Sains Malaysiana 50(7)(2021): 2079-2084
http://doi.org/10.17576/jsm-2021-5007-21

COVRATIO Statistic as A Discrimination Method for Multivariate Normal Distribution

(Statistik *COVRATIO* sebagai Suatu Kaedah Diskriminasi untuk Taburan Multivariat Normal)

NORLI ANIDA ABDULLAH*, AFERA MOHAMAD APANDI, MOHD IQBAL SHAMSUDHEEN & YONG ZULINA ZUBAIRI

ABSTRACT

The COVRATIO statistic has been used to identify the presence of outlier in data, which is based on deletion approach, where the determinant of covariance matrix for the full dataset excludes i*-th row. This study proposes a novel discrimination method for the multivariate normal (MVN) distribution using the idea of COVRATIO statistic, denoted as* $\text{COVRATIO}_{(+i)}$. *The linear discrimination function (LDF) for MVN distribution will be compared to the* $\text{COVRATIO}_{(+i)}$ *statistic. Simulation results showed that the* $\text{COVRATIO}_{(+i)}$ *as discrimination method performs better than the LDF with lower misclassification probabilities in all cases considered. The interest in the discrimination method arose in connection with the study of an application to discriminate the shape of the human maxillary dental arches, thus* $\text{COVRATIO}_{(+i)}$ *statistic may be considered as an alternative.*

Keywords: COVRATIO statistic; dental arch; discrimination method; linear discrimination function; multivariate normal distribution

ABSTRAK

Statistik COVRATIO *telah digunakan untuk mengenal pasti kehadiran data luar dengan menggunakan kaedah penghapusan, dengan baris i dari penentu matriks kovarians dikeluarkan daripada set data penuh. Kajian ini mencadangkan kaedah diskriminasi baru untuk taburan normal multivariat (MVN) menggunakan idea daripada statistik* COVRATIO*, yang dikenali sebagai* $\text{COVRATIO}_{(+i)}$. *Fungsi diskriminasi linear (LDF) untuk taburan MVN akan dibandingkan dengan kaedah tersebut. Hasil simulasi menunjukkan bahawa statistik diskriminasi* $\text{COVRATIO}_{(+i)}$ *adalah lebih baik daripada LDF dengan kebarangkalian salah pengelasan yang lebih rendah dalam semua kes yang dipertimbangkan. Kepentingan kaedah diskriminasi timbul dalam kajian membezakan bentuk arkus pergigian maksila manusia dan statistik* $\text{COVRATIO}_{(+i)}$ *ini boleh digunakan sebagai alternatif.*

Kata kunci: Arkus pergigian; fungsi diskriminasi linear; kaedah diskriminasi; statistik COVRATIO*; taburan multivariat normal*

INTRODUCTION

The study of discriminant analysis has been widely carried out in different areas of statistics such as biology, medicine, computer vision, bankruptcy prediction, and security. By definition, discrimination is a multivariate technique concerned with separating sets of objects to categories or classes (Johnson & Wichern 1992). Describing, comparing, classifying, recognizing, and discriminating the shape of natural and man-made objects are of interest in many disciplines (Dass & Li 2009; Lacko et al. 2015; Laganà et al 2019; Yergin et al 2001). In the literature, principal component analysis (PCA) is the most common method used to provide information for discrimination as it explains the variability of the data that quantifies shape differences. However, the main limitation of PCA is that it does not consider class separability since there is no mention of class label of the shape. Hence, assigning new objects to a particular class may be difficult to accomplish (Costa & Cesar 2009). The nearest neighbor method is a non-parametric and simplest discrimination approach. It basically identifies the sample in a set of N samples that is closest to a new object and takes its class. However, the performance of this method is generally inferior to the parametric discriminant as it does not supply information

about the probability density function of the class (Costa & Cesar 2009).

