Annals of Data Science (2020) 7(4):697–712
https://doi.org/10.1007/s40745-020-00293-x

Copula Approach for Developing a Biomarker Panel for Prediction of Dengue Hemorrhagic Fever

Jong-Min Kim[1] · **Hyunsu Ju**[2] · **Yoonsung Jung**[3]

Received: 12 August 2019 / Revised: 9 May 2020 / Accepted: 21 May 2020 / Published online: 10 June 2020

Abstract

The choice of variable-selection methods to identify important variables for binary classification modeling is critical for producing stable statistical models that are interpretable, that generate accurate predictions, and have minimal bias. This work is motivated by the availability of data on clinical and laboratory features of dengue fever infections obtained from 51 individuals enrolled in a prospective observational study of acute human dengue infections. Our paper uses objective Bayesian method to identify important variables for dengue hemorrhagic fever (DHF) over the dengue data set. With the selected important variables by objective Bayesian method, we employ a Gaussian copula marginal regression model considering correlation error structure and a general method of semi-parametric Bayesian inference for Gaussian copula model to estimate, separately, the marginal distribution and dependence structure. We also carry out a receiver operating characteristic (ROC) analysis for the predictive model for DHF and compare our proposed model with the other models of Ju and Brasier (Variable selection methods for developing a biomarker panel for prediction of dengue hemorrhagic fever. BMC Res Notes 6:365, 2013) tested on the basis of the ROC analysis. Our results extend the previous models of DHF by suggesting that IL-10, Days Fever, Sex and Lymphocytes are the major features for predicting DHF on the basis of blood chemistries and cytokine measurements. In addition, the dependence structure of these Days Fever, Lymphocytes, IL-10 and Sex protein profiles associated with disease outcomes was discovered by the semi-parametric Bayesian Gaussian copula model and Gaussian partial correlation method.

✉ Yoonsung Jung
yojung@pvamu.edu

[1] Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN 56267, USA

[2] U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, USA

[3] Cooperative Agricultural Research Center, College of Agriculture and Human Sciences, Prairie View A&M University, Prairie View, TX 77446, USA

Keywords Copula · Variable selection · ROC · Biomarker · Dependence

1 Introduction

Acute dengue viral infections, hyperendemic to the tropics, are major cause of morbidity in tropical countries. There is no drug therapy or vaccine treatment now. Dengue hemorrhagic fever (DHF) is a late-stage complication of acute dengue infection primarily associated with capillary leakage, hemorrhage, circulatory shock, and representing life-threatening complications. Early detection of DHF may help identify patients that would benefit from more intensive therapy. Ahmad [2] proposed hyperbolic Sine Rayleigh distribution capable of modeling bladder cancer susceptibility data with unimodal failure rate function. Ju and Brasier [11] proposed variable selection methods of discriminative features that were identified and that were associated with DHF. Patients' profile of biomarkers such as blood chemistries and cytokine measurements can be associated with DHF outcome. However, correlations of these feature variables are known to be high, and which members of the complex panel of biomarkers will result in the most stable and robust classifier is not known. The variable selection methods proposed by Ju and Brasier [11] did not provide clear guidelines for predicting the development of DHF. The motivation of this paper is to provide a copula model [21] considering the dependence structure between correlated variables of the development of DHF with high predictive accuracy because copula is a useful device to express joint distributions of two or more random variables and explain the dependence structure between variables by eliminating the influence of the marginal distributions of the individual variables, and a copula function does also not require a normal distribution and independent, identical distribution assumptions. Furthermore, the invariance property of copula has been attractive which is especially in the finance area. The other reason is to provide semi-parametric Bayesian Gaussian copula model of Hoff [9] to confirm the result which can be interpretable for the important variable coefficients to DHF over the dengue data set. However, most copulas have a limitation which fails to satisfy the copula properties when extended from bivariate to multivariate cases. To overcome the limitation, Aasa et al. [1] proposed pair-copula constructions of multiple dependence. Gaussian bivariate copula uses the conditional distributions to find a partial correlation. The partial correlation coefficient has been measure by the Gaussian copula. To accomplish this research purpose, we employ the Bayarri et al. [3] objective Bayesian method, the probability of a proposition corresponds to a reasonable belief which can be justified by requirements of rationality and consistency, for identifying important variables to DHF over the dengue data set, and copula methods which are suitable for modeling highly correlated variables under a dependent structure regardless of the form of the marginal distributions.

