International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN (P): 2319-3972; ISSN (E): 2319-3980
Vol. 10, Issue 1, Jan–Jun 2021; 13–20
© IASET

International Academy of Science,
Engineering and Technology
IASET Connecting Researchers; Nurturing Innovations

FEATURE SELECTION USING MATRIX CORRELATIONS AND ITS APPLICATIONS IN AGRICULTURE

B. K. Hooda[1] ***& Ekta Hooda***[2]

[1]*Professor, Department of Mathematics & Statistics, CCS Haryana Agricultural University, Hisar, India*

[2]*Research Associate, Directorate of Extension Education, CCS Haryana Agricultural University, Hisar, India*

ABSTRACT

Dimensionality reduction techniques are broadly categorized as feature extraction and feature selection. Feature extraction techniques select features in the transformed space while feature selection techniques consist of finding a subset of original features or variables that is optimal for a given criterion for adequate representation of the whole data. Principal Component Analysis (PCA) is often the most common choice for reducing dimensionality of multivariate data through feature extraction. However, dimensionality reduction using PCA does not provide a real reduction of dimensionality in terms of the original variables, since all of the original variables are used in projection to the lower dimensional. Several criteria have been proposed for selecting the best subset of features which can preserve the structure and variation of the original data. However, little is known about the applications feature selection techniques in agricultural and biological research where many measurements are taken on each individual. In the present study, applicability of matrix correlation based feature selection techniques has been examined for identification of informative and redundant features in wheat data. RV-coefficient (Robert and Escoffier, 1976) and Yanai's Generalized Coefficient of Determination (Ramsay et al., (1984) have been used to measure the similarity between two data matrices. Subsets selected using different criteria have been compared in terms of the measure of overall predictive efficiency. For identification of important features, secondary data of 67 wheat genotypes recorded for 14 characters have been used. Models built with subset of best features are expected not only to reduce the model complexity but also require less time and resources.

KEYWORDS: *Feature Selection, Feature Extraction, Dimensionality, PCA, Matrix Correlation, Agriculture*

Article History

Received: 12 Jan 2021* | *Revised: 19 Jan 2021* | *Accepted: 04 Feb 2021

INTRODUCTION

In many situations, practical as well as theoretical considerations compel us to reduce data dimensionality or select variable subsets prior to the desired analysis. The existing dimensional reduction approaches are broadly categorized as feature extraction and feature selection. Feature extraction techniques select features in the transformed space while feature selection techniques find an optimal subset of original variables which according to some given criterion adequately represent the whole data. Principal Component Analysis (PCA) is an optimal statistical tool for feature extraction in multivariate analysis. It replaces the initial set of variables by a small number of linear combinations of the original variables called principal components (PCs) that together explain most of the variation in the data. However, the dimensionality reduction via principal component analysis does not provide a real reduction of dimensionality in terms of the original variables, as all original variables are still required to define even a single PC. Thus, for better interpretation one can reduce the dimensionality of the space in terms of the original variables without disturbing the main features of the whole data set. Also, in applications where interpretable features are desired, feature/variable selection techniques are more appropriate. The important contributions to the problem of variable selection in PCA setting are due to Jolliffe (1972, 2002), McCabe (1984), Krzanowski (1987), Cadima and Jolliffe (2001). Hooda and Hooda (2006, & 2008) used Shannon's entropy and mutual information for variable selection in multivariate analysis under the assumption of normality of data. Hooda *et al.* (2017) used principal component analysis (PCA) and canonical correlation analysis techniques in an attempt towards identification of principal agricultural and socio-economic dimensions in Haryana. In the present study, applicability of matrix correlation based feature selection techniques has been examined for identification of informative and redundant features in wheat data. RV-coefficient (Robert and Escoffier, 1976) and Yanai's Generalized Coefficient of Determination (Ramsay *et al*., (1984) have been used to measure the similarity between two data matrices. Subsets selected using different criteria have been compared in terms of the measure of overall predictive efficiency. For identification of important features, secondary data of 67 wheat genotypes recorded for 14 characters have been used.

MATERIAL AND METHODS

Data

Secondary data on growth and yield characters of 67 wheat genotypes was used for selection of variable subsets according to their importance. The data was generated in an experiment conducted at research farm of the Department of Genetics and Plant Breeding CCS HAU-Hisar with 6 row/entry and row length of 6m. The detail of the genotypes and recorded variables on wheat crop is given in Table-1.