The linear discrimination function (LDF) is the most common discriminant method with MVN and equal covariance matrices assumptions. Let $f_k(\mathrm{v})$ and $f_j(\mathrm{v})$ be the probability density function of the k-th and j-th population, respectively, and $\mathbf{v}$ be the new observation. The LDF is given as:

$$h_{kj}(\mathbf{v}) = \log\left(\frac{f_k(\mathbf{v})}{f_j(\mathbf{v})}\right) = (\bar{\mathbf{v}}(k)-\bar{\mathbf{v}}(j))'\mathbf{S}_{pooled}^{-1}\mathbf{v} - \frac{1}{2}(\bar{\mathbf{v}}(k)-\bar{\mathbf{v}}(j))'\mathbf{S}_{pooled}^{-1}(\bar{\mathbf{v}}(k)+\bar{\mathbf{v}}(j)). \quad \because \mathbf{S}_k = \mathbf{S}_j \tag{1}$$

where $\mathbf{S}_{pooled} = \dfrac{(n_k-1)\mathbf{S}_k + (n_j-1)\mathbf{S}_j}{n_k+n_j-2}$, n_k and n_j are the respective sample size and $k \neq j$. The discriminant rule for the 2 populations is as follows (Johnson & Wichern 1992; Mardia et al. 1979)

$$\text{if } h_{kj}(\mathbf{v}) > 0 \text{, then } \mathbf{v} \in \Pi_k \text{,}$$

$$\text{if } h_{kj}(\mathbf{v}) < 0 \text{, then } \mathbf{v} \in \Pi_j \text{, and} \tag{2}$$

The Gaussian and equality of covariance matrices assumptions for the discriminant function are particularly restrictive and sometimes difficult to satisfy. If violated, it may reduce the performance of discriminant analysis. Hence, this study proposes a new discriminant based on COVRATIO statistic as an alternative discrimination method, which uses the determinant ratio of two covariance matrices that are not necessarily from Gaussian distribution.

This paper is organized as follows. First section begins with the introduction of this article. Next section proposes COVRATIO as a new discrimination method. Subsequent section presents a comparison between the LDF and COVRATIO via simulation study on performance of the test and followed by its application to real data in the following section. Finally, conclusion of this paper is provided in the last section.

PROPOSED $COVRATIO_{(+i)}$ AS DISCRIMINATION METHOD

The COVRATIO procedure dates back to Belsley et al. (1980). They proposed a numerical statistic to identify the presence of influential observation in linear regression models. This numerical statistic is based on the determinantal ratio given as:

$$COVRATIO(-i) = \frac{|COV(-i)|}{|COV|} \text{ ,} \tag{3}$$

where $|COV|$ is the determinant of covariance matrix for full data set and $|COV(-i)|$ is that for the reduced data set by excluding the i^{th} observation. If the ratio is close to 1, then there is no significant difference between them. In other words, the i^{th} observation is consistent with the other observations. Alternatively, if the value of $|COV(-i)|$ is close to or larger than a derived cut off point, then it indicates that the i^{th} observation is a candidate of an outlier. This procedure has been extended and employed for other models such as the simple linear functional relationship model and circular regression model (Ghapor et al. 2014; Ibrahim et al. 2013; Rambli et al. 2015).

In this study, the idea of COVRATIO statistic for outlier detection was extended for the purpose of discrimination. Let II_1 and II_2 be two groups of pre-defined populations with known probability distribution where the mean and covariance matrix exist. Also let Σ_1 and Σ_2 be their $(p \times p)$ covariance matrix, respectively. The determinant of Σ_1 and Σ_2, denoted as $|\Sigma_1|$ and $|\Sigma_2|$ gives the descriptive measures of multivariate variability (Peña & Rodríguez 2003). Next, instead of using reduced observation as formulated by Belsley et al. (1980), define the ratio of covariance determinant for Π_1 as:

$$COVRATIO_{1(+i)} = \frac{\left|\mathbf{\Sigma}_{1(+i)}\right|}{\left|\mathbf{\Sigma}_1\right|}, \tag{4}$$

where $|\Sigma_{1(+i)}|$ is the determinant of covariance matrix $\Sigma_{1(+i)}$ which is estimated by adding new i^{th} observation in II_1.

The same i^{th} observation will be added in II_2 and its $COVRATIO_{2(+i)}$ is obtained using

$$COVRATIO_{2(+i)} = \frac{\left|\mathbf{\Sigma}_{2(+i)}\right|}{\left|\mathbf{\Sigma}_2\right|}. \tag{5}$$

Smaller value of $COVRATIO_{(+i)}$ (which is close to 1) indicates no variation when new observation was included in a particular population. In other words, the i^{th} observation is consistent with the other observations and therefore belongs to the population.