Since Genest et al. [8] proposed a semiparametric estimation procedure of dependence parameters in multivariate families of distributions, copula methods have gotten more attention in the areas of finance, actuarial science, biomedical studies, and engineering. The copula method arises from Sklar's theorem [21], which allows researchers to piece together joint distributions with marginal distribution of

individual variables. Kim et al. [13] applied a copula method for modeling directional dependence of genes as an alternative method for a Bayesian network, which is a probabilistic graphical method. Kim and Kim [12] also proposed an improved copula method for modeling the directional dependence of genes. A copula determines the dependence relationship by joining the marginal distributions together to form a joint distribution. The scaling and the shape are entirely determined by the marginals. In contrast to correlation the copula function can be applied when variables are heavily tailed. Standard references for a detailed overview of copula applications include the books by Joe [10] and Nelsen [18]. We have now applied these new copula methods to investigate the relationship with the development of DHF and important biomarker variables that will be obtained by the Bayarri et al. [3] objective Bayesian method.

The remainder of this paper is organized as follows. Section 2 presents the Bayesian variable selection to find important variables to DHF over the dengue data set. Section 3 discusses Gaussian copula marginal regression models, semi-parametric Bayesian Gaussian copula estimation and Gaussian copula partial correlation with the selected important variables. Our conclusions are presented in last section.

2 Statistical Method for Variable Selection

In order to improve the performance of copula methods proposed in Sect. 3, we seek to reduce the dimensionality of dengue data set. Feature selection is a machine learning technique in data mining, which reduces the number of input variables when developing a predictive model [19, 20]. Feature reduction removes meaningless features which are not related to a studied disease, leading to overfitting of our proposed model with the dengue data set. The previous work by Ju and Brasier [11] tried to identify biomarker variables that are associated with disease outcome could predict the development of DHF. Ju and Brasier [11] found that the important variables identified to DHF from the dengue data set were Interleukin-6 (IL-6), Interleukin-10 (IL-10) and Platelets through five different feature reduction classification methods, including generalized path seeker, multivariate adaptive regression splines, TreeNet, Boosting, and Random Forest and Bayesian moving averaging method identified three variables—IL-10, Lymphocytes, and Platelets having a high probability of predicting DHF. From the results, Ju and Brasier [11] suggested that IL-10, Platelets, and Lymphocytes counts are the major features for predicting DHF on the basis of blood chemistries and cytokine measurements. In this study, we also used the same dengue data set which can be downloaded from the NIAID Clinical Proteomics Center Web site at https://bioinfo.utmb.edu/CPC/Projects/default.jsp.

The data are 51 dengue infected subjects identified at participating clinics and hospitals, or at a community-based active surveillance study in Maracay, Venezuela. Applying the 2009 World Health Organization (WHO) criteria, it was revealed that 13 subjects had developed dengue hemorrhagic fever. So the laboratory values of 51 individuals are 38 dengue fever (DF) and 13 DHF. The previous works by Brasier et al. [4] and Ju and Brasier [11] motivate us to identify biomarker variables that are necessary for predicting the development of DHF

with high probability. Bayarri et al. [3] proposed criteria in determining objective model selection priors by considering their application to the problem of variable selection in normal linear models and obtained the methodological results in a new model selection objective prior with a number of compelling properties. Based on Bayarri et al. [3], Garcia-Donato and Forte [6] developed R package *BayesVarSel* conceiving to calculate Bayes factors, model choice and variable selection in linear models. To determine the optimal model for important variable selection of the dengue data set, we run a Gibbs sampling on the data set by using R package *BayesVarSel*. The GibbsBvs command in the *BayesVarSel* package provides approximate computation of summaries of the posterior distribution using a Gibbs sampling algorithm to explore the model space, and frequency of visits to construct the estimates. We set the possible prior distribution for regression parameters within each model as constant and set the possible prior distribution over the model space as gZellner which corresponds to the *g*-prior probability distribution function in Zellner [23]. The number of iterations is 100,000 times after the 3000 number of iterations at the beginning of the Markov Chain Monte Carlo (MCMC) that are thrown away. The most complex model has 26 covariates, plus the intercept, so there are a total of 67,108,864 competing models. All these models are kept and are used in the estimates. Table 1 shows the result of the inclusion probabilities of 26 feature variables in the data set. We can find a meaning result that Sex has the highest inclusion probability (0.91745), followed by Interleukin-10 (IL-10, 0.9137), Lymphocytes (0.66225) and Days Fever (0.63847). The highest posterior probability model and median probability model in Table 1 indicate significant features in the Bayesian model selection. We identified Days Fever, IL-10, Lymphocytes, and Sex as important variables to DHF. Figure 1 verifies the result from Table 1.