Table 1: Wheat Genotype

S No	Genotype	S No	Genotype	S No	Genotype	S No	Genotype
1	AL 1	18	AL 18	35	AL 35	52	WH 542
2	AL 2	19	AL 19	36	AL 36	53	WH 711
3	AL 3	20	AL 20	37	AL 37	54	WH 1105
4	AL 4	21	AL 21	38	AL 38	55	WH 1124
5	AL 5	22	AL 22	39	AL 39	56	UP 2338
6	AL 6	23	AL 23	40	AL 40	57	HD 2687
7	AL 7	24	AL 24	41	AL 41	58	WH 1080
8	AL 8	25	AL 25	42	AL 42	59	PBW 343
9	AL 9	26	AL 26	43	AL 43	60	DPW621-50
10	AL 10	27	AL 27	44	AL 44	61	PBW 550
11	AL 11	28	AL 28	45	AL 45	62	DBW 17
12	AL 12	29	AL 29	46	AL 46	63	HD 2967
13	AL 13	30	AL 30	47	AL 47	64	HD 2851
14	AL 14	31	AL 31	48	AL 48	65	RAJ 3765
15	AL 15	32	AL 32	49	AL 49	66	PBW 373
16	AL 16	33	AL 33	50	HD3086	67	PBW 590
17	AL 17	34	AL 34	51	WH 1025		

The observations were recorded on the following 14 characters:

- DTH: No. of Days to Heading
- PH: Plant Height (cm)
- SL: Spike Length (cm)
- SPLET: Spikelet/Spike
- TILL: No. of Tillers/ Meter
- SWT: Spike Weight (g)
- GWpS: Grain Weight (g/ spike)
- FLL: Flag Leaf Length (cm)
- FLB: Flag Leaf Breath (cm)
- FLA: Flag Leaf Area (cm^2)
- GWT: 1000 Grain Weight (g)
- BY: Biological Yield (kg /plot)
- HI: Harvest Index (%)
- GY: Grain Yield (kg/plot)

Variable Selection Method

If **A** and **B** are two nxp (non-zero) matrices, then cosine of the angle gives the correlation between them (Ramsay et al., 1984). Thus, correlation between the matrices **A** and **B** is given by

$$\cos(\mathbf{A},\ \mathbf{B}) = \frac{\langle \mathbf{A}, \mathbf{B} \rangle}{\|\mathbf{A}\| . \|\mathbf{B}\|} \tag{1}$$

Where, $\langle \mathbf{A}, \mathbf{B} \rangle$ = Trace ($\mathbf{A}^t\ \mathbf{B}$); $\|\mathbf{A}\| = \sqrt{Trace(\mathbf{A}^t\mathbf{A})}$ and $\|\mathbf{B}\| = \sqrt{Trace(\mathbf{B}^t\mathbf{B})}$ represent the inner product and the norms induced by the inner products.

In the present notations, n×p data matrix **X** is the standardized data on p characters observed on each of the n wheat genotypes. **Y**(n×q) denote an arbitrary subset of q columns of **X** and R = $\mathbf{X}^t\mathbf{X}$/n is the correlation matrix of p variables.

Rv-Coefficient

The RV-coefficient was introduced by Escoufier (1973) as a measure of similarity between squared symmetric and semi-definite matrices and as a theoretical tool to analyse multivariate techniques. Let **A**(n×g) be the PC scores of first k principal components based on the complete data set and **B**(n×k) be the scores of PCs based on a subset of k (here, g = k) variables measured on same set of n individuals. In order to compare rectangular matrices using the RV -Coefficient we first transform them into square matrices. Let **S** and **T** be two positive definite matrices of same dimensions obtained as

S = **AA**^{t}, **T** = **BB**^{t}. RV-Coefficient measures the distances between the corresponding points of these two configurations and is cos(**AA**^{t}, **BB**^{t}) and can be defined as:

$$RV(\mathbf{X},\mathbf{Y}) = \frac{Trace(\mathbf{S}.\mathbf{T})}{\sqrt{Trace(\mathbf{S}.\mathbf{S})\,.Trace(\mathbf{T}.\mathbf{T})}} = \frac{Trace(\mathbf{XX}'.\mathbf{YY}')}{\sqrt{Trace(\mathbf{XX}'.\mathbf{XX}')\,.Trace(\mathbf{YY}'.\mathbf{YY}')}} \tag{2}$$

The RV-coefficient is used as the actual measure of closeness of **X** and **Y**. The value of RV(B,**Y**) varies from 0 to 1. RV(**X**,**Y**) = 0 if and only if the two sets of variable are independent. The closer to 1 the RV(**X**,**Y**) is, more similar are the two configurations. Thus, for selection of best subset of a given size we maximize the RV-coefficient between the two configurations.

