Prediction of dengue infection severity using classic and robust discriminant approaches

by Windarto Windarto

Submission date: 22-Feb-2023 12:38PM (UTC+0800)
Submission ID: 2020203241
File name: on_severity_using_classic_and_robust_discriminant_approaches.pdf (515.76K)
Word count: 4716
Character count: 23388

Prediction of dengue infection severity using classic and robust discriminant approaches

Cite as: AIP Conference Proceedings **2329**, 060021 (2021); https://doi.org/10.1063/5.0042127
Published Online: 26 February 2021

Toha Saifudin, and Windarto

Prediction of Dengue Infection Severity Using Classic and Robust Discriminant Approaches

Toha Saifudin [1,a)] and Windarto [2,b)]

[1]*Study Program of Statistics, Faculty of Science and Technology, Universitas Airlangga, Indonesia*
[2]*Study Program of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Indonesia*

[a)]Corresponding author: tohasaifudin@fst.unair.ac.id
[b)] windarto@fst.unair.ac.id

Abstract. Dengue infection is one of feared diseases in the public because it often results in death in sufferers. Patients suspected of dengue infection are usually routinely drawn their blood to be checked in the laboratory examination. Unfortunately, death can be caused by a lack of speed and proper handling according to the severity of the patient. Refer to this problem, it is necessary to predict dengue infection severity based on blood diagnose results. This is important to prepare the precise treatment according to the severity of patients in order to reduce the number of death from this disease. Because the patient's blood examination result is a multivariate dataset then in this paper the prediction was solved using multivariate method, namely discriminant analysis. In this method, the parameter estimation was carried out using Maximum Likelihood (ML) method. This leads to classic discriminant analysis. Unfortunately, the ML method is heavily influenced by outlier so the estimator becomes less precise when data has been contaminated by outliers. To overcome this problem, a robust estimation method using Minimum Covariance Determinant (MCD) was used. This leads to the robust discriminant analysis. This study used a sample of dengue infection patient medical record data from Surabaya Hajj Hospital. The result of this study showed that the appropriate analysis for sample data was the quadratic discriminant analysis. Furthermore, the robust quadratic model with MCD estimator produced better prediction than the classic quadratic model with ML estimator. The robust quadratic model produced percentage of classification accuracy of 87.2% in the male patient training data which is greater than the classic quadratic model accuracy of 85.7%. In the female patient training data, the robust quadratic model produced percentage of classification accuracy of 88.7% which is greater than the classic quadratic model accuracy of 80.7%. In addition, the MCD estimator was able to detect more outlier data than the ML estimator.

INTRODUCTION

World Health Organization (WHO) stated that dengue infection is a tropical disease that spreads fastest and as a 'threat of a new pandemic'. Dengue infection has developed into a serious problem in several tropical countries, especially Indonesia. It was included in the ten viruses that were said to be the most deadly in the world [1].

According to WHO's term in case of dengue virus infection, there are three diseases severity that can be suffered by patients [2]. They are Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF), and Dengue Shock Syndrome (DSS). Those diseases are caused by the dengue virus which is transmitted through the bite of mosquitoes Aedes aegypti or Aedes albopictus. DHF is a more severe level than DF. Plasma leakage or the occurrence of bleeding is a factor that can distinguish the DHF and DF. While DSS is the most dangerous type of dengue [3]. To enforce DF, DHF, and DSS diagnosis, it is necessary the laboratory examination. Hematokrit and thrombocyte levels are the main parameters used to determine the diagnosis on laboratory criteria levels [4].

Knowing dengue infection severity is important to performed as consideration the medical officer to determine the therapy and observation of the patient [5]. Lack of speed in handling or inappropriate handling with the severity of dengue infection can cause death of patients. So, prediction of dengue infection severity is needed because it pertains to the follow-up of patient handling. This is important to prepare the precise treatment according to the severity of patients in order to reduce the number of death from this disease. Therefore, we were interested in

International Conference on Mathematics, Computational Sciences and Statistics 2020
AIP Conf. Proc. 2329, 060021-1–060021-9; https://doi.org/10.1063/5.0042127
Published by AIP Publishing. 978-0-7354-4073-9/$30.00

finding the model classification of DF, DHF, and DSS based on the patient's blood examination using a statistical approach.

There are several methods in statistical approach. Because the patient's blood examination result is a multivariate dataset, so the authors used discriminant analysis approach. Discriminant analysis is one of the multivarat analysis techniques that aims to separate some groups with discriminant functions. Here, assumed that data come from multivariate normal distribution. When multivariate normal distribution assumptions have been met, then to estimate the parameters can be used the method of Maximum Likelihood (ML) [6]. This analysis can be called classic discriminant analysis. Unfortunately, the ML method is heavily influenced by outlier so the estimator becomes less precise when data has been contaminated by outliers. To overcome this problem, robust estimation methods can be used. One of robust estimation methods in discriminant analysis is Minimum Covariance Determinant (MCD) [7]. Therefore, in this research the authors discuss prediction of dengue infection severity using discriminant analysis with ML and MCD estimator.

