www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2022; SP-11(8): 07-10

www.thepharmajournal.com
Received: 07-05-2022
Accepted: 11-06-2022

Alisha Mittal
Department of Mathematics & Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

Manoj Kumar
Department of Mathematics & Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

Corresponding Author
Alisha Mittal
Department of Mathematics & Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

Multivariate statistical techniques in agriculture sciences: A review

Alisha Mittal and Manoj Kumar

Abstract
Multivariate analysis (MVA) comprises of observation and analysis of more than one variable at a time. MVA has a great application in field of agriculture like selection of appropriate variety to grow, quantity of fertilizers, pesticides and insecticides and to assess soil fertility. In the present review we had tried to explain the use of multivariate statistical methods such as principal component analysis (PCA), factor analysis (FA) and canonical correlation analysis (CCA) to explain relationships among different variables and making decisions for future works relating to the agriculture and livestock sciences.

Keywords: Principal component analysis, factor analysis, canonical correlation

Introduction
All scientific, psychological, sociological, political, economic, biology, zoology and botany make decisions on the basis of results obtained after successful analysis of data. Unless the data are summarized by some methods and suitable interpretations have been made, almost all of data in agricultural science are in bulk and by themselves they offer only little help. It is not possible to make any interpretation just by looking at the raw data, a careful scientific analysis of these data is necessary to provide enormous amount of valuable information. If the data is more complex, then more sophisticated statistical techniques are required to analyse it. It is seen that sometimes, data are collected on a number of units and on each unit many variables are measured but not just one. Further, when dealing with many variables, in order to obtain more precise and more definite information, scientists are required to use further complex analyses to get maximum information that can be acquired from data (Everitt and Dunn, 1992) [9]. In case of univariate data, when there is only one variable under our study, we usually summarize it by the population or sample mean, standard deviation, variance, moments, skewness, kurtosis etc. (Anderson, 1984) [2]. These are the basic statistics used for data description. On the other hand, multivariate statistics is a form of statistics covering the simultaneous observation and analysis of more than one statistical variable. Some methods of bivariate statistics like simple linear regression and correlation are special cases of multivariate statistics in which two variables are involved (Steel and Torrie, 1960) [31]. Multivariate statistics helps in understanding the different aims and scope of the study, and it can explain how different variables are inter-related with each other or one another. When we implement multivariate statistics to a particular problem practically, it may involve various types of univariate and multivariate analysis in order to understand the relationships among variables and their importance to the actual problems being studied (Johnson and Wicheren, 1996) [18]. There are different types of multivariate statistical techniques for analysis are available in literature such as multivariate analysis of variance (MANOVA), multivariate regression analysis, principal components analysis (PCA), factor analysis (FA), canonical correlation analysis (CCA), discriminant and clustering analysis. In present study, a detailed description of review of literature of three techniques viz. principal components analysis (PCA), factor analysis (FA), canonical correlation analysis (CCA) has been done to facilitate future work in these areas.

Principal Component Analysis (PCA)
Principal component analysis (PCA) is the oldest and one of the best-known data reduction techniques in multivariate analysis. It was first devised by Pearson (1901) [25], and later developed by Hotelling (1933) [16]. As like many multivariate statistical methods, it was not widely accepted or used due to computational challenges.

But after the advent of electronic computers, it is now well rooted in every statistical software packages. Principal Component Analysis (PCA) uses sophisticated mathematical principles to transform a number of possibly correlated variables into a smaller number of variables called principal components. The main aim of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, retaining maximum variation (as possible) present in the data set. This reduction is obtained by transforming complete data set to a new set of variables, called the principal components, which are uncorrelated, and ordered in a way that the first few contain most of the variation present in all of the original variables. PCA depends upon two things, first is the eigen value decomposition of positive semi-definite variance- covariance matrices and second is the singular value decomposition (SVD) of rectangular matrices. So, PCA is a way of recognising patterns in data, and to express the data in such a way to focus on similarities and dissimilarities. Since it is very difficult to find patterns in data of high dimensions where the facilities of presenting data graphically are not available. PCA proves to be a great tool in such cases. The roots of statistical techniques are very challenging to trace. Preisendorfer and Mobley (1988) [26] noticed that the singular value decomposition (SVD) in a form that described PCA was derived independently by Beltrami (1873) [4] and Jordan (1874) [20]. The SVD was used by Fisher and Mackenzie (1923) [10] in the background of a two-way analysis of an agricultural trial. However, the earliest explanations of the technique known as PCA were given by Pearson (1901) [25] and Hotelling (1933) [16]. The Hotelling's paper had two parts. The first is the most important, along with Pearson's paper, was among the group of papers edited by Bryant and Atchley (1975) [6]. Both the papers followed different approaches, using the standard algebraic derivation near to that introduced by Hotelling (1933) [16]. On the other hand, for better fitting of the set of points in p-dimensions, Pearson (1901) [25] was concerned with finding lines and planes. Geometric optimization problems considered by him also lead to PCs. No much relevant information seemed to be published in the 32 years between Pearson and Hotelling papers. Later, it was indicated by Rao (1964) [27] that a similar kind of approach to that of Pearson was adopted by Frisch (1929) [13]. Hotelling approach defined PCA as different in character from that of factor analysis. Another paper of Hotelling (1936) [17] highlighted a power method for obtaining PCs. Some alternative derivations for PCs were also adopted by Girshick (1936) [14]. Anderson (1963) [1] gave most theoritical information regarding sampling distributions of the coefficients and variances of PCs for large samples. Rao (1964) [27] gave a large number of ideas relating to uses, inferences and extension works of PCA. A link between PCA and other statistical techniques was discussed by Gower (1966) [15].

