Nigerian J. Anim. Sci. 2019, 21 (2): 1-8

Application of Principal Component Analysis (PCA) for correcting multicollinearity and dimension reduction of morphological parameters in Bunaji Cows

[1]Alphonsus, C and [2]Raji, A.O.

[1]*Dept of Animal Science, Faculty of Agriculture, Kaduna State University, Kafanchan campus.*
[2]*Dept of Animal Science, Faculty of Agriculture, University of Maiduguri, Borno state*

Cyprian.alphonsus@kasu.edu.ng

Target Audience: Researchers, Animal Geneticist and Breeders

Abstract

This paper presents the application of Principal Component Analysis (PCA) on the dimension reduction of morphological variables. Sixteen morphological variables were measured from 50 multiparous Bunaji cows. The correlation amongst most of the morphological variables was very high suggesting severe multicollinearity. Therefore, PCA was applied to verify whether the collinear variables could be combined to form composite scores. The application of the PCA effectively reduced the dimensionality of the 16 morphological variables into four artificial composite variables (called principal components) which were uncorrelated and independent of each other with standardized means of zero and standard deviation of one and explained 90.45% of the variation in the original morphological data set. Therefore, PCA can be used to correct the problem of multicollinearity and dimension reduction of morphological data in multiple regression analysis.

Keywords: principal component, correlation, communality, body indices, orthogonal varimax

Description of Problem

Emphasis have shifted over the years from subjective method of appraising cattle to more objective method like the use of linear body measurement of different body parts of the animal. The linear body measurement can be taken at a relatively lower cost with high relative accuracy and consistency (1), and they have moderate to high heritability (2, 3). Since these conformation traits have genetic component they could be used as correlated traits in predicting the direct and correlated responses due to selection. In animals, one trait is often associated with other traits, it may therefore be necessary to consider more than one trait for selection and improvement at a time (4). However, one of the limitations in applying multiple regression analysis to the morphological data is that of the multicollinearity (3, 4). Multicollinearity is simply a high degree of correlation among predictive variables in multiple regressions (5). A high degree of multicollinearity amongst predictive variables increases the variance in estimation of the regression coefficients (6) and compromise the basic assumption of multiple regressions which states that 'the predictors are uncorrelated and independent of each other'. When predictors suffer from multicollinearity, using multiple regression leads to inflation of regression coefficients thereby compromising the integrity and reliability of the resultant models. These coefficients could fluctuate in signs and/ or magnitude as a result of slight change in one variable (7, 8).

One of the ways of solving the problem of multicollinearity is the application of Principal Component Analysis (PCA). Principal Component Analysis is a traditional multivariate statistical method commonly used to reduce the number of predictive variables and solve the multicollinearity problem (9). The PCA aims at explaining as much of the variation in the data by finding linear combinations that are independent of each other without losing too much information in the process.

Therefore, the aim of this study was to determine whether Principal Component Analysis can be apply to solve the problem of multicollinearity and dimension reduction of morphological variables of Bunaji cows.

Materials and Methods

Data collection:

The data used for this study were collected from 50 multiparous Bunaji cows at the dairy herd of National Animal Production Research Institute (NAPRI) Shika, Kaduna state, Nigeria, located between latitude 11^0 and 12^0N at an altitude of 640 m above sea level, and lies within the Northern Guinea Savannah Zone (10). Eight morphological traits comprises of stature (ST), chest width (CW), withers height (WH), heart girth (HG), body length (BL), body depth (BD) rump width (RW) and body weight (BW) were measured. Cows were housed in tie stalls, and standard position of the cow was defined to take measurements. The morphological traits were measured in centimeter (cm) using graduated measuring stick and flexible meter tape, while the body weight was measured using weighbridge. The eight original morphological variables were used to calculate the other eight body indices as shown in Table 1. The details of the measurements and definition of the traits are also presented in Table 1. Each cow was measured 3 times for the complete lactation length; the frequency of the measurements was early-, mid- and late lactation, commencing one week post- partum.

