Vol. 8 January 2017
CHED Accredited Research Journal, Category A-1
Print ISSN 2094-5019 • Online ISSN 2244-0461
doi: http://dx.doi.org/10.7828/ajob.v8i1.999

Asian Journal of Biodiversity
This Journal is in the Science Master Journal List of
Thomson Reuters (ISI) Zoological Record

Euclidean Distance Index Based on Few Orthogonal Principal Components for Biological Taxonomic Purposes

VINCENT T. LAPINIG
ORCID No. 0000-0003-1838-5300
vinceginipal82@gmail.com
Northwestern Mindanao State College of Science and Technology
Tangub City

ROBERTO N. PADUA
ORCID No. 0000-0002-2054-0835
rnpadua@yahoo.com
Liceo de Cagayan University
Cagayan de Oro City

ABSTRACT

The paper proposed an alternative measure for classification of biological organisms in taxonomy. The alternative measure is based on a normalized Euclidian distance index that is derived from a few orthogonal principal components of the input trait matrix. Using an exploratory, descriptive method of research, data from two (2) species of the biting fly were analyzed. Result revealed that the alternative measure successfully distinguished between the two (2) species of biting fly *L. torrens* and *L. carteri* using only the wing dimensions of the flies. The other information such as palp length, palp width, and length of antennal segment were not useful in discriminating between two (2) fly species. The implication is that for taxonomic purposes, only a few traits are sufficient to distinguish between organisms coming from the same genera.

Keywords: survey, threatened, endemic plants, Mindanao, Philippines

INTRODUCTION

In taxonomy, morphological traits of biological organisms are often used as initial bases for classifying organisms into their "closest" species (Brown 2000). Let X_j be a p-variate vector of morphological traits (e.g., body mass, length of tail, etc.) for j=1,2,…,n individuals. The vectors $\{X_j\}$ are used as inputs to a cluster analysis algorithm with k = m suspected cluster (species). The species to which the individual with traits X_j belong is identified based on the dendrogram of the cluster analysis output (Johnson and Wichern 2007).

The vector X_j need not contain only morphological characteristics of the organisms but may also include physiological traits (metabolic rates) and ecological traits (speed) (Cheverud 1982). Static allometry finds relationships that may exist among the p –traits contained in the vector X_j at the same developmental stage for the same species or the same population (genera). In Lapinig et al. (2016), it is demonstrated that if an allometric relationship exists among the p traits, then the first principal component of the covariance structure of X_j accounts for more than 90% of the total variance. This being the case, the authors suggested using with the first principal component as the single numerical index $\{I_j\}$ as the discriminant feature for purposes of classification.

The use of a single numerical index $\{I_j\}$ for classification purposes has the obvious advantage of being simpler to use and interpret. In fact, Lapinig et al. (2016), derived the simple linear discriminant rule:

$$\text{Allocate } X_j \text{ to } \pi_2 \text{ if } I_j > \frac{S_1\bar{x}_2+S_1\bar{x}_1}{S_1+S_2} \qquad (1)$$

where $\bar{x}_1, \bar{x}_2$ are the sample means and S_1, S_2 are the sample standard deviations.

The lynchpin for the successful use of (1) is the fact that the p – traits are linearly related through the static allometric equation. The researchers pose a more general problem: Suppose that subsets of the p traits are allometrically related but no linear relationship exists among all the p traits, how may a single numerical index be formulated? The obvious consequence of this situation is that the first principal component alone cannot sufficiently explain the total system variance. This means that the first $m < p$ principal component will have to be used in the construction of the allometric index.

OBJECTIVES OF THE STUDY

This study aimed to propose an alternative measure for classification of biological organisms in taxonomy.