Factor Analysis (FA)

Factor analysis (FA) is a statistical method which is used to explain variation between observed and correlated variables in terms of a lesser number of unobserved variables called factors. According to Spearman (1904) [30], the main aim of FA is to discover simple patterns of inter-relationships among the variables. Sometimes, it can happen that variability in three or four observed variables reflect the variations in fewer such unobserved variables. FA is used to find for such joint variations inresponse to some unobserved latent variables according to Anderson (1984) [2]. The observed variables are written as linear function of the potential factors and error terms. The information obtained from inter-dependence between observed variables can be used later to reduce the set of variables in a dataset as described by Manly (2001) [24]. FA is sometimes related to principal component analysis (PCA), but the two are different approaches. Latent variable models, including factor analysis, use regression modeling techniques to test hypotheses producing error terms, while PCA is a descriptive statistical technique as shown by Dunetman (1989) [7]. Even if the independent variables are not measured directly, FA is used to study the inter- relationships among many dependent variables with the aim of finding something new about the nature of dependent variables. The first step of FA is to extract a set of factors from a data set. These factors must be orthogonal and are ordered in a way that these factors explain the proportion of the variance of the original data. Generally, a small subset of factors is kept for further consideration and the remaining factors are considered as irrelevant i.e., they are assumed to explain measurement error or noise. In order to make the interpretation of the factors having some importance, the first selection step is generally followed by a rotation of the factors that are retained. Generally, there are two main types of rotation. One is orthogonal; when the new axes are also orthogonal to each other and the other is oblique; when the new axes need not be orthogonal to each other. The part of variance explained by the total subspace after rotation is the same as it was before rotation according to Kaiser (1958) [21].

The researcher must first choose which factor model to employ in the analysis. According to Ford *et al.* (1986) [11], factor analysis comprises of two different approaches: common factor analysis and component analysis. In the component analysis model, no assumption regarding unique or error variance is made in the data. On the other hand, the common factor analysis model requires some assumption on the variance in a variable leading to division of variance into common and unique components. Rummel (1970) [28] divided the unique variance into specific and random error variance. Tucker *et al.* (1969) [34] stressed that a serious thought to the appropriate factor model should be given by the researchers while dealing with the study. When our objective is to maximize the ability to explain the variance of observed variables, the component model is more useful. Ford *et al.* (1986) [11] and Tucker *et al.* (1969) [34] showed that common factor analysis is more useful when the measured variables are a linear function of a set of latent variables. According to Kenny (1970) [23] application of component analysis when the objective is to estimate relationships among latent variables can lead to unsuitable solutions which do not contribute to fundamental theory.

Canonical Correlation Analysis (CCA)

Canonical correlation analysis is a technique for which tells the relationship between two sets of variable by finding linear combinations that are maximally correlated (Thompson, 1984) [33]. According to Bilgin *et al*. (2003) [5], canonical correlation analysis can be used to deal with two variable sets simultaneously and to yield both structural and spatial meaning. Formally, dependence refers to any position in which random variables do not fulfil a mathematical condition of probabilistic independence (Steel and Torrie, 1960) [31]. The applications of canonical correlation analysis such as

estimation of the relationship between some traits measured pre and post slaughtering, production performance and body measurements, milk and reproductive traits, , milk and wool yield traits, head and scrotum measurements, testicular and body measurements or measurements in different periods (sucking and fattening) and body measurements etc. were discussed in the previous livestock studies (Al-Kandari and Jolliffe, 1997; Fourie *et al.*, 2002; Tatar and Elicin, 2002; Emsen and Davis, 2004) [3, 12, 32, 8].

Canonical variables which in the study by Sahin (2011) [29] were needed to represent the association between morphologic traits measured at weaning and at the six-month age from 72 lambs of merino, were so formed that the first pair has the largest correlation of any linear combination of the original variables. Subsequent pairs also have maximized correlations subject to the constraint that they are uncorrelated with each previous pair (Johnson and Wichern, 2002) [19]. Kaldor (1967) [22] has emphasized that the industrial growth leads to the overall growth. A positive correlation between the rates of growth of GDP and the rates of growth of manufacturing output was founded by him in his study of 12 industrially advanced countries during the period 1953-54 to 1963-64. It was also observed that the rates of economic growth are almost unvaryingly related to the fast rate of growth of the secondary sector like manufacturing.