Statistical Analysis

The correlation matrix of all the morphological parameters and their indices was first determined using PROC CORR procedure of SAS (15) to determine the level of the multicollinearity among the morphological variables.

Because of the large correlations between most of the morphological variables, principal component analysis was applied. Principal component analysis is a method for transforming the variables in a multivariate data set X_1, X_2,.X_n, into new variables, Y_1, Y_2,…..Y_n, which are uncorrelated and account for decreasing proportions of the total variance of the original variables, defined as follows;

$Y_1 = P_{11}X_1 + P_{12}X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + P_{1n}X_n$

$Y_2 = P_{21}X_1 + P_{22}X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + P_{2n}X_n$

$Y_3 = P_{n1}X_1 + P_{n2}X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + P_{nn}X_n$

With the coefficient being chosen so that Y_1, Y_2, ……..Y_n account for decreasing proportion of the total variance of the original variables X_1, X_2 …..X_n (16).The principal component analysis was run using PROC Factor SAS software (SAS, 15)

Results and Discussion

The first step in applying Principal Components Analysis (PCA) to a multiple regression data is to determine the correlation matrix of the predictive variables, as this will suggest whether there is multicollinearity problem amongst the predictors. In the present study, the correlation matrix showed high degree of correlations amongst the morphological variables (Table 2), hence an indication of multicollinearity (16, 17). Multicollinearity is a serious problem in multiple regression analysis because it violates the basic assumption of regression that requires the predictors to be independent and uncorrelated with each other’s (18, 19). It also compromises the integrity and reliability of the regression models (20).

Table 1: Details of measurements of morphological traits and calculation of body indices

	Measurements	Abbrev	Description	Instrument
Original morphological measurements (adopted from Fisher, 11 and IHFA, 12)				
1	Stature	ST	Measured from top of the spine in between hips to ground	Measuring stick
2	Height-at-withers	HW	Highest point over the scapulae vertically to the ground or measured from the highest point on the dorsum of the animal to the ground surface at the level of front legs	Measuring stick
3	Heart Girth	HG	Measured as a circumference of the body at a point immediately behind the fore legs, perpendicular to the body axis	Flexible tape
4	Chest width	CW	Measured from the inside surface between the top of the front legs.	Flexible tape
5	Body depth	BD	Distance between the top of spine and bottom of barrel at last rib, the deepest point independent of stature.	Flexible tape
6	Body length	BL	Measured from the point of shoulder to the ischium	Flexible tape
7.	Rump width	RW	The distance between the most posterior point of pin bones	Flexible tape
8	Body weight	BW	Live weight of the animal	Weigh bridge
Body indices and their mode of calculations (Alderson, 13; Sarma, 14).				
1	Height slope	HS	Withers height - status	Calculated
2	Width slope	WS	Rumps width/ chest width	Calculated
3.	Length index	LI	Body length / withers height	Calculated
4	Depth index*	DI	Body depth/withers height	Calculated
5	Foreleg length*	FL	Withers height- body depth	Calculated
6	Body index	BI	(Body length/heart girth)x 100	Calculated
7	Height index	HI	Withers height/body length	Calculated
8	Weight index	WI	Body weight x withers height	Calculated

*= in the original formula chest depth was used instead of body depth

Therefore, the morphological data were subjected to principal component analysis using ‘one’ as a prior communality estimates. The number of the principal components (PCs) retained for varimax orthogonal rotation was determined using eigenvalue criteria of one and the cumulative percentage of variance explained by the PCs retained (21). Using these criteria it was obvious that the first four PCs displayed eigenvalues equal to or greater than one, and explained over 90% of the variation in the morphological data set. This suggested that the morphological variables can be reduced into four composite variables (Principal components) without losing much of the information in the original data set. Therefore, the four PCs were retained for interpretation. The morphological variables and the corresponding loadings are presented in Table 3. In the interpretation of the rotated factor-loading pattern, a parameter was said to load heavily on a given PC if the factor loading was greater or equal to 0.60. There was a clear grouping of the morphological parameters evident by the loading pattern of the parameters on the PCs. Most of the original morphological parameters loaded heavily on the first PC, which was subsequently labeled as ‘body size measures’. Also, most of the index values loaded heavily on the second PC and were labeled ‘body indices’. Other parameters like chest width (CW), body length (BL) and weigh slope (WS) loaded heavily on the third PC, which were labeled as ‘body balance measures’. Lastly, height slope (HS) which could be term as measure of ‘ascendency’ was the only parameter that heavily loaded on the fourth PC, suggesting that HS is not strongly correlated with any of the morphological parameter measured (Table 1) and could therefore be treated as independent variable in subsequent multiple regression analysis (19).