REVIEW OF LITERATURES

Dengue Virus and Its Diagnosis Enforcement

Dengue virus is artropod borne virus (arbovirus) that belong to Flavivirus genus, flaviviridae family. Dengue virus is cause of dengue infection which transmitted by biting Aedes aegypti and Aedes albopictus mosquitoes. This disease can be divided into three severity levels, namely DF, DHF, and DSS [3].

On laboratory criteria, thrombocyte and hematocrit levels are used as parameter to determine the diagnosis of dengue infection. Decreasing of platelets number in the blood is called thrombocytopenia. Thrombocytopenia of dengue infection occur through the mechanism of bone marrow suppression and destruction as well as shortening the life span of platelets. The normal level of thrombocyte is 150,000 - 400,000/mm^3 [4].

Hematocrit is volume of erithrocyte cells in 100 mm^3 blood notated in percentage. Normal hematocrit levels for men and women are different. The normal ranges for hematocrit are depend on age and sex of the individual. The normal ranges for adult males are about 42% – 54%, while for adult females are about 38% – 46%. When suffering from dengue fever, the patient's hematocrit increased. However, after receiving fluid therapy, hematocrit levels of dengue patients typically have declined by more than 20% [8].

Multivariate Normal Distribution and Its Goodness of Fit Test

A vector of $\boldsymbol{x} = (x_1, x_2, \dots, x_p p$ is distributed Multivariate Normal with mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ if it has probability density function in the form [6]

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (1)$$

where usually notated by $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

A goodness of fit test for Multivariate Normal distribution based on skewness and kurtosis measurements was developed [9]. If $\bar{\boldsymbol{x}}$ and $\mathbf{S}$ are estimate of mean vector and covariance matrix, respectively, skewness and kurtosis are formulated by

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \mathbf{S}^{-1} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})\right)^3, \text{ and} \quad (2)$$

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\left((\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} \mathbf{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\right)^2, \quad (3)$$

respectively. The test statistic for skewness, $\frac{n}{6}(b_{1,p})$ is approximately χ^2 distributed with $p(p+1)(p+2)/6$ degrees of freedom. Similarly, the test statistic for kurtosis, $b_{2,p}$ is approximately Normally distributed with mean $p(p+2)$ and variance $8p(p+2)/n$. A correction term into the skewness test statistic was introduced, usually when $n < 20$, in order to control type I error [10]. The corrected skewness statistic for small samples is $\frac{nk}{6}(b_{1,p})$, where $k = (p+1)(n+1)(n+3)/(n(n+1)(p+1)-6)$. This statistic is also distributed as χ^2 with $p(p+1)(p+2)/6$ degrees of freedom.

Under the null hipothesis (H_0) that data comes from multivariate normal distribution, the null hipothesis is accepted when

$$\frac{n}{6}(b_{1,p}) \leq \chi^2_{\alpha,\frac{p(p+1)(p+2)}{6}}\ , \tag{4}$$

and

$$\frac{b_{2,p}-p(p+2)}{\sqrt{\frac{8p(p+2)}{n}}} \leq z_\alpha, \tag{5}$$

where $\chi^2_{\alpha,\frac{p(p+1)(p+2)}{6}}$ is the quantile $(1-\alpha)$% of Chi-Square distribution with degree of freedom of $\frac{p(p+1)(p+2)}{6}$, and z_α is the quantile $(1-\alpha)$% of Standard Normal distribution [11].

Discriminant Analysis

Discriminant analysis is one method in multivariate analysis that aims to separate some groups with a discriminant functions. Discriminant functions are able to describe the differences between groups through mathematical equations. Discriminant analysis was first discovered and developed later by Fisher. Suppose that x_{ijr} is the value of i^{th} observation, j^{th} group and r^{th} independent variable for $r = 1, 2, \cdots, p$, and $\boldsymbol{x_{ij}} = \begin{pmatrix} x_{ij1} & x_{ij2} & \cdots & x_{ijp} \end{pmatrix}$ is a vector of i^{th} observation from j^{th} group. Based on normality assumption, the linear discriminant score for the j^{th} group to be

$$d_j(\boldsymbol{x}) = \boldsymbol{\mu_j}^{\mathrm{T}}\boldsymbol{\Sigma^{-1}}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu_j}^{\mathrm{T}}\boldsymbol{\Sigma^{-1}}\boldsymbol{\mu_j} + \log p_j \text{ for } j = 1, 2, \ldots, k, \tag{6}$$

