ_p , the smallest eigenvalue. If some of the eigenvalues are small, we can neglect them and represent the points fairly well with fewer than p dimensions. For example, if $p = 3$ and λ_3 is small, then the swarm of points is an “elliptical pancake,” and a two-dimensional representation will adequately portray the configuration of points.

Because the eigenvalues are variances of the principal components, we can speak of “the proportion of variance explained” by the first k components:

$$\text{Proportion of variance} = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\sum_{j=1}^p s_{ij}} \quad \dots\dots 5$$

Since $\sum_{i=1}^p \lambda_i = \text{tr}(S)$. Thus we try to represent the p -dimensional points $(y_{i1}, y_{i2}, \dots, y_{ip})$ with a few principal components $(z_{i1}, z_{i2}, \dots, z_{ik})$ that account for a large proportion of the total variance. If a few variables have relatively large variances, they will figure disproportionately in $\sum_j s_{ij}$ and in the principal components.

However, in the case of discriminant functions and canonical variates, the eigenvalues are not variances, as they are in principal components.

If the variables are highly correlated, the essential dimensionality is much smaller than p . In this case, the first few eigenvalues will be large, and (5) will be close to 1 for a small value of k . On the other hand, if the correlations among the variables are all small, the dimensionality is close to p and the eigenvalues will be nearly equal. In this case, the principal components essentially duplicate the variables, and no useful reduction in dimension is achieved.

Any two principal components $z_i = a'_i y$ and $z_j = a'_j y$ are orthogonal for $i \neq j$; that is, $a'_i a_i = 0$, because a_i and a_j are eigenvectors of the symmetric matrix S. Principal components also have the secondary property of being uncorrelated in the sample; that is, the covariance of z_i and z_j is zero:

$$s_{z_i z_j} = a'_i S a_j = 0 \text{ for } i \neq j \quad \dots\dots 6$$

Discriminant functions and canonical variates, on the other hand, have the weaker property of being uncorrelated but not the stronger property of orthogonality. Thus when we plot the first two discriminant functions or canonical variates on perpendicular coordinate axes, there is some distortion of their true relationship because the actual angle between their axes is not 90° . If we change the scale on one or more of the y 's, the shape of the swarm of points will change, and we will need different components to represent the new points. Hence the principal components are not scale invariant. We therefore need to be concerned with the units in which the variables are measured. If possible, all variables should be expressed in the same units. If the variables have widely disparate variances, we could standardize them before extracting eigenvalues and eigenvectors. If one variable has a much greater variance than the other variables, the swarm of points will be elongated and will be nearly parallel to the axis corresponding to the variable with large variance. The first principal component will largely represent that variable, and the other principal components will have negligibly small variances. Such principal components (based on S) do not involve the other $p - 1$ variables, and we may prefer to analyze the correlation matrix R.

III. Case Study

The pollution studies of Hisar City for pollution control has been carried out. The data was collected from the pollution control board, Hisar on the live variables, given below in table 1 for twenty different samples taken from the different places located in Hisar city of Haryana state of India.

X ₁	pH value
X ₂	COD(Chemical Oxygen Demand)
X ₃	BOD(Biological Oxygen Demand)
X ₄	SS(Suspended Solids)
X ₅	Oil and Grease

Table 1: Data of twenty industries about various pollution parameters.

Code	Type of Industry	X ₁ (pH)	X ₂ (COD) mg/l	X ₃ (BOD) mg/l	X ₄ (SS) mg/l	X ₅ (Oil and Grease) mg/l
A	Surgical unit	6.96	232.5	577.6	369	0
B	Card Board Industry	8.02	145	361.6	397	0
C	Textile unit	7.06	110	373.6	736	0
D	Cold Rolling	2.24	29	515.2	710	24
E	Steel Unit(Picling)	1.35	23.5	348	163	0
F	Steel Unit(Hot Rolling)	2.16	36	278.34	119	0

G	Textile unit	10.9	24	310.4	386	0
H	Surgical unit	8.71	119	492	249	0
I	Cold Rolling	1.18	27	294.4	510	12
J	Sewerage of Hisar	6.83	390	611.2	918	0
K	Textile unit	7.86	390	513.6	692	0
L	Steel Unit(Picling)	2.6	25	286.4	118	0
M	Distillery Unit	6.72	2750	10000	976	10
N	Steel Unit (Cold Rolling)	1.79	17	692	576	56
O	Surgical unit	10.13	220	395.2	396	0
P	Steel Unit(Picling)	6.34	22.5	497.6	170	0
Q	Agro Industry	8.82	92.5	296	190	0
R	Leather Unit	7.91	135	291.2	160	0
S	Milk Plant	6.9	320	579.2	463	0
T	Textile unit	10.91	48	286	396	0

Multivariate Analysis

Principal Component Analysis: the researchers first evaluated the mean of the five variables as given in the table 2.

Table 2: Shows the mean of the five variables

X ₁	6.2695
X ₂	257.8
X ₃	899.977
X ₄	464.7
X ₅	5.1

After evaluating the mean, the sum of the deviation squares is calculated and produced the half matrix shown in the table 3.

Table 3: Deviation Squares and products of matrix

	X ₁	X ₂	X ₃	X ₄	X ₅
X ₁	197.4073	4027.7122	43534.4763	3422.447	404.129
X ₂		6824125	2401624	1560423.75	3176.4
X ₃			87485600	5121207	62851.95
X ₄				1606136.25	17776.6
X ₅					3435.8
	Trace of this matrix=95919494				

After on the transformation of matrix into variance covariance matrix as shown in table 4.