Table 2: Pearson's correlation of morphological variables and their indices

Variables	BW	ST	CW	BD	HW	HG	BL	RW	HS	WS	DI	FL	BI	HI
Body weight (BW)	-													
Status (ST)	**0.841**	-												
Chest width (CW)	**0.678**	**0.643**	-											
Body depth (BD)	**0.790**	**0.731**	0.299	-										
Withers Height (WH)	**0.876**	**0.989**	**0.657**	**0.736**	-									
Heart girth (HG)	**0.926**	**0.790**	**0.624**	**0.719**	**0.833**	-								
Body length (BL)	**0.659**	**0.536**	**0.424**	**0.598**	**0.524**	**0.634**	-							
Rump width (RW)	**0.666**	**0.596**	**0.641**	0.352	**0.639**	**0.633**	0.210	-						
Height slope (HS)	0.325	0.030	0.157	0.107	0.177	0.371	-0.025	0.359	-					
Width slope (WS)	-0.253	-0.279	**-0.681**	-0.062	-0.256	-0.218	-0.362	0.122	0.129	-				
Depth index (DI)	-0.084	-0.316	**-0.469**	0.399	-0.327	-0.124	0.124	-0.372	-0.103	-0.261	-			
Foreleg length (FL)	0.328	**0.564**	**0.600**	-0.131	**0.575**	0.350	0.045	**0.511**	0.129	-0.301	**-0.961**	-		
Body index (BI)	**-0.778**	**-0.672**	**-0.522**	**-0.563**	**-0.735**	**-0.889**	-0.214	**-0.654**	**-0.496**	0.0062	0.211	-0.397	-	
Height index (HI)	**0.526**	**0.750**	**0.445**	**0.412**	**0.773**	**0.434**	-0.135	**0.585**	0.232	-0.027	-0.473	**0.635**	**-0.691**	-
Weight index (WI)	**-0.995**	**0.889**	**0.694**	0.107	**0.918**	**0.928**	**0.648**	**0.682**	0.294	**-0.590**	-0.259	0.386	**-0.784**	**0.583**

Since the PCs were labeled according to the sizes of their variances, the first Principal component (PC_1) explained the largest amount of variation (54.61 %) in the morphological data set, while the subsequent principal components PC_2, PC_3 and PC_4 accounted for decreasing proportion 16.88%, 11.68% and 7.26%, respectively of the original variables. Also the eigenvalues of the PCs followed similar trend with that of the percentage of variance explained by each PC. This agreed with the earlier findings on the applications of principal component analysis (16, 18, 19)

The communality estimates (h) which is the percentage of variance explained in each of the original morphological variables explained by the extracted PC was very high ranging from 69.41% to 98.22%.

The PCs displayed varying levels of correlations with the morphological parameters similar to the loading pattern of the morphological parameter on the PCs (Table 4). Thus, confirming the loading pattern of the Principal Component Analysis. However, the correlations amongst the PCs were zero. This shows that the PCA successfully transform the 16 morphological variables into four artificial composite variables which were uncorrelated and independent of each other with standardized means of zero and standard deviation of one (Table 5). This indicated that the PCA completely removed the multi-collinearity amongst the PCs and could therefore be used with high degree of reliability in multi-regression analysis (20, 21).