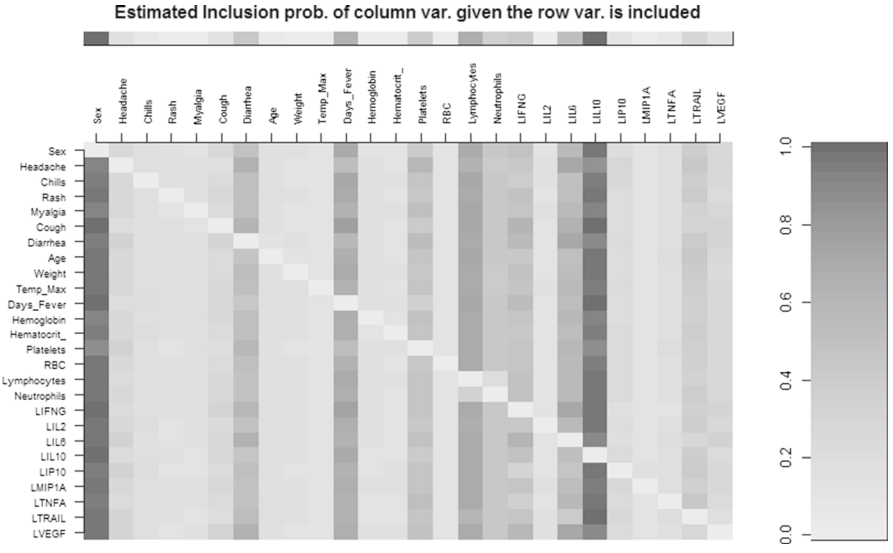

Fig. 1 A Bayesian variable selection plot for DHF

Table 1 Bayesian variable selection

	Incl. prob.	Highest posterior probability	Median posterior probability
Intercept	1.00000	*	*
Sex	**0.91745**	*	*
Headache	0.25941		
Chills	0.18360		
Rash	0.14631		
Myalgia	0.15174		
Cough	0.24120		
Diarrhea	0.48210		
Age	0.15691		
Weight	0.14867		
Temperature Max	0.13248		
Days Fever	**0.63847**	*	*
Hemoglobin	0.14742		
Hematocrit	0.14381		
Platelets	0.44191		
RBC	0.13603		
Lymphocytes	**0.66225**	*	*
Neutrophils	0.41384		
IFNG	0.45733		
IL-2	0.14376		
IL-6	0.53249		*
IL-10	**0.91370**	*	*
IP10	0.22165		
MIP1A	0.13333		
TNFA	0.18261		
TRAIL	0.36322		
VEGF	0.24498		

Bold values indicates Correlation between Class and Variable

*Significant biomarker to DHF

3 Copula Methods

A copula is a multivariate distribution function defined on the unit $[0, 1]^n$, with uniformly distributed marginals. In this paper, we focus on a bivariate (two-dimensional) copula, where $n = 2$. Sklar [21] shows that any bivariate distribution function, $F_{XY}(x, y)$, can be represented as a function of its marginal distribution of X and Y, $F_X(x)$ and $F_Y(y)$, by using a two-dimensional copula $C(\cdot, \cdot)$. More specifically, the copula may be written as

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)) = C(u, v),$$

where u and v are the continuous empirical marginal distribution functions $F_X(x)$ and $F_Y(y)$, respectively. Note that u and v have an uniform distribution $U(0, 1)$.

Therefore, the copula function represents how the function, $F_{XY}(x, y)$, is coupled with its marginal distribution functions, $F_X(x)$ and $F_Y(y)$. Using copula is the general way to describes the dependence mechanism between correlated random variables by eliminating the influence of the marginals or any monotone transformation of the marginals. It also be useful for constructing joint distributions, especially with non-normal random variables.

Let X, Y be random variables with continuous distribution functions $F_X(x)$ and $F_Y(y)$, respectively, let X and Y be continuous random variables with copula C and marginal distribution functions $F_X(x)$ and $F_Y(y)$ so that $X \sim F_X(x)$, $Y \sim F_Y(y)$, and $(X, Y) \sim F_{XY}(x, y)$, and let $u = F_X(x)$, $v = F_Y(y)$, and $(u, v) \sim C$. Then Spearman's ρ and Kendall's τ are given, respectively, by

$$\rho_C = 12 \int_0^1 \int_0^1 \Big[C(u, v) - uv\Big] dudv,$$

and

$$\tau_C = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1$$

[18].