Table 4: variance covariance matrix

	X ₁	X ₂	X ₃	X ₄	X ₅
X ₁	10.39	211.98	2291.29	180.13	21.269
X ₂		359164.47	12401.2632	82127.56	167.179
X ₃			4604505.26	269537.21	3303.997
X ₄				84533.486	935.61
X ₅					180.83

The computed latent root of the matrix is shown in table 5 and corresponding latent vector shown in table 6.

Table 5: Eigen Value

PC#	Eigen value	%Variance	Cumulative Variance
PC ₍₁₎	94403334.938	98.419	98.419
PC ₍₂₎	1316835.917	1.373	99.792
PC ₍₃₎	197130.769	0.206	99.998
PC ₍₄₎	2091.728	0.002	100.00
PC ₍₅₎	125.291	0.000	100.00

Table 6: Eigen Vector

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
VAR 1	0.000	0.003	0.013	-0.123	0.992
VAR 2	0.265	0.111	0.995	0.075	-0.003
VAR 3	0.963	-0.090	-0.255	-0.020	0.001
VAR 4	0.058	0.990	-0.129	-0.019	-0.003
VAR 5	0.001	0.009	-0.079	0.989	0.124

The computed Principal Component loadings are shown in table 7.

Able 7: Principal Component loadings

Code	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
A	638.812	339.178	26.989	-2.191	5.469
B	409.332	376.572	-5.079	-5.074	6.497

C	431.126	707.124	-85.467	-14.394	4.524
D	544.477	659.909	-197.520	1.451	3.213
E	350.574	132.702	-87.413	-8.619	1.060
F	284.303	96.795	-52.001	-5.509	1.901
G	327.355	356.864	-106.096	-13.306	9.724
H	519.436	215.508	-43.994	-6.911	7.901
I	319.896	481.445	-116.280	-2.083	1.137
J	744.496	896.990	97.722	-1.587	2.993
K	637.540	682.077	151.893	4.639	4.683
L	289.088	93.863	-64.426	-6.538	2.385
M	10410.288	373.655	-51.895	-5.609	5.569
N	703.779	510.386	-239.259	31.248	7.404
O	461.488	380.893	58.122	-0.339	8.381
P	494.707	126.105	-127.368	-12.487	6.141
Q	320.360	171.766	-11.664	-3.799	8.100
R	325.274	147.218	34.015	0.198	7.154
S	668.950	441.775	97.959	2.569	4.803
T	345.349	965.445	-155.940	-22.792	7.580

The Principal Components are :

$$PC_{(1)} = 0.0X_1 + 0.265X_2 + 0.963X_3 + 0.058X_4 + 0.001X_5$$

$$PC_{(2)} = 0.003X_1 + 0.111X_2 - 0.09X_3 + 0.99X_4 + 0.009X_5$$

$$PC_{(3)} = 0.013X_1 + 0.955X_2 - 0.255X_3 - 0.129X_4 - 0.079X_5$$

$$PC_{(4)} = -0.123X_1 + 0.075X_2 - 0.020X_3 - 0.019X_4 + 0.989X_5$$

$$PC_{(5)} = 0.992X_1 - 0.003X_2 + 0.001X_3 - 0.003X_4 + 0.124X_5$$

Now from the table 5, it is clear that the first latent root correspond to 98.419% total variance of the given data.

So it is enough to consider just first eigen vector which gives:

$$PC_{(1)} = 0.0X_1 + 0.265X_2 + 0.963X_3 + 0.058X_4 + 0.001X_5$$

Based on the first principal component we considered the component loadings (Table 7) and arranged them in an order as given below:

Order No	Code	Type of Industry
1	M	Distillery Unit
2	J	Sewerage of Hisar
3	N	Steel Unit (Cold Rolling)
4	S	Milk Plant
5	A	Surgical unit
6	K	Textile unit
7	D	Cold Rolling
8	H	Surgical unit
9	P	Steel Unit(Picling)
10	O	Surgical unit
11	C	Textile unit
12	B	Card Board Industry
13	E	Steel Unit(Picling)
14	T	Textile unit
15	G	Textile unit
16	R	Leather Unit
17	Q	Agro Industry(Card Board)
18	I	Cold Rolling
19	L	Steel Unit(Picling)
20	F	Steel Unit(hot Rolling)

Thus the distillery unit is most polluted.

References:

- [1] Meredith A. Goldwasser and Aarrett A. Fitzmaurice, "Multivariate Linear Regression Analysis of Childhood Psychopathology using Multiple Informant Data", International Journal of Methods in Psychiatric Research.
- [2] Fitzmaurice, GM, Laird, NM, Zahner, GEP, Daskalakis, C. Bivariate logistic regression analysis of childhood psychopathology ratings using multiple informants. American Journal of Epidemiology 1995;142(11):1194-1203.
- [3] Alvin C. Rencher "Methods of Multivariate Analysis", Second Edition, Brigham Young University, a John Wiley & sons, Inc. Publication.
- [4] Rencher, A. C. (1998), *Multivariate Statistical Inference and Applications*, New York: Wiley.
- [5] Lavine, B., and Carlson, D. (1987), "European Bee or Africanized Bee?: Species Identification through Chemical Analysis," *Analytical Chemistry*, 59 (6), 468-470.
- [6] Mosteller, F., and Wallace, D. L. (1984), *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, New York: Springer-Verlag.
- [7] Lachenbruch, P. A. (1975), *Discriminant Analysis*, New York: Hafner.