Therefore, the discrimination rule using $COVRATIO_{(+i)}$ for 2 populations is given as:

Allocate new i^{th} observation in Π_1 if $COVRATIO_{1(+i)} < COVRATIO_{2(+i)}$,
Else, allocate to Π_2 (6)

These result can be generalized for g populations, given as:

Allocate new i^{th} observation in Π_j if
$COVRATIO_{j(+i)} < ... < COVRATIO_{h(+i)}$, (7)
where $j, h = 1, 2, ..., g$ and $j \neq h$.

SIMULATION STUDY FOR COMPARING PERFORMANCE OF LDF AND $COVRATIO_{(+i)}$

The performance of LDF and $COVRATIO_{(+i)}$ was examined *via* simulation study. Random samples from two MVN populations were generated and re-assigned to their corresponding population using LDF and $COVRATIO_{(+i)}$. The misclassification probabilities using these discrimination methods were then compared. This procedure was done as follows:

Step 1 Two sets of random samples with sample size n from p-variate normal distributions; Π_1: $\mathbf{X}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and Π_2: $\mathbf{X}_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ were generated using R random number generator.
Step 2 Generate another two random samples **A** and **B** which were drawn from Π_1 and Π_2 respectively.
Step 3 Calculate the LDF in Equation (1) using random sample **A** and **B**, denoted as h_{12}(A) and h_{12}(B), respectively.
Step 4 Calculate $COVRATIO_{1(+i)}$ (A), $COVRATIO_{1(+i)}$ (B), $COVRATIO_{2(+i)}$ (A) and $COVRATIO_{2(+i)}$ (B) using (4) and (5), by adding random sample **A** and **B** in both sets of random samples generated from Π_1 and Π_2 in step 1.
Step 5 Discrimination rule for LDF and $COVRATIO_{(+i)}$ (in (2) and (6)) were used to determine the membership of **A** and **B**.
Step 6 The above steps were repeated for s = 10000 times.
Step 7 The proportion of **A** and **B** which does not assign to their corresponding true population when discriminated using LDF and $COVRATIO_{(+i)}$ may be regarded as the misclassification probability.

Table 1 shows the performance of LDF and $COVRATIO_{(+i)}$ when different values of n, p, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were considered. In general, the misclassification probability for LDF and $COVRATIO_{(+i)}$ reduces when a particular value of $\boldsymbol{\mu}_2$ was chosen to indicate larger separation between Π_2 and Π_1. The dispersion of the data in a particular population also dictates the misclassification probability, whereby larger value of $\boldsymbol{\Sigma}$ may result in higher chances of a sample to be assigned to an incorrect population. Increasing value of n gives slightly higher misclassification probability where similar findings can be observed in Moawed and Osman (2017). On the other hand, increasing value of p gives smaller misclassification probability. However, at $p = 8$, the LDF breaks down and shows extremely high misclassification probability due to the curse of dimensionality (Trunk 1979).

Overall, the $COVRATIO_{(+i)}$ method gives smaller probability misclassification as compared to the LDF in all cases, indicating better performance of discrimination. The proposed $COVRATIO_{(+i)}$ is preferable, as it does not require Gaussian and equality of covariance matrices assumptions, which may be difficult to attain. It also uses the raw data instead of descriptive statistics when discriminating therefore retaining information about the population. Moreover, due to its simplicity, the $COVRATIO_{(+i)}$ can be employed when discriminating more than 2 groups. The drawback of $COVRATIO_{(+i)}$ method is that it is computationally more expensive compared to LDF; however, with technology advancement in available statistical software, such computation is straightforward and incredibly fast.

APPLICATION TO REAL DATA

Obtaining the shape feature, discriminating groups of shape and modelling categories of shape of the dental arch has clinical importance in orthodontic and prosthetic dentistry. The data of the maxillary dental arch shape from Rijal et al. (2011) were considered. The membership of the 47 control samples which belong to each of the 3 shape models of human dental arch were tracked from the dendrogram (Rijal et al. 2012, 2011). Then, the LDF and $COVRATIO_{(+i)}$ were employed to re-assign these casts into one of 3 populations of the shape models. Their misclassification probability was calculated to ensure that these methods are deemed good for shape discrimination.