Yanai's Generalized Coefficient of Determination (GCD)

Yanai's Generalized Coefficient of Determination (Ramsay et al., 1984) measures the degree of similarity between two subspaces and is defined as the cosine of the angle between the matrices of the orthogonal projections on those subspaces. Given a data set and a subset of k of its principal components, the GCD gives a measure of similarity between the principal subspace spanned by the first k principal components and the subspace spanned by a given k-variables subset of the original variables (Cadima and Jollife, 2001). The GCD is the correlation between the matrix $\mathbf{P}_k$ of orthogonal projections on the subspace spanned by a given k-variable subset and the matrix $\mathbf{P}_g$ of orthogonal projections on the subspace spanned by the given principal components of the full data set.

$$GCD = \cos(\mathbf{P}_k, \mathbf{P}_g) = \frac{1}{k}\sum_{i=1}^{k}(R_m)_i \tag{3}$$

Where $(R_m)_i$ is the multiple correlation between the ith PC of the full data set and the k selected variables. Maximization of GCD corresponds to the selection of k variables that span a subspace that is as close as possible to the principal subspace spanned by the g principal components. The GCD for the subspaces has values between 0 (means subspaces are orthogonal) and 1 (if the two sets of PCs coincide). According to Ramsay *et al.* (1984), GCD is the average of the squared canonical correlations between two sets of variables spanning each of the subspaces.

RESULTS AND DISCUSSION

Coefficient of variation indicated that the character GWpS (Grain weight (g/spike) has maximum variability (22.94%) followed by SL (Spike length), TILL (No. of tillers/ meter) and FLL (Flag leaf length) with CV values equal to 14.72%, 14.23% and 14.03%, respectively. Since the characters are measured in different units so correlation matrix is more appropriate than the covariance matrix for selection of important variables. Correlation matrix for the 14 characters of wheat is presented below:

Table 2: Correlation Matrix

	DTH	PH	SL	SPLET	TILL	SWT	GWpS	FLL	FLB	FLA	GWT	BY	HI
PH	0.268												
SL	0.084	0.208											
SPLET	0.214	0.156	-0.031										
TILL	0.052	0.150	0.216	0.101									
SWT	-0.045	0.085	-0.217	-0.112	0.100								
GWpS	-0.092	0.054	-0.220	-0.114	0.109	0.962							
FLL	-0.077	-0.016	0.307	0.154	-0.198	-0.036	-0.026						
FLB	0.265	0.143	0.143	0.241	0.278	-0.085	-0.056	-0.235					
FLA	0.073	0.061	0.374	0.281	-0.031	-0.080	-0.053	0.838	0.330				
GWT	-0.089	-0.045	0.031	-0.348	-0.127	0.103	0.177	-0.084	-0.272	-0.227			
BY	-0.174	0.176	-0.206	-0.040	0.182	0.103	0.127	-0.041	-0.194	-0.153	0.197		
HI	-0.017	-0.171	0.139	0.046	-0.046	-0.047	-0.084	-0.093	0.043	-0.067	-0.091	-0.438	
GY	-0.141	-0.061	-0.031	0.032	0.104	0.033	0.011	-0.147	-0.080	-0.191	0.020	0.268	0.743

Correlation matrix given below indicates that except a few entries all elements are small. The characters FLL and FLA have very high positive correlation (0.838) indicating that these characters provide overlapping information. Similarly, HI has high positive correlation (0.743) with grain yield (GY). First seven principal components were retained based on the average criterion (eigenvalue >1) for PCA with correlation matrix as input. Percent variation explained and cumulative for these components is given below:

Table 3

PC	Variance	Variation (%)	Cumulative Variation (%)
1	2.66	18.98	18.98
2	2.04	14.58	33.56
3	1.84	13.17	46.73
4	1.65	11.80	58.54
5	1.35	9.66	68.19
6	1.22	8.69	76.88
7	1.01	7.18	84.06