Table 3: relationship among morphological measures express as loadings in Principal Component Analysis

Items	Principal Components (PCs)				
	PC_1	PC_2	PC_3	PC_4	h
Cumulative variance (%)	54.61	71.49	83.17	90.45	
Body weight	**0.87**	0.06	0.33	0.30	96.51
Stature	**0.90**	0.36	0.21	0.09	98.22
Chest width	0.44	-0.49	**0.63**	0.20	86.13
Body depth	**0.93**	0.33	0.06	-0.02	97.96
Height-at-withers	**0.90**	0.36	0.19	0.05	97.63
Heart girth	**0.82**	0.08	0.31	0.39	93.46
Body length	0.51	-0.31	**0.71**	0.03	85.76
Rump width	**0.65**	0.39	0.01	0.48	68.73
Height slope	0.11	0.06	-0.10	**0.90**	84.27
Width slope	-0.05	-0.26	**0.80**	-0.04	96.45
Length index	**-0.68**	**-0.63**	0.32	0.20	76.13
Depth index	0.08	**0.94**	-0.17	-0.10	93.83
Foreleg length	0.20	**0.92**	0.20	0.10	94.05
Body index	**-0.73**	-026	-0.03	-0.49	85.26
Height index	**0.67**	**-0.65**	-0.32	0.04	96.49
Weight index	**0.89**	0.13	0.32	0.26	97.94
Individual Variance explain (%)	54.61	16.88	11.68	07.26	
Eigen values	8.74	2.70	1.87	1.16	

h=communality estimates

Table 4: Pearson's Correlations between principal components (PCs) and morphological parameters

Morphological variables	Principal components (PCs)			
	PC_1	PC_2	PC_3	PC_4
Body weight	0.870**	0.065	0.334	0.304
Stature	0.898**	0.355	0.207	-0.086
Chest width	0.438	0.489	0.626	0.197
Body depth	0.929**	-0.330	0.062	-0.022
Height-at-withers	0.899**	0.358	0.188	0.048
Heart girth	0.818**	0.076	0.312	0.395
Body length	0.505	-0.309	0.714**	0.026
Rump width	0.555	0.393	0.008	0.481
Height slope	0.108	0.056	-0103	0.901**
Width slope	-0.049	-0.261	-0.804**	0.201
Length index	-0.679**	-0.635**	0.316	-0.036
Depth index	0.079	-0.944**	-0.168	-0.103
Foreleg length	0.195	0.923**	0.201	0.096
Body index	-0.733**	-0.258	0.029	-0.488
Height index	0.669**	0.646**	-0.316	0.038
Weight index	0.893**	0.126	0.316	0.257
PC1	1.000	0.000	0.000	0.000
PC2	0.000	1.000	0.000	0.000
PC3	0.000	0.000	1.000	0.000
PC4	0.000	0.000	0.000	1.000

*=P<0.05; **=P<0.01

Table 5: descriptive statistics of the Principal Components

Principal components (PCs)	N	Means	SD	Minimum	Maximum
PC_1	40	0.00	1.00	-1.257	1.759
PC_2	40	0.00	1.00	-2.244	1.663
PC_3	40	0.00	1.00	-2.472	1.759
PC_4	40	0.00	1.00	-1.069	3.092

N= number of animals; SD= standard deviation

Conclusion and Applications

1. The application of Principal Component Analysis (PCA) to the morphological data effectively reduced the dimensionality of the 16 morphological variables into four artificial composite variables (called principal components) which are uncorrelated and independent of each other with standardized means of zero and standard deviation of one and explained 90.45% of the variation in the original morphological data set.
2. Therefore, PCA can be used to correct the problem of multicollinearity and dimension reduction of morphological data in multiple regression analysis.