Table 2 shows the misclassification probabilities when the samples were re-assigned to Π_1^*,Π_2^* or Π_3^* using the LDF and $COVRATIO_{(+i)}$. Relatively low misclassification probabilities for LDF were obtained, however $COVRATIO_{(+i)}$ outperformed the LDF method by giving zero misclassification probabilities. These results demonstrate the ability of the $COVRATIO_{(+i)}$ to discriminate the dental arch shape correctly; thus, it can be regarded as a better discrimination method.

TABLE 1. Misclassification probability for LDF and $COVRATIO_{(+i)}$ when different sample size, dimension, mean vector and covariance matrices were considered. Values in the parentheses indicate the standard error of the simulation

Parameters		Population Π_1		Population Π_2	
LDF		LDF	$COVRATIO_{(+i)}$	LDF	$COVRATIO_{(+i)}$
$n_1 = n_2 = 20, p = 2,$ $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$	0.0741 (0.0026)	0.0622 (0.0024)	0.0672 (0.0025)	0.0577 (0.0023)
	$\boldsymbol{\mu}_2 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$	0.0141 (0.0012)	0.0049 (0.0007)	0.0149 (0.0012)	0.0046 (0.0007)
	$\boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$	0.0015 (0.0004)	0 (0.0000)	0.0013 (0.0004)	0 (0.0000)
$n_1 = n_2 = 20, p = 2,$ $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$ $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}$	0.0022 (0.0005)	0.0005 (0.0002)	0.0093 (0.0010)	0.0005 (0.0002)
	$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$	0.0026 (0.0005)	0.0012 (0.0003)	0.0203 (0.0014)	0.0021 (0.0005)
	$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 2.5 & 0 \\ 0 & 2.5 \end{pmatrix}$	0.0031 (0.0006)	0.0022 (0.0005)	0.0301 (0.0017)	0.0070 (0.0008)
	$\boldsymbol{\Sigma}_2 = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$	0.0031 (0.0006)	0.0031 (0.0006)	0.0460 (0.0021)	0.0114 (0.0011)
$p = 2,$ $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$n_1 = n_2 = 10$	0.0640 (0.0024)	0.0530 (0.0022)	0.0634 (0.0024)	0.0504 (0.0022)
	$n_1 = n_2 = 20$	0.0741 (0.0026)	0.0622 (0.0024)	0.0672 (0.0025)	0.0577 (0.0023)
	$n_1 = n_2 = 50$	0.0820 (0.0027)	0.0663 (0.0025)	0.0799 (0.0027)	0.0636 (0.0024)
$n_1 = n_2 = 20,$ $\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ \vdots \\ 2 \end{pmatrix},$ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$	$p = 2$	0.0741 (0.0026)	0.0622 (0.0024)	0.0672 (0.0025)	0.0577 (0.0023)
	$p = 4$	0.0177 (0.0013)	0.0087 (0.0009)	0.0163 (0.0013)	0.0104 (0.0010)
	$p = 6$	0.0040 (0.0006)	0.0018 (0.0004)	0.0050 (0.0007)	0.0015 (0.0004)
	$p = 8$	0.4000 (0.0049)	0 (0.0000)	0.9000 (0.0030)	0 (0.0000)

TABLE 2. Misclassification probability when 47 dental casts were re-assigned into either one of the populations of the shape model using the LDF and $COVRATIO_{(+i)}$

Discrimination / True population	Π_1^*		Π_2^*		Π_3^*	
	LDF	$COVRATIO_{(+i)}$	LDF	$COVRATIO_{(+i)}$	LDF	$COVRATIO_{(+i)}$
Π_1	-	-	0.18	0	0.18	0
Π_2	0.09	0	-	-	0.09	0
Π_3	0.21	0	0.07	0	-	-

CONCLUSION

This paper proposed a novel discrimination method for the multivariate normal (MVN) distribution using the idea of COVRATIO statistic, denoted as $COVRATIO_{(+i)}$. The $COVRATIO_{(+i)}$ statistic gives smaller probability misclassification as compared to the LDF in all cases indicating better performance of discrimination. Therefore, it can be an accurate alternative method for discrimination.