MATERIALS AND METHODS

Let X be a p – variate random vector of traits with mean μ and covariance matrix Σ. Principal Components Analysis (PCA) seeks to find linear combinations $Y=a^TX$ with maximum variance. More specifically,

$$Max\ var(Y) = var\ (a^TX) = a^T\Sigma a \quad (2)$$
$$a \neq \theta$$
$$\text{Subject To: } a^Ta \leq 1.$$

It is known that the solution **a** corresponds to the eigenvectors **e** of Σ**.** Since Σ is positive – definite, there is an orthogonal matrix P such that:

$$\Sigma = P^T\mathrm{DP} \quad (3)$$

where $D=diag(\lambda_i)$ is a diagonal matrix of the eigenvalues of Σ and $P=[e_1:e_2:\ldots:e_p]$is the matrix whose columns are the corresponding eigenvectors. The eigenvectors $\{e_i\}$ are orthogonal:

$$\langle e_i/e_j\rangle = 0 \quad (4)$$

Since $P^T\ P=PP^T=I$. The ith principal component is given by:

$$Y_i = e_i^TX \quad (5)$$

and:

$$var(Y_i) = e_i^T\Sigma e_i = e_i^Te_i\lambda_i = \lambda_i \quad (6)$$

The proportion of the total variance explained by the ith principal component is provided by:

$$Proportion\ of\ Variance\ Explained\ by\ Y_i = \frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p} \quad (7)$$

In Lapinig et al. (2016), an allometric relationship exists among all the p component of X, hence:

$$Proportion\ of\ the\ total\ variance\ Explained\ by\ Y_i = \frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p} > 90\% \quad (8)$$

In the event that only subsets of the p traits are allometrically related, then the first m – principal component, $m < p$, will be needed:

$$\begin{matrix}Cumulative\ Proportion\ of\ the\ total\\ variance\ explained\ by\ the\ first\ m\\ principal\ components\end{matrix} = \frac{\sum_{j=1}^{m}\lambda_j}{\sum_{i=1}^{p}\lambda_i} > 90\% \quad (9)$$

Note, however, that the principal component themselves are orthogonal:

$$\begin{aligned}\langle Y_i/Y_j\rangle &= Cov(e_i^T \mathrm{X}, e_j^T \mathrm{X}) \\ &= e_i^T \Sigma e_j \\ &= e_i^T(\lambda_j e_j) \\ &= \left(e_i^T \cdot e_j\right)\lambda_j = 0 \end{aligned} \tag{10}$$

The orthogonality of the first m principal components suggest the following one – dimensional index:

$$d_{j,k}^2 = Y_{1,k}^2 + Y_{2,k}^2 + \cdots + Y_{m,k}^2 \quad ,k = 1, 2, \ldots, n \tag{11}$$

Or:

$$d_{j,k} = \sqrt{Y_{1,k}^2 + Y_{2,k}^2 + \cdots + Y_{m,k}^2}, \; k = 1, 2, \ldots, n \tag{12}$$

To account for the variability across principal component, the researchers modify (12) into:

$$d_{j,k} = \sqrt{\frac{Y_{1,k}^2}{\lambda_1} + \frac{Y_{1,k}^2}{\lambda_2} + \cdots + \frac{Y_{1,k}^2}{\lambda_m}} \;, k = 1, 2, \ldots, n \tag{13}$$

Suppose next that there are two (2) populations π_1 and π_2 for which the indices (13) are computed:

$$\pi_1 = \left\{d_{j,k}^{(1)} : k = 1, 2, \ldots, n_1\right\} \text{ with } E\left(d_{j,k}^{(1)}\right) = \mu_1$$

$$\pi_2 = \left\{d_{j,k}^{(2)} : k = 1, 2, \ldots, n_1\right\} \text{ with } E\left(d_{j,k}^{(2)}\right) = \mu_2,$$