The first 7 PCs explained 84.06% of the total variability. The first principal component explained 18.98% of variability followed by 14.58% and 13.17% variability explained by PC2 and PC3 respectively. The discarded 7 PCs explained only about 16% of the total variation. Principal component loading for the first 7 PCs are given below:

Table 4: Principal Component Loadings Matrix for First 7 PCs

Variable	Component						
	1	2	3	4	5	6	7
DTH	0.324	0.085	0.414	-0.229	-0.233	0.196	0.537
PH	0.148	0.340	0.408	-0.187	0.344	0.171	0.481
SL	0.524	0.017	-0.024	0.170	0.374	0.624	-0.086
SPLET	0.461	0.066	0.343	0.070	0.010	-0.558	0.221
TILL	0.022	0.101	0.603	-0.039	0.401	0.124	-0.486
SWT	-0.558	0.540	0.295	0.486	-0.207	0.054	0.036
GWpS	-0.568	0.566	0.264	0.476	-0.188	0.064	-0.013
FLL	0.498	0.422	-0.529	0.448	0.191	-0.115	0.093
FLB	0.439	0.034	0.605	-0.075	-0.174	0.086	-0.296
FLA	0.726	0.430	-0.176	0.394	0.085	-0.059	-0.080
GWT000	-0.457	0.057	-0.312	-0.076	0.199	0.488	0.203
BY	-0.411	0.254	0.021	-0.227	0.696	-0.353	0.031
HI	0.062	-0.720	0.186	0.629	-0.049	0.120	0.145
GY	-0.241	-0.585	0.242	0.502	0.447	-0.158	0.162

Jolliffe (1972, 1973) gave several methods for selection and discarding of variables in principal component analysis. According to Jolliffe (1972) B2 criterion, if we associate one variable to each of the first 7 PCs starting from PC1, then variables in order of their importance were found to be FLA, HI, FLB, GY, BY, SL, and DTH.

Variable subsets of various sizes selected using RV-Coefficient and GCD criterion have been presented in Table-5 and Table-6 respectively.

Table 5: Subsets of Various Sizes Selected using RV-Coefficient Criterion

Size	Rv	Rv-Square	Cardinality of Selected Variables
1	0.452	0.204	10
2	0.626	0.392	7, 10
3	0.731	0.535	7, 10, 13
4	0.792	0.628	7, 8, 9, 13
5	0.842	0.709	7, 8, 9, 12, 14
6	0.878	0.771	2, 7, 8, 9, 12, 14
7	0.909	0.826	2, 4, 7, 8, 9, 12, 14
8	0.937	0.877	2, 4, 5, 7, 8, 9, 12, 13
9	0.958	0.918	1, 2, 4, 5, 7, 8, 9, 12, 13
10	0.979	0.959	1, 2, 4, 5, 7, 8, 9, 11, 13, 14
11	0.997	0.994	1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 14
12	0.999	0.999	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12. 14
13	1.000	1.000	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14

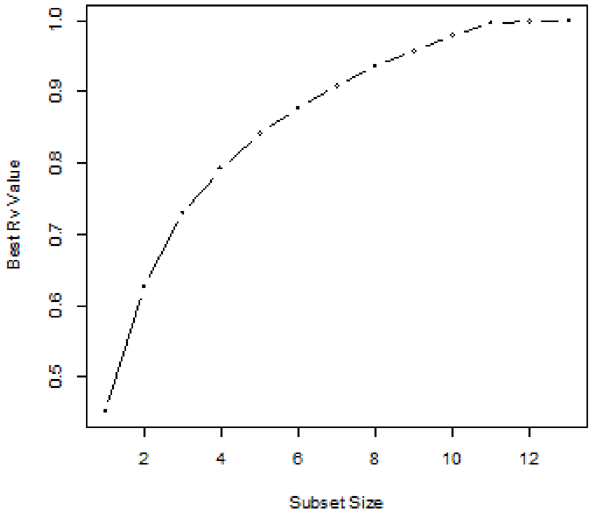

Figure 1: Plot of Best Rv Values against Subset Size.