ACKNOWLEDGEMENTS

This study was supported by a Fundamental Research Grant Scheme from Ministry of Higher Education Malaysia (FP042-2017A) and Bantuan Kecil Penyelidikan (BK003-2017).

REFERENCES

Belsley, D.A., Kuh, E. & Welsch, R.E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Costa, L.D.F. & Cesar, R.M. 2009. *Shape Analysis and Classification: Theory and Practice.* Florida: CRC Press.

Dass, S.C. & Li, M. 2009. Hierarchical mixture models for assessing fingerprint individuality. *The Annals of Applied Statistics* 3(4): 1448-1466.

Ghapor, A.A., Zubairi, Y.Z., Mamun, A. & Rahmatullah Imon, A.H.M. 2014. On detecting outlier in simple linear functional relationship model using *COVRATIO* statistic. *Pakistan Journal of Statistics* 30(1): 129-142.

Ibrahim, S., Rambli, A., Hussin, A.G. & Mohamed, I. 2013. Outlier detection in a circular regression model using *COVRATIO* statistic. *Communications in Statistiscs-Simulation and Computation* 42(10): 2272-2280.

Johnson, R.A. & Wichern, D.W. 1992. *Applied Multivariate Statistical Analysis: Discrimination and Classification.* 3rd ed. Englewood Cliffs, New Jersey: Prentice Hall.

Lacko, D., Huysmans, T., Parizel, P.M., De Bruyne, G., Verwulgen, S., Van Hulle, M.M. & Sijbers, J. 2015. Evaluation of an anthropometric shape model of the human scalp. *Applied Ergonomics* 48: 70-85.

Laganà, G., Di Fazio, V., Paoloni, V., Franchi, L., Cozza, P. & Lione, R. 2019. Geometric morphometric analysis of the palatal morphology in growing subjects with skeletal open bite. *European Journal of Orthodontics* 41(3): 258-263.

Mardia, K.V., Kent, J.T. & Bibby, J.M. 1979. *Multivariate Analysis*. New York: Academic Press.

Moawed, S.A. & Osman, M.M. 2017. The robustness of binary logistic regression and linear discriminant analysis for the classification and differentiation between dairy cows and buffaloes. *International Journal of Statistics and Applications* 7: 304-310.

Peña, D. & Rodríguez, J. 2003. Descriptive measures of multivariate scatter and linear dependence. *Journal of Multivariate Analysis* 85(2): 361-374.

Rambli, A., Yunus, R.M., Mohamed, I. & Hussin, A.G. 2015. Outlier detection in a circular regression model. *Sains Malaysiana* 44(7): 1027-1032.

Rijal, O.M., Abdullah, N.A., Isa, Z.M., Noor, N.M. & Tawfiq, O.F. 2012. A probability distribution of shape for the dental maxillary arch using digital images. In *34th Annual International Conference of the IEEE Engineering-in-Medicine-and-Biology-Society (EMBS)*. IEEE. pp. 5420-5423.

Rijal, O.M., Abdullah, N.A., Isa, Z.M., Davaei, F.A., Noor, N.M. & Tawfiq, O.F. 2011. A novel shape representation of the dental arch and its applications in some dentistry problems. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE*. pp. 5092-5095.

Trunk, G.V. 1979. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3: 306-307.

Yergin, E., Ozturk, C. & Sermet, B. 2001. Image processing techniques for assessment of dental trays. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society 3*. IEEE. pp. 2571-2573.

Norli Anida Abdullah* & Yong Zulina Zubairi
Centre for Foundation Studies in Science
University of Malaya, Jalan Universiti
50603 Kuala Lumpur, Federal Territory
Malaysia

Afera Mohamad Apandi
Institstute of Advanced Studies
University of Malaya, Jalan Universiti
50603 Kuala Lumpur, Federal Territory
Malaysia

Mohd Iqbal Shamsudheen
Department of Statistical Science
University College London
London
United Kingdom

*Corresponding author; email: norlie@um.edu.my

Received: 21 February 2020
Accepted: 19 November 2020