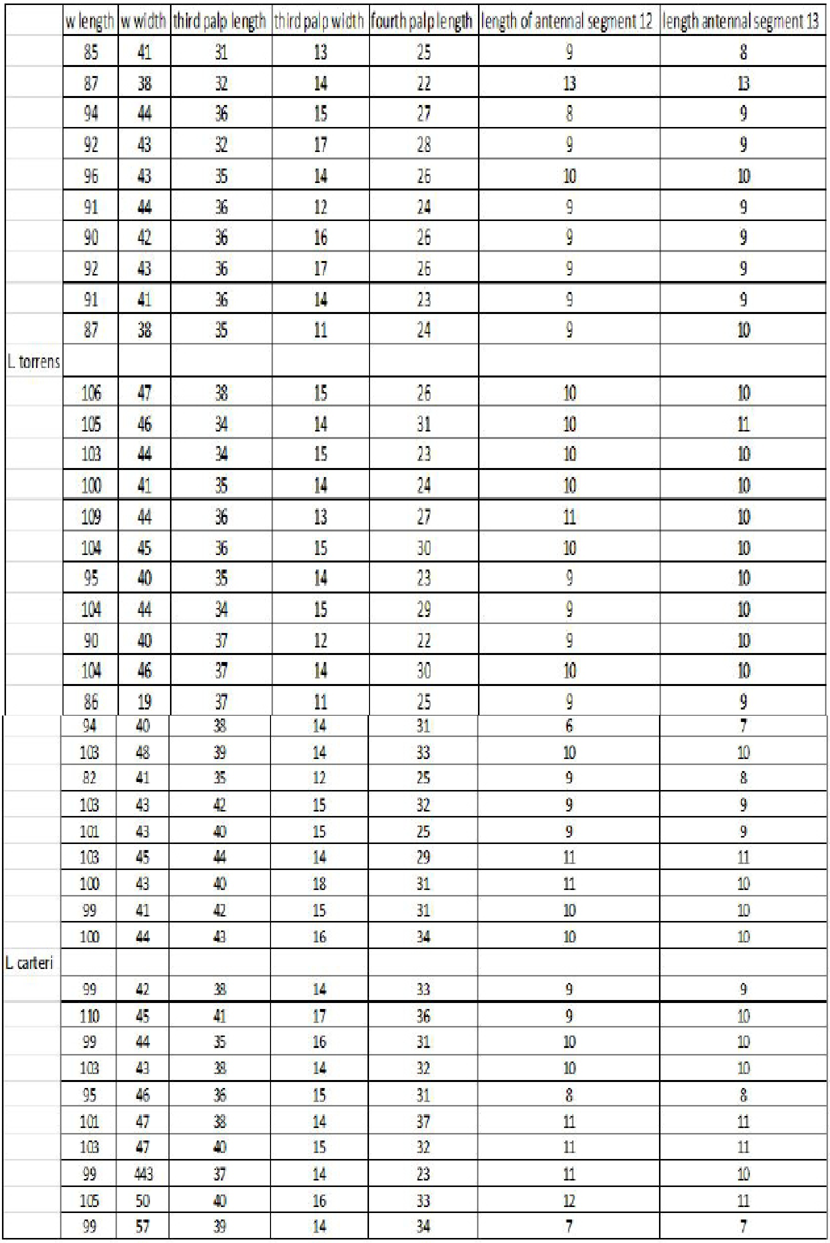
Then the researchers have the following allocation rule:

$$\text{Allocate} X_j \text{ to } \pi_2 \text{ if } d_j > \frac{S_1\mu_2 + S_1\mu_1}{S_1 + S_2} \quad \text{if } \mu_2 > \mu_1 \tag{14}.$$

RESULTS AND DISCUSSION

Two species of biting flies (genus *Leptoconops*) are very similar morphologically so that for many years they were actually thought to be the same. Later, biological differences such as sex differences of emerging flies and biting habits were found to exist. Data on some morphological characteristics of two species of the biting flies: *L. carteri* and *L. torrens* were obtained by W. Atchely from the website (www.prenhall.com/statistics) and are reproduced here for convenience:

Table 1: Bitting - Fly Data



	w length	w width	third palp length	third palp width	fourth palp length	length of antennal segment 12	length antennal segment 13
	85	41	31	13	25	9	8
	87	38	32	14	22	13	13
	94	44	36	15	27	8	9
	92	43	32	17	28	9	9
	96	43	35	14	26	10	10
	91	44	36	12	24	9	9
	90	42	36	16	26	9	9
	92	43	36	17	26	9	9
	91	41	36	14	23	9	9
	87	38	35	11	24	9	10
L torrens							
	106	47	38	15	26	10	10
	105	46	34	14	31	10	11
	103	44	34	15	23	10	10
	100	41	35	14	24	10	10
	109	44	36	13	27	11	10
	104	45	36	15	30	10	10
	95	40	35	14	23	9	10
	104	44	34	15	29	9	10
	90	40	37	12	22	9	10
	104	46	37	14	30	10	10
	86	19	37	11	25	9	9
	94	40	38	14	31	6	7
	103	48	39	14	33	10	10
	82	41	35	12	25	9	8
	103	43	42	15	32	9	9
	101	43	40	15	25	9	9
	103	45	44	14	29	11	11
	100	43	40	18	31	11	10
	99	41	42	15	31	10	10
	100	44	43	16	34	10	10
L carteri							
	99	42	38	14	33	9	9
	110	45	41	17	36	9	10
	99	44	35	16	31	10	10
	103	43	38	14	32	10	10
	95	46	36	15	31	8	8
	101	47	38	14	37	11	11
	103	47	40	15	32	11	11
	99	443	37	14	23	11	10
	105	50	40	16	33	12	11
	99	57	39	14	34	7	7

A principal components analysis was performed on the correlation matrix to extract the significant orthogonal linear combinations. Figure 1 shows the Scree Plot which suggests the extraction of four or five principal components:

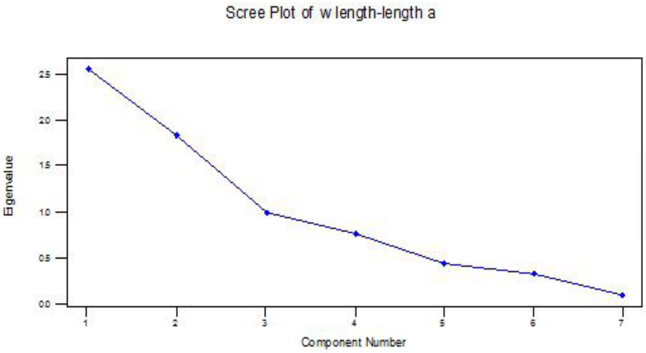
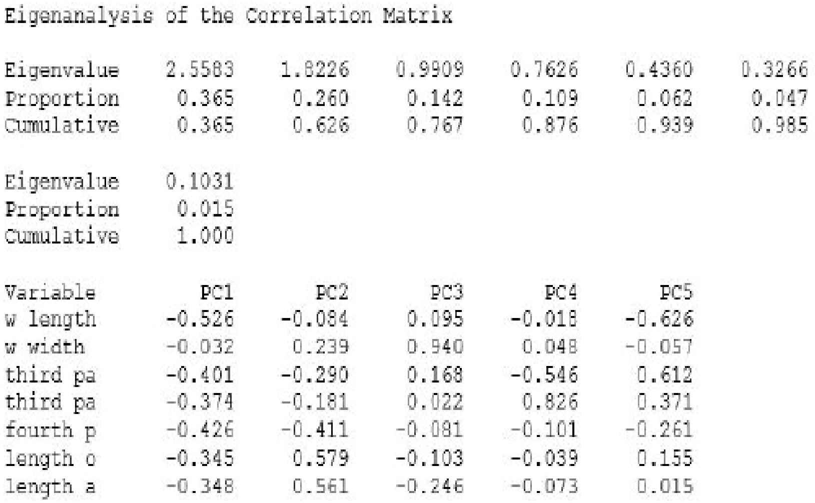