3.1 Gaussian Copula Marginal Regression Model

To provide a stable statistical model considering correlation error structure with important variables (Days Fever, IL-10, Lymphocytes, and Sex) to DHF by objective Bayesian variable selection method, we used the Gaussian copula marginal regression models described by Song [22] and Masarotto and Varin [17], because such models provide a flexible general framework for modeling dependent responses of any type. Since we have mixed binary and continuous data types in our data, we have determined that the Gaussian copula marginal regression (GCMR) model is suitable for our analysis. The R package *gcmr* in Masarotto and Varin [17] fits Gaussian copula marginal regression models. Inference is performed through a likelihood approach. Computation of the exact likelihood is possible only for continuous responses. Otherwise, the likelihood function is approximated by importance sampling. See Masarotto and Varin [17] for details.

We compared three different models, as shown in Table 2. Model 1 consists of the important variables (IL-6, IL-10, Platelets) identified by learning ensemble, Model 2 with the important variables (IL-10, Lymphocytes, and Platelets) identified by Bayesian moving averaging method, and Model 3 with the important variables (Days Fever, IL-10, Lymphocytes and Sex) found in Sect. 2. Ju and Brasier [11] proposed Models 1 and 2, and suggested that the important variables (Days Fever, IL-10, Lymphocytes and Sex) in Model 2 are the major features for predicting DHF on the basis of blood chemistries and cytokine measurements.

Table 2 Proposed model vs. two models by Ju and Brasier [11]

Model	Variables
Model 1	IL-6, IL-10, Platelets
Model 2	IL-10, Lymphocytes, Platelets
Model 3 (Proposed Model)	Days Fever, IL-10, Lymphocytes, Sex

Bold values indicates Correlation between Class and Variable

Based on Table 2, we compared three models considering correlation structure in Table 3. Model 3 was the best of the these three models in terms of the Akaike information criterion (AIC). To find the optimal residual error structure between variables on Table 2, ARMA-type process has been tested and the optimal dependence structure was a ARMA(0, 0) for Model 3, meaning that no correlation structure was needed. This allows us to apply the GCMR model without auto-correlation structure to the important variables (Days Fever, IL-10, Lymphocytes and Sex) in Sect. 2. The result shown in Table 3 is very different from the results from the variable selection methods described by Ju and Brasier [11] because Model 3 including Sex (male and female) and Days Fever are better models than Models 1 and 2 without the Sex and Days Fever variables.

Logistic regression is appropriate when the response variable is categorical with two possible outcomes (DF, DHF) in Class target variable over the dengue data set. Binary variables can be represented using an indicator variable Y_i, taking on values 0 (DH) or 1 (DHF), and modeled using a binomial distribution with probability $P(Y_i = 1) = \pi_i$, where $i = 1, 2, \ldots, 51$. Logistic regression using GCMR with ARMA(0, 0) for Class (DH, DHF) models this probability as a function of Days Fever, IL-10, Lymphocytes and Sex explanatory variables. We are interested in determining the probability that a dengue infected subject develops DHF given the subject's selected information (Days Fever, IL-10, Lymphocytes and Sex). According to the output in Table 4, the model is

$$\begin{aligned} logit(\pi_i) = &-71.479 + 49.366 Days\,Fever + 63.866 IL\text{-}10 \\ &- 19.709 Lymphocytes + 23.403 Sex + \epsilon_i, \end{aligned}$$

where the residual, ϵ_i, follows the normal distribution with mean zero and variance, σ_ϵ^2 and ϵ_i does not need a correlation structure for the GCMR model with important variables (Sex, Days Fever, Lymphocytes, and IL-10) for the class (DH, DHF). To test $H_0 : \beta_1 = 0$, we use z = 8.253 (p-value < 2e−16).