The results in Table-5 and also the Fig-1 indicate that RV-coefficient changes slightly when the number of features is greater than 7. The best subset of 7 features selected using the RV-coefficient is {PH, SPLET, GWpS, FLL, FLB, BY, GY}. The column 3 of the Table-5 is equivalent to the proportion of total variance that is preserved if the p variables are orthogonally projected onto the subspace spanned by a given subset of k variables. Thus, selected best subset of size 7 explained 82.6% variability which very close to the variability explained by the same number of PCs based on complete data.

Variable subsets of various sizes selected using Yanai’s Generalized Coefficient of Determination (GCD) criteria have been presented in Table-6. The subsets of sizes one and two are same for both the criteria. However, subsets of sizes 3 or more have many variables in with that selected via RV-coefficient. The best subset of 7 variables selected using the

GCD criterion is {PH, SPLET, TILL, SWT, FLL, BY, HI}. The majority of the variables selected vide RV-coefficient and GCD criteria are same. In some cases substitution has taken place due the high correlation between variables (for example, GY was selected by RV-coefficient while HI by GCD criterion).

Table 3: Subsets of Various Sizes Selected using GCD-Criterion

Size	GCD	Cardinality of Selected Variables
1	0.528	10
2	0.671	7, 10
3	0.664	7, 8, 9
4	0.846	7, 8, 9, 13
5	0.833	7, 8, 9, 12, 13
6	0.850	3, 4, 7, 8, 12, 14
7	0.829	2, 4, 5, 6, 8, 12, 13
8	0.869	2, 3, 7, 8, 9, 11, 13, 14
9	0.921	1, 2, 5, 7, 8, 9, 11, 13, 14
10	0.971	1, 2, 4, 5, 7, 8, 9, 11, 13, 14
11	0.998	1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 14
12	0.999	1, 2, 3, 4, 5, 6, 8, 9, 11, 12. 14
13	1.000	1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14

From Table 5 & 6 a subset of desired number of important variables can be selected for wheat. Further, the analysis can serve as a guide for the experimenter to select a relatively more informative subset of variables or to discard the redundant variables for future studies related to this crop. The selection of more informative variables is expected to be economical on cost and time aspects in future experiments. However, final decision regarding inclusion or exclusion of any variable rests on the judgment of the experimenter and objectives of his research.

CONCLUSIONS

Importance of variables selection in has been emphasized over the dimensionality reduction through principal component analysis while interpreting research data. Variables selection based on GCD and RV- coefficient criteria have been applied for feature selection in wheat. Subsets of various sizes have been obtained by both the criteria. Best subsets of various sizes have been determined using both the criteria.

REFERENCES

1. *Cadima, J. and Jolliffe, I. T. (2001). Variable selection and the interpretation of principal subspaces. J. Agricultural, Biol. Environ. Statist. 6 (1): 62-79.*
2. *Escoufier, Y. (1973). Le traitement des variables vectorielles. Biometrics, 29: 751-760.*
3. *Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial Data. Applied Statistics, 21 : 160-173.*
4. *Jolliffe, I. T. (2002). Principal Component Analysis. Springer-Verlag, New York.*
5. *Hooda, B. K. and Hooda, D. S. (2006). Dimension reduction in multivariate analysis using maximum entropy criterion. Journal of Statistics and Management Systems, 9 (1): 175-183.*
6. *Hooda, D. S. and Hooda, B. K. (2001). On measurement of stochastic dependence in multivariate data. Indian Journal Pure Applied Mathematics, 32 (6): 801-815.*

7. *Hooda, E., Hooda, B. K. and Manocha, V. (2017). Dynamics of inter-district developmental disparities in Haryana. Journal of Applied and Natural Science, 9(2): 983-991. https://doi.org/10.31018/jans.v9i2.1307*

8. *Krzanowski, W. J. (1987). Selection of variables to preserve multivariate Data Structure, using principal components. Applied Statistics, 36: 22-33.*

9. *McCabe, G. P. (1984). Principal variables. Technometrics, 26 (2): 137-144.*

10. *Ramsay, J. O., Berge, J. and Styan, G. P. H. (1984). Matrix correlation. Psychometrika, 49 (3): 403-423.*

11. *Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. Applied Statistics, 25 (3): 257-265.*

12. *Yanai, H. (1974). Unification of various techniques of multivariate analysis by means of generalized coefficient of determination (G.C.D.), J. of Behaviormetrics, 1, 45–54.*