Figure 1: Scree-Plot

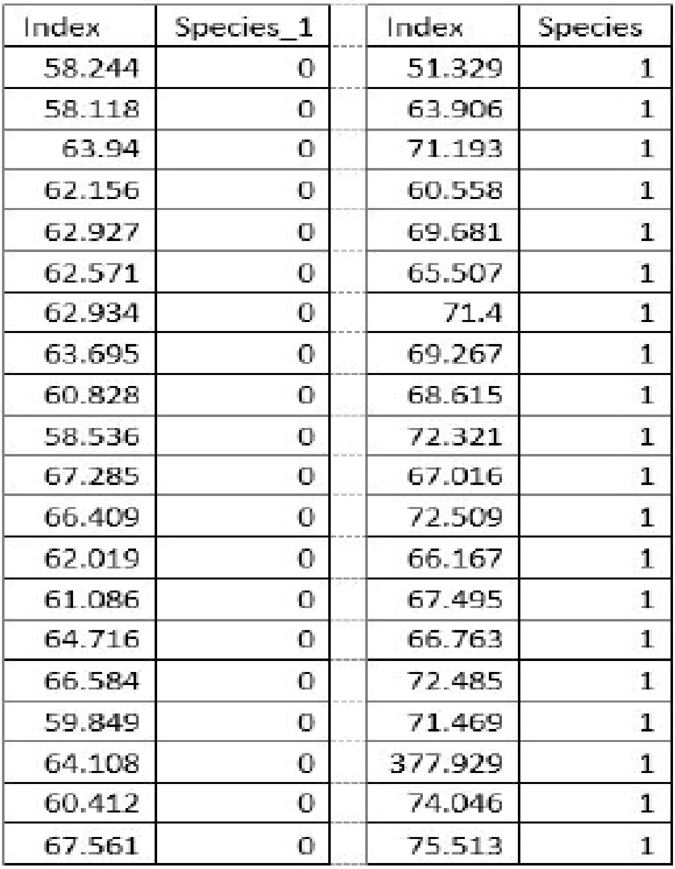
Table 2: Principal Components Analysis of the Biting Fly Data



```
Eigenanalysis of the Correlation Matrix

Eigenvalue    2.5583   1.8226   0.9909   0.7626   0.4360   0.3266
Proportion     0.365    0.260    0.142    0.109    0.062    0.047
Cumulative     0.365    0.626    0.767    0.876    0.939    0.985

Eigenvalue    0.1031
Proportion     0.015
Cumulative     1.000

Variable         PC1      PC2      PC3      PC4      PC5
w length      -0.526   -0.084    0.095   -0.018   -0.626
w width       -0.032    0.239    0.940    0.048   -0.057
third pa      -0.401   -0.290    0.168   -0.546    0.612
third pa      -0.374   -0.181    0.022    0.826    0.371
fourth p      -0.426   -0.411   -0.081   -0.101   -0.261
length o      -0.345    0.579   -0.103   -0.039    0.155
length a      -0.348    0.561   -0.246   -0.073    0.015
```

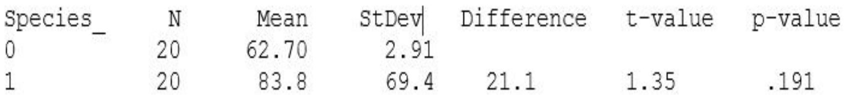
Tabular values reveal that the first four principal components accounted for 87.6% of the total variance whereas the first five principal components reproduced 93.9% of the total variance. The researchers chose the first five principal components as basis for computing the allometric index. The indices for the individual specimen flies are provided in Table 3:

Table 3: Computed Allometric Index for the Individual Flies



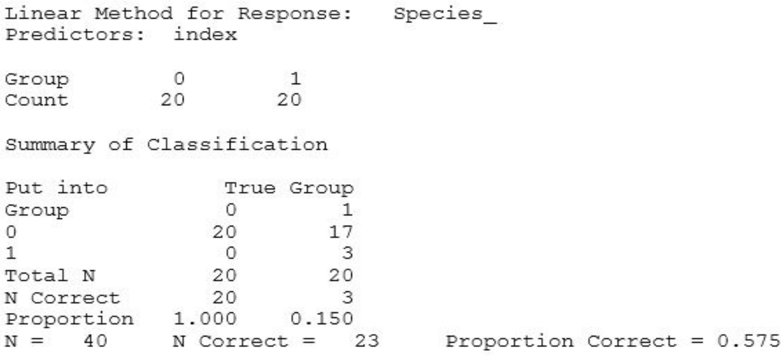
Index	Species_1	Index	Species
58.244	0	51.329	1
58.118	0	63.906	1
63.94	0	71.193	1
62.156	0	60.558	1
62.927	0	69.681	1
62.571	0	65.507	1
62.934	0	71.4	1
63.695	0	69.267	1
60.828	0	68.615	1
58.536	0	72.321	1
67.285	0	67.016	1
66.409	0	72.509	1
62.019	0	66.167	1
61.086	0	67.495	1
64.716	0	66.763	1
66.584	0	72.485	1
59.849	0	71.469	1
64.108	0	377.929	1
60.412	0	74.046	1
67.561	0	75.513	1

