

Dimension reduction in multivariate analysis using maximum entropy criterion

B. K. Hooda*

Department of Mathematics and Statistics

CCS Haryana Agricultural University

Hisar 125 004

India

D. S. Hooda†

Jaypee Institute of Engineering and Technology

A. B. Road, Raghogarh 473 226

District Guna, Madhya Pradesh

India

Abstract

In the present communication dimension reduction criteria in Multivariate data with no external variables are studied by using Entropy Optimization Principles. Maximum entropy criterion is provided and its relation with other criteria for selection of principal variables in multivariate analysis is established. A comparative study of performance of principal variables with the corresponding number of principal components is made by considering empirical data set.

Keywords : *Principal variables and components, covariance matrix, information loss, singular transformation and correlation matrix.*

1. Introduction

Researchers often record several characters in their research experiments where each character has a special significance to the experimenter.

*E-mail: bkhoda@hau.ernet.in

†E-mail: ds_hooda@rediffmail.com

Journal of Statistics & Management Systems

Vol. 9 (2006), No. 1, pp. 175–183

© Taru Publications

In real life problems the variation in one character may be concomitant with others so that the additional information supplied independently of the others may be negligible. It is thus useful to identify the variables, which simply complicate the analysis, and hardly supply any extra information. Principal component analysis is often used for reducing dimensions in multivariate data. According to McCabe (1984) principal components fulfill most of the desirable properties of a dimension reducing transformation, but they frequently fail to provide useful results and their interpretation is rather difficult. As principal components are linear combinations of all the original variables, so selection of a few principal components can reduce the dimensionality of the space, however, one has to interpret the results in terms of original number of variables. Thus, it would be useful if one can reduce the dimensions of the space, as well as the number of variables that are considered important for future measurements.

Selection of variables in regression setting has been investigated extensively in statistical literature, e.g. Aitkin (1974), Hocking (1976) and Draper and Smith(1998). For multivariate data with no predictor or external variable Jolliffe (1972) suggested the use of conditional variance-covariance matrix of the discarded variables given that of selected variables. Kapur and Kesavan (1992) showed that the information theoretic measures of stochastic dependence among a set of variates can be employed in the solution of the feature extraction problem. Hooda and Hooda (2001) defined a measure of generalized dependence and studied its application in pattern recognition.

The present paper focus on dimensional reduction in multivariate data with no external variables by using entropy optimization techniques. Maximum entropy criterion is provided and its relation with other criteria for selection of principal variables in multivariate analysis is established. Under the assumption of multivariate normality the variable selection criteria have been applied to a real data set generated and used by Parkash (2001). Performance of principal variables selection criteria is compared with the corresponding method of principal components.

2. Variable selection transformation

Suppose, each individual is characterized by p characters represented by the random vector $\mathbf{X} = (X_1, X_2, \dots, X_p)'$. Further suppose that \mathbf{X} has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We wish to transform

the p -dimensional random vector \mathbf{X} in to a m -dimensional vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)'$ in a m -dimensional space where $m \leq p$. For reduction in dimensions through the selection of m principal variables we can use a linear transformation

$$\mathbf{Y} = \mathbf{A}_m \mathbf{X} \tag{2.1}$$

where $\mathbf{A}_m = [a_{ij}]_{m \times p}$.

The transformation from \mathbf{X} to \mathbf{Y} involves a loss of information. This loss in information can be measured in terms of the differences in Σ_{xx} and Σ_{yy} . So we wish to determine the transformation matrix \mathbf{A}_m that minimizes the information loss.

Principal Component Analysis gives the optimal solution with eigen vectors corresponding to the m largest eigen roots of Σ_{xx} forming the columns of a matrix \mathbf{A} . However, if we wish to select the m original variables out of p observed variables the dimensional reduction transformation matrix \mathbf{A}_m must be of the form

$$\mathbf{A}_m = (\mathbf{I}_m, \mathbf{0}_{m \times (p-m)}) \tag{2.5}$$

or a matrix obtained by permuting the columns of \mathbf{A}_m .

For $p = 5$ and $m = 4$ we get the subset (x_1, x_2, x_3, x_4) and

$$\mathbf{A}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

By permuting the columns of \mathbf{A}_4 we get the matrices:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

which are given by the combinations (x_1, x_2, x_3, x_5) , (x_1, x_2, x_4, x_5) , (x_1, x_3, x_4, x_5) and (x_2, x_3, x_4, x_5) , respectively.