where $\boldsymbol{\mu}_j$ is the mean vector of j^{th} group, $\boldsymbol{\Sigma}$ is the general covariance matrix, and p_j is the prior probability of j^{th} group [6]. An object characterized by $\boldsymbol{x}$ is classified in to $j_0{}^{th}$ group if $d_{j_0}(\boldsymbol{x})=\max\{d_j(\boldsymbol{x})\}$; $j = 1,2, \ldots, k$. In practice, the classification rule is implemented by substituting the sample quantities $\bar{\boldsymbol{x}}_j$ and $\mathbf{S}$ for $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}$, respectively. Those sample quantities will be given in section 2.5 and 2.6 below. Whereas, the prior probability of j^{th} group can be used two options. Firstly, it is estimated by an equal value for each group, i.e. $\hat{p}_j = \frac{1}{k}$ for $j = 1,2, \ldots, k$. The second option is based on group size, i.e. $\hat{p}_j = \frac{n_j}{n}$, where n_j is the sample size of j^{th} group, and n is the total sample size, i.e. $n = \sum_{j=1}^{k} n_j$.

In fact, the assumption of similarity covariance matrix between groups for linear discriminant analysis is not always being met. Quadratic discriminant analysis is emerging as one of the alternative methods in discriminant analysis which used when the group follows a multivariate normal distribution with unequally covariance matrix between groups, by forming quadratic discriminant functions. The quadratic discriminant score for the j^{th} group as follows

$$d_j(\boldsymbol{x}) = -\tfrac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) + \log p_j, \text{ for } j = 1, 2, \ldots, k, \tag{7}$$

where $\boldsymbol{\mu}_j$ is the mean vector of j^{th} group, $\boldsymbol{\Sigma}_j$ is the covariance matrix of j^{th} group, and p_j is the prior probability of j^{th} group. Here, the sample quantities and classification rule are similarly with linear discriminant analysis. In addition, the sample quantities $\mathbf{S}_j$ is used to substitute $\boldsymbol{\Sigma}_j$ [6].

Parameter Estimation

Usually, mean vectors and covariance matrix for j^{th} group and the general covariance matrix above are estimated using Maximum Likelihood (ML) method. The ML's estimator for mean vectors are [6]

$$\bar{\boldsymbol{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}\text{, for } j = 1, 2, \ldots, k. \tag{8}$$

While, the general covariance matrix $\boldsymbol{\Sigma}$ is estimated by pooled covariance matrix as follow

$$\mathbf{S} = \frac{\sum_{j=1}^{k}(n_j - 1)\mathbf{S}_j}{n-k}, \tag{9}$$

where

$$\mathbf{S}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)^T, \text{ and } n = \sum_{j=1}^{k} n_j. \tag{10}$$

In here, the discriminant analysis with ML estimator is called classical or non robust discriminant analysis.

Minimum Covariance Determinant (MCD) is a robust estimation of mean vector and covariance matrix of multivariate data. The MCD algorithm estimates a mean vector and covariance matrix of multivariate data based on the determinant of the smallest covariance matrix. Suppose that a multivariate sample data X is viewed as a set of n observations with p variables, estimator of MCD will be searched based on a subset of X which consists of a number of h observations, where $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$, that is the integer part of the division of $(\boldsymbol{n} + \boldsymbol{p} + \boldsymbol{1})$ by 2. So, there is a number of combinations that should be inspected to obtain MCD estimator [7]. If the discriminant parameters are estimated by using a robust method, we called the robust discriminant analysis.

DATA AND METHODOLOGY

The data used in this research is results of blood tests and final diagnosis of dengue infection patients in 2017-2018. A number of 167 observations were obtained from Surabaya Hajj Hospital, East Java, Indonesia. They are 90 observations of male patients, and 77 observations of female patients. There were 67, 12, and 11 male patients with DF, DHF, and DSS levels, respectively. On the other hand, a number of 51, 17, and 9 female patients with DF, DHF, and DSS levels, respectively.

In this study, the dependent variable is the classification of dengue infection severity with code 1 for DF, 2 for DHF, and 3 for DSS. Whereas, the independent variables were age (x_1) in years, the decreasing of hematocrit levels at the first measurement after the patients hospitalized to the day later (x_2) in per mm^3, and the decreasing of thrombocyte levels at the first measurement after the patients hospitalized to the day later (x_3) in percentage.