A two-sample t-test was performed to determine if the indices differed significantly between the two groups. Table 4 shows the results of the t-test:



```
Species_      N       Mean     StDev   Difference    t-value    p-value
0             20      62.70     2.91
1             20       83.8     69.4     21.1          1.35       .191
```

Since the computed t-value failed to reach the value for significance at the .05 level, hence, it is concluded that the indices are statistically the same for the two species of flies. Hence, the allometric index computed would not be effective in discriminating members of the two fly species. A discriminant analysis was likewise performed to confirm this statement. Table 5 summarizes the result of the discriminant analysis.



```
Linear Method for Response:   Species_
Predictors:   index

Group          0        1
Count         20       20

Summary of Classification

Put into           True Group
Group                0        1
0                   20       17
1                    0        3
Total N             20       20
N Correct           20        3
Proportion       1.000    0.150
N =    40      N Correct =    23      Proportion Correct = 0.575
```

Discriminant analysis confirmed what was initially suspected. The probability of correct classification registered a 57.5% efficiency rate which implies a probability of misclassification of about 42.3%.

Closer inspection of the t-test results revealed that the small t-value computed is due to the disproportionately large variance in the indices computed for the second species (*L. torrens*). Likewise, this large variance accounted for the very low correct classification rate for flies belonging to this species (probability of correctly classifying a fly in = 15%).

Having obtained this unsatisfactory result, the researchers proceeded to re-analyze the data set using the original covariance matrix as an input. The Scree Plot obtained for the principal components analysis with the covariance matrix as an input is shown below:

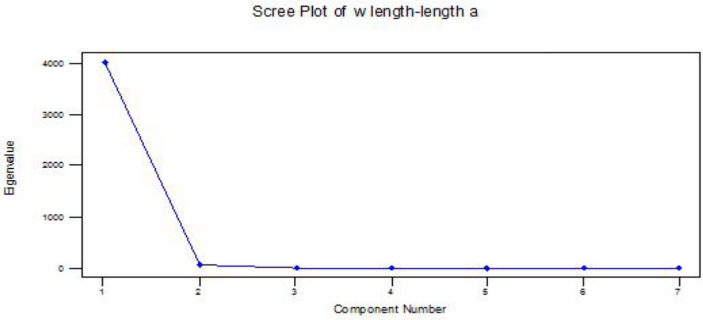
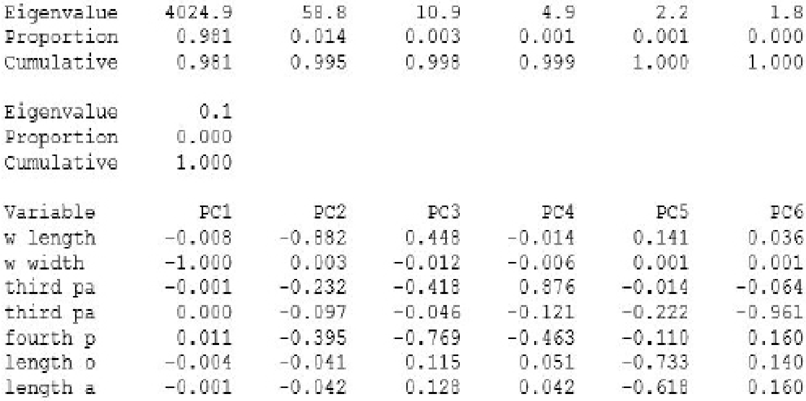