Instead of permuting the columns of A_m , in practice, it may be more convenient to permute the elements of X as

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}$$

where $X^{(1)}$ is a vector of m selected variables and $X^{(2)}$ be the vector of $(p - m)$ discarded variables. Accordingly, we partition Σ_{xx} as

$$\Sigma_{xx} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Here Σ_{11} is the $m \times m$ covariance matrix of $X^{(1)}$ and other elements have similar meaning. Thus, the selection of a set of m original variables is equivalent to selection of Σ_{11} from ${}^p C_m$ possible choices.

3. Principal variables selection criteria

In this section we describe various criteria for principal variable selection by using entropy optimization techniques:

3.1 Maximum generalized variance criterion

Out of all possible sets of m variables from the p -original variables we select the set that maximizes the generalized variance. Let Σ_{11} be the covariance matrix of the selected variables and $\Sigma_{22.1}$, the conditional covariance matrix of the remaining variables given the selected variables. The generalized variance $|\Sigma_{11}|$ represent the variation retained by the selected variables and the conditional covariance matrix $\Sigma_{22.1}$ contains the information left in the remaining variables given the selected ones. Writing the generalized variance $|\Sigma|$ as

$$\begin{aligned} |\Sigma| &= |\Sigma_{11}| \cdot |\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}| \\ &= |\Sigma_{11}| \cdot |\Sigma_{22.1}|. \end{aligned} \quad (3.1)$$

Hence, maximizing $|\Sigma_{11}|$ is equivalent to minimizing $|\Sigma_{22.1}|$, i.e. maximizing the retained variability represented by $|\Sigma_{11}|$, is equivalent to minimizing the lost or reduced information represented by $|\Sigma_{22.1}|$. Further McCabe (1984) showed that the maximum generalized variance criterion is equivalent to

$$\text{Minimize } \prod_{i=1}^{p-m} \lambda_i \text{ where } \lambda_1, \lambda_2, \dots, \lambda_{p-m} \text{ are eigen values of } \Sigma_{22.1}.$$

3.2 Maximum entropy criterion

We always lose information while simplify a complex model involving large number of variables into a simpler and less detailed model by discarding the unnecessary details or by aggregating these variables among themselves. We therefore, should choose the transformation $\mathbf{Y} = \mathbf{A}_m \mathbf{X}$ that results in minimum loss of information for a given degree of simplification.

Let $S(\mathbf{X})$ and $S(\mathbf{Y})$ be the entropies for the distribution of \mathbf{X} and \mathbf{Y} , respectively. Since entropy is a measure of expected information so $S(\mathbf{X}) \geq S(\mathbf{Y})$, the information loss is then defined as

$$\text{Information Loss} = S(\mathbf{X}) - S(\mathbf{Y}). \quad (3.2)$$

The Information Loss in (3.2) is greater than or equal to zero. The equality is attained if and only if \mathbf{A}_m is non-singular and in this case we get no reduction in dimensions. Thus, for reduction in dimensions \mathbf{A}_m must be singular. According to the entropy criterion we find \mathbf{A}_m for which information loss is minimum or equivalently, the entropy of $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)$ is maximum for a given value of m .

If $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{Y} = \mathbf{A}_m \mathbf{X} \sim N(\mathbf{A}_m \boldsymbol{\mu}, \mathbf{A}_m \boldsymbol{\Sigma} \mathbf{A}_m)$. It can be shown that the entropies of the distribution of \mathbf{X} and \mathbf{Y} are respectively given by

$$S(\mathbf{X}) = \frac{1}{2} \ln(|\boldsymbol{\Sigma}|) + \frac{p}{2} \ln(2\pi e)$$

and

$$S(\mathbf{Y}) = \frac{1}{2} \ln(|\mathbf{A}_m \boldsymbol{\Sigma} \mathbf{A}_m|) + \frac{m}{2} \ln(2\pi e).$$

Let, $\lambda_1, \lambda_2, \dots, \lambda_m$ be the eigen values of the matrix $\mathbf{A}_m \boldsymbol{\Sigma} \mathbf{A}_m$, then

$$|\mathbf{A}_m \boldsymbol{\Sigma} \mathbf{A}_m| = \lambda_1 \lambda_2 \dots \lambda_m$$

so that

$$\ln(|\mathbf{A}_m \boldsymbol{\Sigma} \mathbf{A}_m|) = \sum_{i=1}^m \ln(\lambda_i)$$

and

$$S(\mathbf{Y}) = \frac{1}{2} \sum_{i=1}^m \ln(\lambda_i) + \frac{m}{2} \ln(2\pi e).$$

3.3 Least generalized dependence criterion

Since m independent variates provide more information than m dependent variates, thus for defining a singular transformation $\mathbf{Y} = \mathbf{A}_m \mathbf{X}$,

we should lose only the redundant information. Consequently, we should select the most independent subset of size m which minimizes the overall dependence among the components of $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)$.