References

1. Essien, A. and O.M. Adesope, O. M.(2003). Linear body measurements of N'dama calves at 12 month in South Western zone of Nigeria. *Livestock Research for Rural Development,* 15:4-9,

2. Kadarmideen H .N and S, Wegmann, (2003).Genetic parameter for body condition score and its relationship with type and production traits in Swiss Holstein. Journal of Dairy Science 86:3685 – 3693
3. De Haas Y, Janss, L.K.G and H.N. Kadarmideen, (2007).Genetic and phenotypic parameters for conformation and yield traits in three dairy cattle breeds. *Journal of Animal Breeding and Genetics, 124 (1):12-19,*
4. Alphonsus, C., Akpa, G.N., Mukasa, C., Rekwot, P.I and P. P Barje, (2011). Genetic Evaluation of Linear udder and Body conformation Traits in Bunaji cows. *Animal Research International.* 8(1): 1366 – 1374www.zoo-unn.org
5. Klainbaum D. G,Kupper LL and K.E. Muller (1998)Applied Regression Analysis and Multivariable Methods. 3rd Edition (Colle Pacific Grove, CA).
6. Yu CH (2008).Multi-collinearity, variance inflation, and orthogonalization in regression. Web link: http://www.creative-wisdom. com/ computer/ sas/collinear.html
7. Leahy K. (2001). Multicollinearity: When the solution is the problem. In Olivia Parr Rud (Ed.) Data Mining Cookbook (pp. 106 - 108). New York: John Wiley & Sons, Inc.
8. Fekedulegn DB, Colbert JJ, Hicks Jr RR and M.E Schuckers (2002).Coping with multicollinearity: An example on application of Principal Components Regression in Dendroecology. Research Paper NE-721, Newton Square PA: United States Department of Agriculture. Forest service. 1-48
   Web link: www. fs.fed. us/ne/ morgantown/4557/dendrochron/rpne721. pdf
9. Bair Eric, Trevor Hastie, Paul Debashis and Robert Tibshirani (2006).Prediction by supervised Principal Components. *Journal of the American Statistical Association*. 473 (19): 119-137
10. Oni OO, Adeyinka IA, Afolayan RA, Nwagu BI Malau-Aduli AEO,Alawa CBI and O.S Lamidi (2001). Relationships between milk yield, post partum body weight and reproductive performance in Friesian x Bunaji Cattle. Asian–Australian Journal of Animal Science. 14(11): 1505 – 1654.
11. Fisher, A.V(1975).Criteria and methods for assessment of carcass and meat characteristics in beef production experiment, EE seminar, Zeist. Pp.43-58
12. IHFA (2006).Irish Holstein Friesian Association; Type Classification Scheme.http://www.ifha.ie/bestofbreed/typeclassificationscheme.htm
13. Alderson G.L.H (1999). The development of system of linear measurements to provide an assessment of type and function of beef cattle. AGRI 25:45-55.
14. Sarma K (2006). Morphological and Craniometrical Studies on the Skull of Kagani Goat (*Capra hircus*) of Jammu Region. Int. J. Morphol. 24(3):449-455, 2006
15. SAS (2006).SAS User‟s Guide Version 8.1. Statistical Analysis system institute Inc, Cary, Nc, USA
16. Lafi, S.Q and J.B Kaneene, (1992). An explanation of the use of principal component analysis to detect and correct for multicollinearity. *Preventive Veterinary Medecine* 13: 261-275.
17. Vaughan TS and Berry KE. (2005). Using Monte Carlo techniques to demonstrate the meaning and implications of multicollinearity. Journal of Statistics Education, 13(1): 1-9. Web link: www.amstat.org/publications/jse/v13n1/vaughan.html

18. Yu CH. (2010). Checking assumptions in regression. Web link: http://www.creativewisdom.com/computer/sas/regression_assumption.html
19. Alphonsus C, Akpa GN, Nwagu BI, Abdullahi I, Zanna M, Ayigun AE, Opoola E, Anos KU, OlaiyaO and O. I Olayinka-Babawale (2014).Application of multivariate principal component analysis on dimensional reduction of milk composition variables. *Journal of Research in Biology* 4(8): 1526-1533http://jresearchbiology.com/documents/RA0489.pdf
20. Maitra S and Yan J. (2008). Principal Component Analysis and Partial Least Squares: two dimension reduction techniques for regression. Casualty Actuarial Society, Discussion paper program. pp.79-90
21. Principal Component Analysis http://support.sas.com/publishing/publicat/chaps/55