Figure 2: Scree Plot Using the Covariance Matrix as Input on the Biting Fly Data

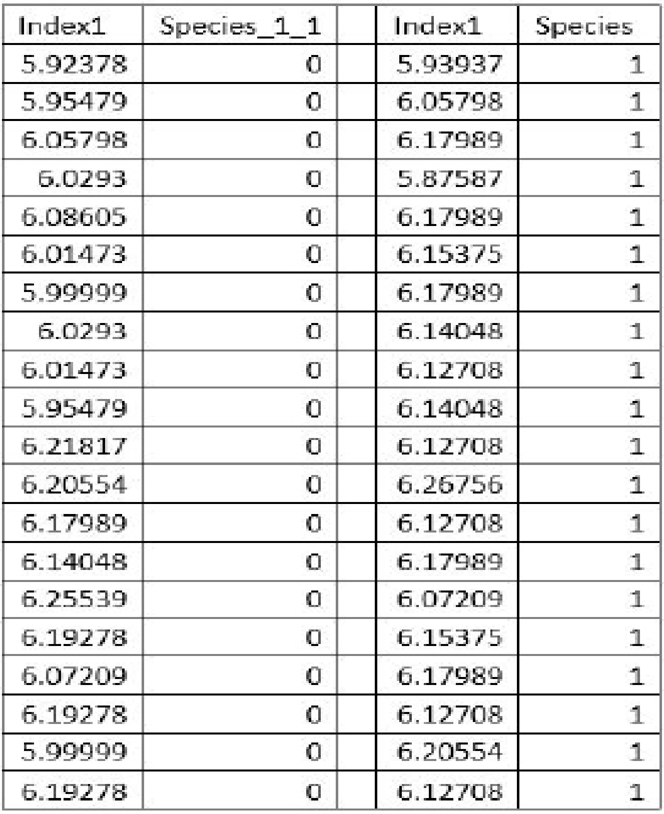
The Scree-Plot suggests taking only the first or at most the first two principal components. The full principal components decomposition is shown in Table 6.



```
Eigenvalue     4024.9     58.8     10.9      4.9      2.2      1.8
Proportion      0.981    0.014    0.003    0.001    0.001    0.000
Cumulative      0.981    0.995    0.998    0.999    1.000    1.000

Eigenvalue        0.1
Proportion      0.000
Cumulative      1.000

Variable          PC1      PC2      PC3      PC4      PC5      PC6
w length       -0.008   -0.882    0.448   -0.014    0.141    0.036
w width        -1.000    0.003   -0.012   -0.006    0.001    0.001
third pa       -0.001   -0.232   -0.418    0.876   -0.014   -0.064
third pa        0.000   -0.097   -0.046   -0.121   -0.222   -0.961
fourth p        0.011   -0.395   -0.769   -0.463   -0.110    0.160
length o       -0.004   -0.041    0.115    0.051   -0.733    0.140
length a       -0.001   -0.042    0.128    0.042   -0.618    0.160
```

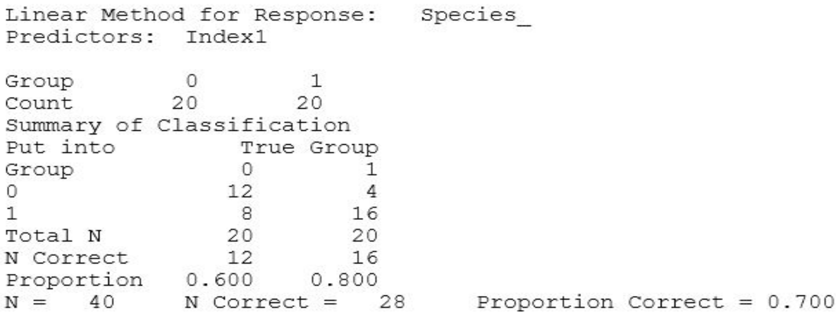
Since 99.5% of the total variance is already assumed for the first two principal components, the researchers choose to ignore the other principal components. Furthermore, the first principal component is mainly dominated by wing width while the second principal component is dominated by wing length. Thus, the researchers obtained an index that essentially considers the wing dimensions of the flies. Furthermore, the researchers transformed the wing dimensions by a logarithmic transformation prior to computing the indices. The new computed allometric indices are shown in Table 7.