Using the concept of mutual information, a scalar measure of generalized dependence (GD) among Y_1, Y_2, \dots, Y_m is given by

$$\text{GD} = \sum_{i=1}^m S(Y_i) - S(\mathbf{Y}). \quad (3.3)$$

Where, $S(\mathbf{Y}) = - \int \dots \int f(y_1 \dots y_m) \ln f(y_1 \dots y_m) dy_1 \dots dy_m$ is the entropy of joint distribution of Y_1, Y_2, \dots, Y_m and

$S(Y_i) = - \int f_i(y_i) \ln f_i(y_i) dy_i$ is the entropy of marginal distribution of Y_i ($i = 1, 2, \dots, m$). If \mathbf{X} is a p -dimensional multivariate normal random vector with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = (\sigma_{ij})_{p \times p}$, then

$$\text{GD} = \frac{1}{2} \ln \left(\frac{\prod_{i=1}^p \sigma_{ii}}{|\boldsymbol{\Sigma}|} \right). \quad (3.4)$$

In terms of correlation matrix we have

$$\text{GD} = -\frac{1}{2} \ln(|\mathbf{R}|).$$

Thus for given m , entropy criterion is further equivalent to the selection of m original variables for which $\sum_{i=1}^m \ln(\lambda_i)$ is maximum. Where, $\lambda_1, \lambda_2, \dots, \lambda_m$ are eigen values of \mathbf{R} .

3.4 Prediction error (PRESS) criterion

In this criterion one assesses the subset of variables selected at every step of the selection procedure in the sense of prediction or model choice. This is achieved through cross validation technique. For more details on this criterion one may refer to Mori et al. (1999).

6. Illustration

For illustration purposes, we make use of a subset of the data generated by Parkash (2001) for assessing the genetic diversity in 107 accessions of cluster bean (Guar). For the present study, we purposely select the following five characters:

1. Number of pods on main shoot (x_1).

2. Number of pods on branches (x_2).
3. Total number of pods per plant ($x_3 = x_1 + x_2$).
4. Number of seeds per pod (x_4).
5. Seed yield per plant (x_5).

For selection of Principal variables the procedure developed by Mishra and Hooda (2005) have been used. Percentage of variation explained by the set of principal variables is computed by the expression given by McCabe (1984).

$$\text{Percentage of Variation Explained} = \frac{m + \sum_{j=m+1}^p R_{x_j, Y}^2}{p} \times 100$$

where m is the number of variables in the selected set, p is the total number of variables considered and $R_{x_j, Y}^2$ is the squared multiple correlation of x_j with the m -dimensional vector Y .

$$\text{Correlation Matrix} = \begin{bmatrix} 1 & 0.041 & 0.278 & 0.095 & 0.160 \\ 0.041 & 1 & 0.967 & -0.053 & 0.750 \\ 0.278 & 0.967 & 1 & -0.026 & 0.755 \\ 0.095 & -0.053 & -0.026 & 1 & -0.082 \\ 0.160 & 0.750 & 0.755 & -0.082 & 1 \end{bmatrix}.$$

Results of principal variable analysis are presented in Table 1. For $m = 4$, all the criteria are optimized for the combination (x_1, x_2, x_4, x_5) confirming that the x_3 is the redundant variable. For $m = 2$ we have (x_3, x_4) as the best subset explaining 71.96% of the total variability. The combination (x_1, x_2) also forms a competent subset of variables. The percentage of variation explained by a given number of principal variables is very near to the cumulative variation explained by the corresponding number of principal components as presented in Table 2.