Suppose that a multivariate sample data $\mathbf{X}$ is viewed as a set of n observations with p variables, estimator of MCD will be searched based on a subset of $\mathbf{X}$ which consists of a number of h observations, where $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$, that is the integer part of the division of $(n + p + 1)$ by 2. This is MCD algorithm which adopted from [7]

(1) Suppose that a sample of multivariate data consists of n observations and p variables with $p \leq n$, notated as $\mathbf{X} = (\boldsymbol{x}_1\ \boldsymbol{x}_2\ \ldots\ \boldsymbol{x}_n)^{\mathrm{T}}$, where $\boldsymbol{x}_i = [x_{i1}\ x_{i2}\ \ldots\ x_{ip}]$, and x_{ij} stated i^{th} observation of j^{th} variable, for $i = 1,2, \ldots, n$ and $j = 1,2, \ldots, p$.

(2) Determine a set of $H_m \subseteq \mathbf{X}$ for iteration index of m = 1, by taking a number of h observations randomly from n observations.

(3) Set $\mathbf{S}_m = \frac{1}{h}\sum_{\boldsymbol{x}_i \in H_m}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_m)^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_m)$ and $\bar{\boldsymbol{x}}_m = \frac{1}{h}\sum_{\boldsymbol{x}_i \in H_m}\boldsymbol{x}_i$.

(4) Compute the determinant of covariance matrix $\mathbf{S}_m$ notated $det(\mathbf{S}_m)$.

(5) Compute the Mahalanobis distance MD_i^2 for each observation $\boldsymbol{x}_i \in \mathbf{X}$, where $MD_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_m)\mathbf{S}_m^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_m)^T$ for $i = 1, \ldots, n$.

(6) Sort those Mahalanobis distances from step 5 with increasing order.

(7) Determine the H_{m+1} by taking a number of h observations according to the first h mahalanobis distances at step 6.

(8) Set $\mathbf{S}_{m+1} = \frac{1}{h}\sum_{\boldsymbol{x_i}\in H_{m+1}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{m+1})^T(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{m+1})$ and $\overline{\boldsymbol{x}}_{m+1} = \frac{1}{h}\sum_{\boldsymbol{x_i}\in H_{m+1}}\boldsymbol{x_i}$.
(9) Compute the determinant of covariance matrix $\mathbf{S}_{m+1}$, notated $det(\mathbf{S}_{m+1})$.
(10) If $det(\mathbf{S}_m) = det(\mathbf{S}_{m+1})$ then stop iteration and continue to step 11, otherwise repeat to step 5 with upgrade m=m+1.
(11) Found that $\overline{\boldsymbol{x}}_{MCD} = \overline{\boldsymbol{x}}_{m+1}$ and $\mathbf{S}_{MCD} = \mathbf{S}_{m+1}$.

Furthermore, the outliers can be detected by the following steps [7]:
(1) Compute the Mahalanobis distance MD_i^2 for each observation $\boldsymbol{x_i} \in \mathbf{X}$, where $MD_i^2 = (\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{MCD})\mathbf{S}_{MCD}^{-1}(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{MCD})^T$ for $i = 1, \ldots, n$
(2) If $MD_i^2 > \chi^2_{(0.025,p)}$ then $\boldsymbol{x_i}$ is outlier data, otherwise it is non outlier data.

In this paper, we performed the following research procedure:
(1) Drawn random sample.
(2) Disaggregate data by sex, i.e: male patients data and female patient data.
(3) Check the discriminant assumptions about normality, mean vector difference, and variance equality for male and female patient data.
(4) Separate data in to training data and validation data for male and female.
(5) Perform the quadratic discriminant analysis using ML and MCD estimator based on training data of male patients, and compute its classification accuracy respectively.
(6) Based on each model, predict classification on validation data and compute its cassification accuracy.
(7) Repeat step 5 and 6, but for female patients data.
(8) Conclude which one better between ML and MCD methods based on criteria of classification accuracy.

RESULTS AND DISCUSSION

Because of difference in normal measurements of hematocrit levels for male and female, data was disaggregated by sex. Then, observations were choosed randomly and separated again in the category of "Training data" and "Validation data". Male patients data was divided into Training and Validation data with 70 and 20 observations, respectively. While, the female patients data was divided into Training and Validation data with 62 and 15 observations, respectively. Among the male training data, there were 53 patients with DF severity, 10 patients with DHF severity, and 7 patients with DSS severity. Meanwhile, among the female training data, there were 41, 14, and 7 patients with DF, DHF, and DSS severity, respectively.