Hence, a dengue infected subject's Days Fever information appears to have a significant positive impact on the probability of developing DHF, while

Table 3 Model comparison of GCMR with correlation structure

AIC	ARMA(0, 0)	ARMA(1, 0)	ARMA(0, 1)	ARMA(1,1)
Model 1	36.4825	37.4079	38.2078	40.0747
Model 2	29.1127	31.0996	28.7083	32.9836
Model 3	**18.4654**	20.4517	20.2364	22.1141

Bold values indicates Correlation between Class and Variable

Table 4 Logistic regression using GCMR with ARMA(0, 0) for Class (DH, DHF)

Variable	Estimate	Std. error	Z-value	p-value
(Intercept)	− 71.479	4.167	− 17.153	< 2e−16
Days Fever	49.366	5.982	8.253	< 2e−16
IL-10	63.866	5.652	11.299	< 2e−16
Lymphocytes	− 19.709	5.278	− 3.734	0.000189
Sex	23.403	2.777	8.427	< 2e−16

controlling for IL-10, Lymphocytes and Sex. To test $H_0 : \beta_2 = 0$, we use z = 11.299 (p-value = 0.0000). The dengue infected subject develops to DHF given the subject's IL-10 information appears to have a significant positive impact on the probability of developing DHF, while controlling for Lymphocytes and Sex but including Days Fever. To test $H_0 : \beta_3 = 0$, we use z = −3.734 (p-value = 0.0002). The dengue infected subject develops to DHF given the subject's Lymphocytes information appears to have a significant negative impact on the probability of developing DHF, while controlling for Sex but including Days Fever and IL-10. To test $H_0 : \beta_4 = 0$, we use z = 8.427 (p-value = 0.0000). The dengue infected subject develops to DHF given the subject's Sex information appears to have a significant positive impact on the probability of developing DHF, including Days Fever, IL-10 and Lymphocytes information. So all variables in Model 3 are statistically significant based on the estimates and standard errors generated by the GCMR with ARMA(0, 0) correlation structure. Furthermore, Days Fever, IL-10, Lymphocytes and Sex are significantly positive influence on DHF and Lymphocytes is significantly negative effect on DHF. It is the most meaningful result made by using our GCMR model.

We also look at the odds ratio (OR) corresponding to Days Fever is 2.75×10^{21} [95% CI (2.23×10^{16}, 3.40×10^{26})] in Table 5. This implies that increasing Days Fever by one unit will increase the odds of developing DHF significantly if we fix IL-10, Lymphocytes and Sex. Likewise, increasing LIL10 or Sex by one unit will increase the odds of developing DHF significantly. But increasing Lymphocytes by one unit will decrease the odds of developing DHF significantly if we fix Days Fever, IL-10 and Lymphocytes.

The best possible test for DHF can be chosen based on the sensitivity, specificity and accuracy. These are widely used to describe a diagnostic test. Sensitivity is defined as the rate of true positives that are correctly identified, and specificity is defined as the rate of true negatives that are correctly identified. Accuracy measures

Table 5 Odds ratio (OR) and 95% confidence interval for OR

	OR	2.5%	97.5%
Intercept	9.06×10^{-32}	2.57×10^{-35}	3.19×10^{-28}
Days Fever	2.75×10^{21}	2.23×10^{16}	3.40×10^{26}
LIL10	5.45×10^{27}	8.42×10^{22}	3.53×10^{32}
Lymphocytes	2.76×10^{-9}	8.86×10^{-14}	8.58×10^{-5}
Sex	1.46×10^{10}	6.31×10^{7}	3.37×10^{12}

how correct a diagnostic test identifies and can be determined from sensitivity and specificity with the presence of prevalence. A receiver operating characteristic (ROC) curve is a graphical presentation of the relationship between both sensitivity and specificity and is created by plotting the true positive rate against the false positive rate at various threshold settings.

An area under the curve (AUC) is used in classification analysis in order to determine which of the used models predicts the classes best. The closer AUC for a model comes to 1, the better it is. So models with higher AUCs are preferred over those with lower AUCs. So we performed ROC analysis which can illustrate the performance of a binary classification by our proposed copula modeling. Table 6 shows the summary of accuracy, sensitivity, specificity and AUC with the three different models for Class (DHF, DH) with Days Fever, IL-10, Lymphocytes and Sex. Our proposed model, Model 3, in Table 6 shows that the accuracy is about 94%, the sensitivity is about 97%, the specificity is about 85% and the AUC is about 91%. Therefore, the accuracy, sensitivity, specificity and AUC of the proposed model are fairly better than other two models (Models 1 and 2).

The test of accuracy was performed by binomial test so that the p-value (0.0003045) of the test in Model 3 is statistically significant. Therefore we can say that our proposed model is a good prediction model for DHF over the dengue data set. Figure 2 shows ROC curves for three different models. The straight line in the ROC curve of Model 3 is $y = x$, which passes through (0, 0) and (1, 1), so the ROC curve are far above the straight line compared to the models by Ju and Brasier [11]. This means that the proposed model illustrates the great performance of a binary classifier.