Table 1
Principal variables analysis results for $m = 2(1)4$

m	Variables in	Entropy	$ \Sigma_{11} $	Dependence	Variation
2	(x_1, x_2)	2.83743	0.99831	0.00085	71.64%
2	(x_1, x_3)	2.79819	0.92295	0.04009	
2	(x_1, x_4)	2.83377	0.99101	0.00451	

(Table 1 Contd.)

m	Variables in	Entropy	$ \Sigma_{11} $	Dependence	Variation
2	(x_1, x_5)	2.82535	0.97448	0.01293	
2	(x_2, x_3)	1.46493	0.06414	1.37335	
2	(x_2, x_4)	2.83686	0.99717	0.00142	
2	(x_2, x_5)	2.4257	0.43817	0.41258	
2	(x_3, x_4)	2.83793	0.99931	0.00034	71.96%
2	(x_3, x_5)	2.41623	0.42994	0.42205	
2	(x_4, x_5)	2.83493	0.99332	0.00335	
3	(x_1, x_2, x_3)	1.81085	0.0075	2.44657	
3	(x_1, x_2, x_4)	4.25041	0.98607	0.00701	91.46%
3	(x_1, x_2, x_5)	3.82463	0.42081	0.43279	
3	(x_1, x_3, x_4)	4.2113	0.9119	0.04612	
3	(x_1, x_3, x_5)	3.79214	0.39433	0.46528	
3	(x_1, x_4, x_5)	4.2351	0.95634	0.02232	
3	(x_2, x_3, x_4)	2.87764	0.06332	1.37978	
3	(x_2, x_3, x_5)	2.45532	0.02721	1.8021	
3	(x_2, x_4, x_5)	3.84142	0.43518	0.416	
3	(x_3, x_4, x_5)	3.83054	0.42581	0.42688	
4	(x_1, x_2, x_3, x_4)	3.22351	0.0074	2.45305	
4	(x_1, x_2, x_3, x_5)	2.79795	0.00316	2.87861	
4	(x_1, x_2, x_4, x_5)	5.23405	0.4127	0.44251	99.84%
4	(x_1, x_3, x_4, x_5)	5.20147	0.38667	0.47509	
4	(x_2, x_3, x_4, x_5)	3.86459	0.02668	1.81197	

Table 2
Comparison of principal variables with principal components

Principal components	Eigen values	Explained variation (%)	Cumulative variation explained (%)	Principal variables	Variation explained
1	2.706	54.12	54.12	—	—
2	1.101	22.02	76.12	(x_3, x_4)	71.96%
3	0.882	17.64	93.78	(x_1, x_2, x_4)	91.46%
4	0.311	6.22	100	(x_1, x_2, x_4, x_5)	99.84%
5	0.000	0.000	100	$(x_1, x_2, x_3, x_4, x_5)$	100

References

- [1] M. A. Aitkin (1974), Simultaneous inference and the choice of variables in multiple regression, *Technometrics*, Vol. 16, pp. 221–227.
- [2] E. M. L. Beale, M. G. Kendall and D. W. Mann (1967), The discarding of variables in multivariate analysis, *Biometrika*, Vol. 54, pp. 357–366.
- [3] R. R. Hocking (1976), The analysis and selection of variables in linear regression, *Biometrics*, Vol. 32, pp. 1–51.
- [4] D. S. Hooda and B. K. Hooda (2001), On measurement of stochastic dependence in multivariate data, *Indian J. Pure Appl. Math.*, Vol. 32 (6), pp. 801–815.
- [5] J. T. Jolliffe (1972), Discarding variables in a principal component analysis, *Applied Statistics*, pp. 121–160.
- [6] J. N. Kapur and H. J. Kesvan (1992), *Entropy Optimization Principles with Application*, Academic Press, New York.
- [7] K. Mishra and B. K. Hooda (2005), Information theoretic procedure for selection of principal variables, *Communicated*.
- [8] T. Mori, M. Iizulka, T. Tarumi and Y. Tanaka (1999), Variable selection in “principal component analysis based on a subset of variables”, *Bulletin of the International Statistical Institute 52nd Session Contributed Papers Book 2*, pp. 333–334.
- [9] G. P. McCabe (1984), Principal variables, *Technometrics*, Vol. 26, pp. 137–144.
- [10] D. F. Morrison (1976), *Multivariate Statistical Methods*, McGraw Hill, New York, p. 68.
- [11] C. R. Rao (1973), *Linear Inference and its Applications*, John Wiley & Sons, Inc., New York.
- [12] S. Parkash (2001), *Evaluation of Phenotypic Variability in Guar*, Unpublished M.Sc. Thesis submitted to CCS Haryana Agricultural University, Hisar.
- [13] S. Watanabe (1969), *Knowing and Guessing*, John Wiley, New York.

Received July, 2005