The multinormality assumption is tested firstly before analyzing the sample using discriminant analysis. Here, the multinormality of data was tested by using a method proposed by [9], and it has been used by [11] or [12]. Here, each group of dengue infection severity, i.e. DF, DHF, and DSS severity should be tested. The method uses two test statistic based on skewness and kurtosis. By using R programming, the values of test statistics for multinormality, i.e. left side of inequality (4) and (5) are available in Table 1. By using significance level of 5%, we find the critical values of Chi-square and Z are 18.307 and 1.645, respectively. Test decisions were taken by comparing test statistic of skewness and kurtosis with their critical values based on inequality (4) and (5) criteria. All of conclusions about multinormality of the data are available in the last column of Table 1. Based on Table 1, we concluded that all of groups in the male and female data followed multivariate normal distribution.

TABLE 1. The values of Multinormality test statistics for sample

Sex	Category	Test statistic of skewness	Test statistic of kurtosis	Decision
Male	1	17.923	-3.052	Accept H_0
	2	5.472	-2.182	Accept H_0
	3	9.894	-0.922	Accept H_0
Female	1	16.942	-1.797	Accept H_0
	2	12.829	-1.785	Accept H_0
	3	0.430	-3.105	Accept H_0

The condition that must be met in the discriminant analysis is the difference in mean vectors between groups. For this reason we tested about difference of mean vectors. The number of DF, DHF, and DSS in male patient data are 67, 12, and 11, respectively. Based on them, the result of some statistical tests for mean vector difference are

presented in Table 2. We can see that we have "Sig" values of 0.000 for all statistical methods, which means $p < 0.05$. Therefore we concluded that there were difference of mean vectors between groups in male patient data.

TABLE 2. Multivariate test of mean vectors for male patient data

Effect	Statistical Methods	Value	F	Hypothesis df	Error df	Sig
CATEGORY	Pillai's Trace	0.881	12.060	9	261.000	0.000
	Wilks' Lambda	0.233	18.821	9	207.018	0.000
	Hotelling's Trace	2.803	26.058	9	251.000	0.000
	Roy's Largest Root	2.624	76.108	3	87.000	0.000

In female patient dataset, the number of DF, DHF, and DSS are 51, 17, and 9, respectively. Using this data, the result of some statistical tests for mean vector difference are presented in Table 3. It can be seen that the "Sig" values are 0.000 for all statistical methods, which means $p < 0.05$. So, we concluded that there were difference of mean vectors between groups in female patient data.

TABLE 3. Multivariate test of mean vectors for female patient data

Effect	Statistical Methods	Value	F	Hypothesis df	Error df	Sig
CATEGORY	Pillai's Trace	0.807	9.085	9	222.000	0.000
	Wilks' Lambda	0.291	12.884	9	175.380	0.000
	Hotelling's Trace	2.101	16.497	9	212.000	0.000
	Roy's Largest Root	1.926	47.496	3	74.000	0.000

Another assumption that have to be tested is equality of covariance matrix across groups. The null hypothesis stated that the observed covariance matrix of the dependent variables are equal across groups. Result of this test for male and female patient data are presented in Table 4.

TABLE 4. Box's test of equality of covariance matrices

data	Box's M	F	df1	df2	Sig
male patient	41.354	3.071	12	3337.689	0.000
female patient	75.199	5.579	12	2584.825	0.000

With significance level $\alpha = 0.05$, the p-value of both tests were less than α. So, we rejected the null hypothesis and concluded that the observed covariance matrix of the dependent variables were unequal across groups for both of male patients and female patients data. Therefore, the appropriate analysis was quadratic discriminant as in the following section.

Quadratic Discriminant Analysis for The Male Patient data

Firstly, we performed quadratic discriminant analysis using ML estimator. We found that the classification accuracy rate for the training data and the validation data of male patients were 85.7% and 70%, respectively. Furthermore, there was 5.66% outlier data in DF level, and there were no outlier in others. The presence of outlier indicated the need for a robust method.

Next, we performed quadratic discriminant analysis using MCD estimator. For the male training data, It was found that the estimator of mean vector for each group using equation (8) and (10) were as follows

$$\begin{cases} \bar{\boldsymbol{x}}_1 = (15.6 \quad 2.575 \quad 6.675) \\ \bar{\boldsymbol{x}}_2 = (19.857 \quad -0.871 \quad 18{,}714.29) \\ \bar{\boldsymbol{x}}_3 = (13.8 \quad 4.5 \quad 7.800) \quad . \end{cases} \tag{11}$$