Table 7: New Allometric Indices Using the Covariance Matrix



Index1	Species_1_1		Index1	Species
5.92378	0		5.93937	1
5.95479	0		6.05798	1
6.05798	0		6.17989	1
6.0293	0		5.87587	1
6.08605	0		6.17989	1
6.01473	0		6.15375	1
5.99999	0		6.17989	1
6.0293	0		6.14048	1
6.01473	0		6.12708	1
5.95479	0		6.14048	1
6.21817	0		6.12708	1
6.20554	0		6.26756	1
6.17989	0		6.12708	1
6.14048	0		6.17989	1
6.25539	0		6.07209	1
6.19278	0		6.15375	1
6.07209	0		6.17989	1
6.19278	0		6.12708	1
5.99999	0		6.20554	1
6.19278	0		6.12708	1

Using the allometric index as a discriminatory feature for the two species of the flies, the researchers performed a discriminant analysis as shown in Table 8:



```
Linear Method for Response:   Species_
Predictors:  Index1

Group         0         1
Count        20        20
Summary of Classification
Put into          True Group
Group              0         1
0                 12         4
1                  8        16
Total N           20        20
N Correct         12        16
Proportion     0.600     0.800
N =    40      N Correct =    28      Proportion Correct = 0.700
```

Tabular values show that the proportion of correctly classified species of flies have increased to an acceptable 70% level (probability of misclassification of about 30%).

CONCLUSION

A single numerical allometric index can be computed based on the Euclidean distance of the first m-principal components, m < p from the origin. The discriminatory power of this index to assign an organism into one of two possible species depends on the input matrix. The" best" input matrix (covariance or correlation matrix) is the input matrix that outputs fewer principal components accounting for more than 90% of the total variance. The method is comparable to the method suggested in the paper of Lapinig et al. (2016) when only the first principal component is used.

LITERATURE CITED

Bonner JT. 2006. Why size matters: from bacteria to blue whales, Princeton, NJ: Princeton University Press.

Calder WA. 1984. Size, function and life history. Cambridge, MA: Harvard University Press.

Cheverud JM. 1970. Relationships among ontogenetic, static and evolutionary allometry. American Journal of Physiological Anthropology 59, 139-149.

Cooper S. 1890. Animal life in the sea and on the land: A zoology for young people. New York, NY: Harper & Brother.

Gayon J. 2000. History of the concept of allometry. American zoologist 40, 748-758.

Gould SJ. 1966. Allometry and size in ontogeny and phylogeny. Biological review of the cambridge philosophical society 41, 587.

Gould SJ. 1971. Geometric similarity in allometric growth: contribution to problem of scaling in evolution of size. American naturalist 105, 113-136.

Huxley JS. 1924. Constant differential growth-ratios and their significance. nature 114, 895-896.

Huxley JS, Tessier G. 1935. Terminology of relative growth nature 137, 780-781.

Kleiber M. 1932. Body size and metabolism. Hilgardia 6, 315-353.

Kolokotrones T, Van et al. 2010. Curvature in metabolic scaling. Nature 464, 753-756.

Lapinig VT, Padua RN. 2017. Construction of allometric indices by principal components analysis with application to biological classification. (Unpublished technical report, Northwestern Mindanao State College 2016)

Miller D. 1973. Growth in uca, ontogeny of asymmetry in ucapugilator (Bosc) (Decapoda, Ocypodidae). Crustaceana, 119-131.

Moore KL. 1983. The developing human. Philadelphia, PA: W. B. Saunders.

Samaras TT. 2007. Human body size and the laws of scaling. Hauppauge, NY: Nova science publishers.

Schmidt-Nielsen K. 1984. Scaling: Why is animal size so important? Cambridge, UK: Cambridge University Press.

Thompson DW. 1917. On growth and form. Cambridge, UK: Cambridge University Press.

West GB, Brown JH et al . 1997. A general model for the origin of allometric scaling laws in biology. Science 276, 122-126.