Conclusions

Most of the observable phenomenons in the agricultural sciences are of a multivariate nature. In medicine recorded observations of subjects in different locations are the basis of consistent diagnoses and medication. This review was to present multivariate data analysis in a way that is understandable by non-mathematicians and consultants who are challenged by statistical data analysis. It discussed on three multivariate statistical techniques, PCA, FA and CCA. A detailed description of review of literature for these techniques is also described to provide an aid for future work in these areas. It was found that there are many motivating applications of these techniques which we are using in our day today life knowingly or unknowingly. A better understanding of these concepts among agricultural educators and better methodology when using these techniques will improve agricultural education research.

References

1. Anderson TW. Asymptotic theory for principal component analysis. Ann. Math. Statist. 1963;34:122-148.
2. Anderson TW. An introduction to multivariate statistical analysis. John Wiley, New York, 1984.
3. Al-Kandari N, Jolliffe IT. Variable selection and interpretation in canonical correlation analysis. Commun Stat-Simul. 1997;C26:873-900.
4. Beltrami E. Sulle funzioni bilineari. Giornale di Mathematiche di Battaglini. 1873;11:98-106.
5. Bilgin OC, Emsen E, Davis ME. An application of canonical correlation analysis to relationships between the head and scrotum measurements in Awassi fat tailed lambs. J. Anim. and Vet. Adv. 2003;2(6):343-349.
6. Bryant EH, Atchley WR. Multivariate Statistical Methods: Within Group Covariation. Halsted Press, Stroudsberg, 1975.
7. Dunetman GH. Principal component analysis. Sage Publication, Newbury Park, 1989.
8. Emsen E, Davis ME. Canonical correlation analyses of testicular and body measurements of awassi ram lambs. J Anim. Vet. Adv. 2004;3:842-845.
9. Everitt BS, Dunn G. Applied Multivariate Data Analysis, Oxford University Press, New York, NY, 1992.
10. Fisher RA, Mackenzie WA. Studies in crop variation II. The manurial response of different potato varieties. J. Agri. Sci. 1923;13:311-320.
11. Ford JK, Mac Callum RC, Tait M. The application of exploratory factor analysis in applied psychology: A critical review and analysis. Personnel Psychology. 1986;B(2):291-314.
12. Fourie PJ, Neser FWC, Oliver JJ, Van der Westhuizen C. Relationship between production performance, Visual Appraisal and Body Measurements of young Dorper Rams.S. Afr. J Anim. Sci. 2002;32:256-262.
13. Frisch R. Correlation and scatter in statistical variables. Nordic Statist. J. 1929;8:36-102.
14. Girshick MA. Principal components. J Amer. Statist. Assoc. 1936;31:519-528.
15. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966:53:325-338.
16. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ. Psychol. 1933;25:417-441.
17. Hotelling H. Simplified calculation of principal components. Psychometrika. 1936;1:27-35.
18. Johnson RA, Wicheren DW. Applied multivariate statistical analysis. Prentice Hall of India, New Delhi, 1996.
19. Johnson RA, Wichern DW. Applied multivariate statistical analysis, 5th ed. Prentice Hill, 2002, 767.
20. Jordan C. Memoire´ sur les forbes billionaires. J Math. Pure. Appl. 1874;19:35-54.
21. Kaiser HF. The varimax criterion for analytic notation in factor analysis. Psychometricka. 1958;23:187-200.
22. Kaldor N. Strategic Factors in Economic Development. New York State School of Industrial and Labour Relations, Cornel I University, Ithaca, 1967, 7-9.
23. Kenny DA. Correlation and causality. New York: Wiley, 1970.
24. Manly BFJ. Statitics for environmental science and management. Chapman and hall/CRC, Boca Raton, 2001.
25. Pearson K. On lines and planes of closest fit to systems of points in space. Philos. Mag. A. 1901;6:559-572.
26. Preisendorfer RW, Mobley CD. Principal Component Analysis in Meteorology and Oceanography. Elsevier, Amsterdam, 1988.
27. Rao CR. The use and interpretation of principal component analysis in applied research. Sankhya A. 1964;26:329-358.
28. Rummel RJ. Applied factor analysis. Evanston, IL: Northwestern University Press, 1970.
29. Sahin M, Cankaya S, Ceyhan A. Canonical correlation analysis for estimation of relationships between some traits measured at weaning time and six-month age in merino lambs. Bulgarian Journal of Agricultural Science. 2011;17(5):680-686.
30. Spearman C. General intelligence, objectively determined and measured. American Journal of Psychology. 1904;15:20l-293.
31. Steel RGD, Torrie JH. Principles and Procedures of Statistics. McGraw Hill Book Co. Inc., New York, 1960.

32. Tatar AM, Elicin A. Ile de France x Akkaraman (G1) Crossbred Male Lambs, Feeding Period of Dairy Intake and Live Weight and Body Size of the relationship between Canonical Correlation Method Investigation. Journal of Agricultural Sciences. 2002;8:67-72.
33. Thompson B. Canonical correlation analysis: uses and interpretation. Sage Publ., CA, USA, 1984, 69.
34. Tucker LR, Koopman RF, Linn RL. Evaluation of factor analytic research procedures by means of correlation matrices. Psvchometrika. 1969;34(4):42l-459.