While, the ML estimator for covariance matrix of the dependent variables were obtained

$$\mathbf{S}_1 = \begin{bmatrix} 88.451 & 2.2 & -74210.26 \\ 2.2 & 5.845 & 26{,}678.85 \\ -74210.26 & 26{,}678.85 & 538{,}789{,}10\ 2.56 \end{bmatrix}, \tag{12}$$

$$\mathbf{S}_2 = \begin{bmatrix} 96.809 & 18.688 & -112{,}214.28\,6 \\ 18.688 & 6.002 & 4{,}209.524 \\ -112{,}214.28\,6 & 4{,}209.524 & 456{,}571{,}42\,8.571 \end{bmatrix}, \text{ and} \tag{13}$$

$$\mathbf{S}_3 = \begin{bmatrix} 92.2 & 37.0 & -236{,}050.0 \\ 37.0 & 57.935 & -336{,}025.0 \\ -236{,}050.0 & -336{,}025.0 & 1{,}958{,}700{,}000.0 \end{bmatrix}. \tag{14}$$

The estimators of prior probabilities based on group size were

$$\hat{p}_1 = 0.757,\ \hat{p}_2 = 0.143, \text{ and } \hat{p}_3 = 0.1. \tag{15}$$

In this sample, consider that $\boldsymbol{x} = (x_1, x_2, x_3)$. Based on equation (7) and by using those MCD estimators in equation (11) until equation (15), the quadratic discriminant scores for dengue infection severity classification of male patients data can be written by the following simple forms:

$$\begin{aligned} d_1(x_1, x_2, x_3) = {} & -15.123 + 0.193\,x_1 + 0.246\,x_2 + 2.675 \times 10^{-5}\,x_3 + \\ & 0.018\,x_1x_2 - 2.862 \times 10^{-6}\,x_1x_3 + 1.466 \times 10^{-5}\,x_2x_3 - \\ & 0.007\,x_1^2 - 0.122\,x_2^2 - 1.488 \times 10^{-9}\,x_3^2, \end{aligned} \tag{16}$$

$$\begin{aligned} d_2(x_1, x_2, x_3) = {} & -111.33 + 6.977\,x_1 - 23.250\,x_2 + 0.002\,x_3 + \\ & 0.827\,x_1x_2 - 6.907 \times 10^{-5}x_1x_3 + 2.3 \times 10^{-4}x_2x_3 - \\ & 0.125\,x_1^2 - 1.451\,x_2^2 - 1.064 \times 10^{-8}\,x_3^2, \end{aligned} \tag{17}$$

$$\begin{aligned} d_3(x_1, x_2, x_3) = {} & -244.439 + 4.006\,x_1 + 68.853\,x_2 + 0.012\,x_3 - \\ & 0.568\,x_1x_2 - 1.031 \times 10^{-4}x_1x_3 + 1.846 \times 10^{-3}x_2x_3 - \\ & 0.023x_1^2 - 5.179x_2^2 - 1.648 \times 10^{-7}x_3^2. \end{aligned} \tag{18}$$

Those quadratic discriminant scores in equation (16), (17), and (18) then applied to training and validation data of male patients. The classification accuracy rate using those quadratic discriminant scores for the training data and the validation data of male patients are 87.2% and 75%, respectively.

Quadratic Discriminant Analysis for The Female Patient data

Similarly, here we also performed quadratic discriminant analysis using ML estimator for female patient data. We found that the classification accuracy rate for the training data and the validation data of female patients were 80.7% and 66.7%, respectively. For outlier checking, there were 7.32% and 7.14% outlier data in DF and DHF levels, and there was no outlier in DSS level. These indicated the need of robust method.

Next, we performed robust quadratic discriminant analysis for the female patient data. The MCD estimators of mean vectors based on female training data were found as follows

$$\begin{cases} \bar{x}_1 = (13 \quad 1.026 \quad 6444.444) \\ \bar{x}_2 = (24.67 \quad 4.067 \quad 1000) \\ \bar{x}_3 = (7.2 \quad 3.34 \quad 63{,}000.0). \end{cases} \tag{19}$$

The MCD estimators of covariance matrix for each group were

$$\mathbf{S}_1 = \begin{bmatrix} 65.846 & 7.396 & -70,846.154 \\ 7.396 & 1.940 & -4,296.581 \\ -70,846.154 & -4,296.581 & 306,025,64\,1.026 \end{bmatrix}, \tag{20}$$

$$\mathbf{S}_2 = \begin{bmatrix} 127.750 & -13.162 & 73,750.0 \\ -13.162 & 1.850 & -13,875.0 \\ 73,750.0 & -13,875.0 & 150,000,00\,0.0 \end{bmatrix}, \text{ and} \tag{21}$$

$$\mathbf{S}_3 = \begin{bmatrix} 4.20 & -8.060 & 80,250.0 \\ -8.060 & 56.878 & -70,875.0 \\ 80,250.0 & -70,875.0 & 1,789,000,000.0 \end{bmatrix}. \tag{22}$$

The estimators of prior probabilities based on group size were

$$\hat{p}_1 = 0.66,\ \hat{p}_2 = 0.23, \text{ and. } \hat{p}_3 = 0.11. \tag{23}$$

In this case, by considering that $\boldsymbol{x} = (x_1, x_2, x_3)$, and entering those estimators above to equations (7), then obtained the discriminant scores for dengue infection severity classification of female patient data in the following simple forms:

$$\begin{aligned} d_1(x_1, x_2, x_3) = {} & -14.627 + 0.387\,x_1 - 0.723\,x_2 + 1.004 \times 10^{-4}\,x_3 + \\ & 0.123\,x_1x_2 - 6.645 \times 10^{-6}\,x_1x_3 + 1.516 \times 10^{-5}\,x_2x_3 - \\ & 0.018\,x_1^2 - 0.476\,x_2^2 - 2.296 \times 10^{-9}\,x_3^2, \end{aligned} \tag{24}$$

$$\begin{aligned} d_2(x_1, x_2, x_3) = {} & -149.144 + 3.443\,x_1 + 45.873\,x_2 + 0.003\,x_3 - \\ & 0.545\,x_1x_2 - 2.646 \times 10^{-5}x_1x_3 - 4.593 \times 10^{-4}x_2x_3 - \\ & 0.024\,x_1^2 - 3.931\,x_2^2 - 1.807 \times 10^{-8}\,x_3^2, \end{aligned} \tag{25}$$

$$\begin{aligned} d_3(x_1, x_2, x_3) = {} & -90.594 + 30.7\,x_1 + 2.879\,x_2 - 0.001\,x_3 - \\ & 0.565\,x_1x_2 + 2.586 \times 10^{-4}x_1x_3 + 2.261 \times 10^{-5}x_2x_3 - \\ & 3.132\,x_1^2 - 0.035\,x_2^2 - 5.632\ \times 10^{-9}x_3^2. \end{aligned} \tag{26}$$

Based on those estimators in (24), (25), and (26), we found that the classification accuracy rate for the training data and the validation data of female patients were 88.7% and 73.3%, respectively.

For ease in seeing the comparison of the results of both, here we show those results in Table 2.

TABLE 2. Percentage of classification accuracy

Data		Quadratic Discriminant	
		ML (%)	MCD (%)
Male	Training	85,7	87,2
	Validation	70	75
Female	Training	80,7	88,7
	Validation	66,7	73,3

Based on Table 2, it can be seen that the robust quadratic discriminant analysis with MCD algorithm was better than the other one. It can improve level of accuracy from ML method about 1.5% and 5% for the training and validation data of male patients, respectively. While, for female patients it can improve about 8% for the training data and 6.6% for the validation data.

In detection of outlier, the ability of MCD compared with ML for male and female patients can be shown in Table 3.

TABLE 3. Percentage of outlier in training data

Methods	Male Patients			Female Patients		
	DF	DHF	DSS	DF	DHF	DSS
ML (%)	5.66	0.00	0.00	7.32	7.14	0.00
MCD (%)	26.41	30.00	28.57	31.71	35.71	28.57

Based on Table 3, the average number of outlier in a group which is detected by MCD and ML are 30.16% and 3.35%, respectively. So, we can say that the MCD method has ability to detect outlier more than the ML method has.

CONCLUSION

Based on percentage of classification accuracy applied to dengue infection severity sample, we can conclude empirically that the quadratic discriminant analysis is more appropriate than the linear discriminant analysis. Furthermore, the MCD estimator is better than ML estimator in quadratic discriminant analysis. In addition, MCD estimator has a higher ability to detect outlier data. Furthermore, in order to classificate the dengue infection severity of new patients, we recommend for using the best discriminant analysis that have been generated in here, i.e. the quadratic discriminant analysis using robust MCD with the discriminant scores writed in equations (16)-(18) for male patient, and equations (24)-(26) for female patients.

REFERENCES

1. https://jakarta.tribunnews.com/2018/12/02/waspada-10-virus-ini-disebut-paling-mematikan-di-dunia, retrieved on 24 December 2018.
2. WHO, *Dengue Haemorrhagic Fever: Diagnosis, Treatment, Prevention, and Control*. *2nd edition* (World Health Organization, Geneva, 1987).
3. Ajlan BA, Alafif MM, Alawi MM, Akbar NA, Aldigs EK, Madani TA., PLoS Negl Trop Dis **13**(8): e0007144 https://doi.org/10.1371/journal.pntd.0007144 (2019).
4. Pusparini, Jurnal Kedokteran Trisakti, April – Juni, **23**(2), 51-56 (2004).
5. Sri Rejeki S Hadinegoro, Paediatr Int Child Health, May , **32**(s1): 33-38 (2012).
6. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, *6th Edition* (Pearson Education, Inc, United States of America, 2007).
7. Rousseeuw, P.J. and Van Driessen, K., Technometrics, **41**(3), 212 – 223 (1999).
8. http://www.emedicinehealth.com/ hematocrit_blood_test/page2_em.htm, retrieved on 21 December 2018.
9. Mardia, K.V., Biometrika, **57**(3), 519–530 (1970).
10. Mardia, K.V., Sankhya: The Indian Journal of Statistics, Series B (1960–2002), **36**(2), 115–128 (1974).
11. Von Eye, A. and Bogat, G.A., Psychology Science, **46**, 243 – 258 (2004).
12. Klar, B., Journal of Multivariate Analysis, **83**, 141-165 (2002).

Prediction of dengue infection severity using classic and robust discriminant approaches

ORIGINALITY REPORT

14%	11%	10%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	eprints.uad.ac.id Internet Source	1%
2	buscador.una.edu.ni Internet Source	1%
3	es.scribd.com Internet Source	1%
4	helmut.knaust.info Internet Source	1%
5	asian-efl-journal.com Internet Source	1%
6	statik.unesa.ac.id Internet Source	1%
7	doczz.net Internet Source	1%
8	www.semanticscholar.org Internet Source	1%
9	E. Andry Dwi Kurniawan, Fatmawati, Miswanto. "Modeling of global warming effect	<1%

	on the melting of polar ice caps with optimal control analysis", AIP Publishing, 2021 Publication	
10	Theory and Applications of Recent Robust Methods, 2004. Publication	<1 %
11	Submitted to University of Queensland Student Paper	<1 %
12	Apichai Khongphatthanayothin, Pornthep Lertsapcharoen, Pentip Supachokchaiwattana, Vidhawas La-orkhun et al. "Myocardial depression in dengue hemorrhagic fever: Prevalence and clinical description*", Pediatric Critical Care Medicine, 2007 Publication	<1 %
13	slideheaven.com Internet Source	<1 %
14	Mathematical and Statistical Methods for Actuarial Sciences and Finance, 2014. Publication	<1 %
15	g3journal.org Internet Source	<1 %
16	Submitted to Mississippi State Board for Community & Junior Colleges Student Paper	<1 %
17	labome.org Internet Source	<1 %

18	vdocuments.mx Internet Source	<1 %
19	Submitted to Victoria University Student Paper	<1 %
20	Eduardo Castaño-Tostado. "Small - sample correction factor of the minimum covariance determinant estimator", Communications in Statistics - Simulation and Computation, 2000 Publication	<1 %
21	article.sapub.org Internet Source	<1 %
22	Siti A. D. Safitri, Fajrina A. Putri, Belindha A. Ardhani, Nur Chamidah. "Co-Kriging method performance in estimating number of COVID-19 positive confirmed cases in East Java Province", AIP Publishing, 2021 Publication	<1 %
23	Wiggins, A.D.. "On the use of the directional derivative in obtaining multivariate extreme values", Statistics and Probability Letters, 198502 Publication	<1 %
24	www.science.gov Internet Source	<1 %
25	worldwidescience.org Internet Source	<1 %

26	www.frontiersin.org Internet Source	<1 %
27	Badi H. Baltagi. "Econometrics", Springer Science and Business Media LLC, 2021 Publication	<1 %
28	JOE H. SULLIVAN, WILLIAM H. WOODALL. "Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations", IIE Transactions, 2000 Publication	<1 %
29	Pitfalls in Diagnostic Radiology, 2015. Publication	<1 %
30	Sirianong Namwongprom, Surangrat Pongpan, Apichart Wisitwong, Chamaiporn Tawichasri, Jayanton Patumanond. "Validation of dengue infection severity score", Risk Management and Healthcare Policy, 2014 Publication	<1 %
31	kurser.math.su.se Internet Source	<1 %
32	nozdr.ru Internet Source	<1 %
33	www.tdx.cat Internet Source	<1 %

34	Miyamura, M.. "Robust Gaussian graphical modeling", Journal of Multivariate Analysis, 200608 Publication	<1%
35	P'ng Loke, Samantha N. Hammond, Jacqueline M. Leung, Charles C. Kim, Sajeev Batra, Crisanta Rocha, Angel Balmaseda, Eva Harris. "Gene Expression Patterns of Dengue Virus-Infected Children from Nicaragua Reveal a Distinct Signature of Increased Metabolism", PLoS Neglected Tropical Diseases, 2010 Publication	<1%
36	Tolmie, Andy, Muijs, Daniel, McAteer, Erica. "EBOOK: Quantitative Methods In Educational And Social Research Using Spss", EBOOK: Quantitative Methods In Educational And Social Research Using Spss, 2011 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Prediction of dengue infection severity using classic and robust discriminant approaches

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10