3.2 Semi-parametric Bayesian Gaussian Copula Model

We want to confirm our results in Sect. 3.1 with a different copula method. Hoff [9] provided the semi-parametric inference for copula models via a type of rank-likelihood function for the association parameters. The semi-parametric inference is based on a generalization of marginal likelihood, called an extended rank likelihood, that does not depend on the univariate marginal distributions of the data.

Table 6 Accuracy, sensitivity, specificity and AUC

		Model 1	Model 2	**Model 3**
Accuracy		0.8627	0.902	**0.9412**
Accuracy	95% CI	(0.7374, 0.943)	(0.7859, 0.9674)	**(0.8376, 0.9877)**
Accuracy	p-value	0.03277	0.004691	**0.0003045**
Sensitivity		0.9211	0.9474	**0.9737**
Specificity		0.6923	0.7692	**0.8462**
Positive predictive value		0.8974	0.9231	**0.9487**
Negative predictive value		0.75	0.8333	**0.9167**
AUC		0.8067	0.8583	**0.9099**

Bold values indicates Correlation between Class and Variable

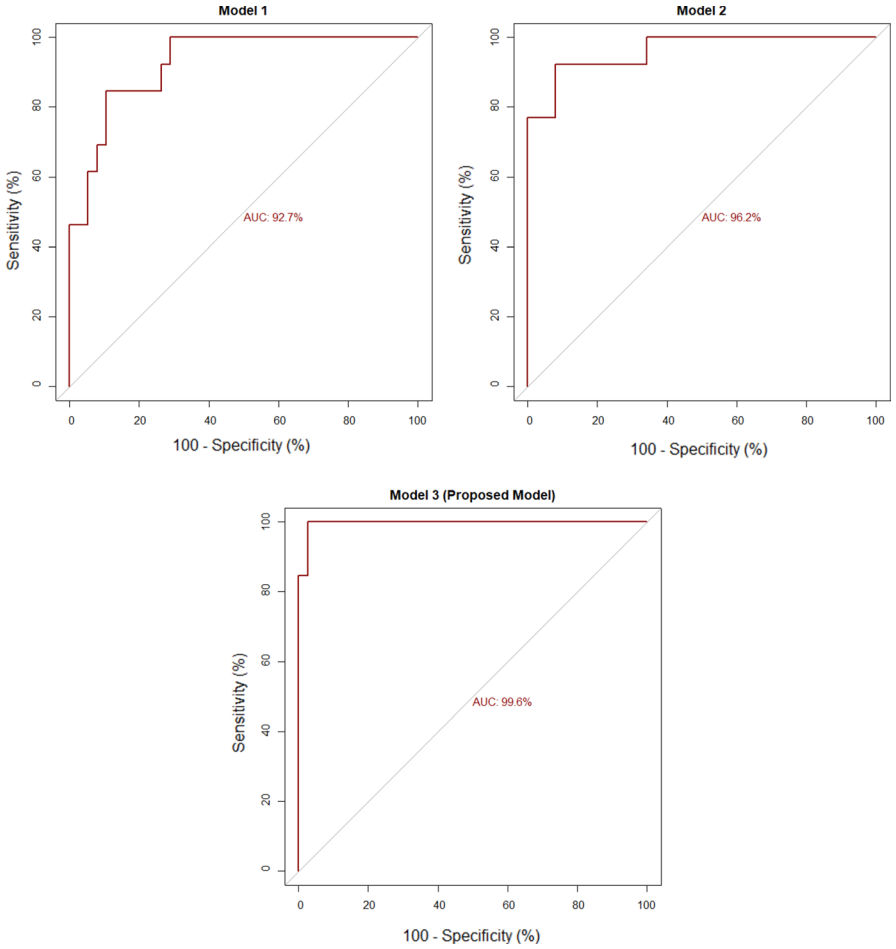

Fig. 2 ROC analysis. Shown are ROC curves for the three predictive models for DHF. Y axis, sensitivity; X axis, 1− specificity

Estimation and inference for parameters of the Gaussian copula are available via a straightforward Markov Chain Monte Carlo algorithm based on Gibbs sampling. Specification of prior distributions or a parametric form for the univariate marginal distributions of the data is not necessary. With the selected important dengue data set (Days Fever, IL-10, Lymphocytes, and Sex) and the Hoff [9] model we want to look at the relationship between Class (DF and DHF) and one of Days Fever, IL-10, Lymphocytes, and Sex is positive, negative or zero. By using the semi-parametric Bayesian Gaussian copula estimation, we made posterior quantiles of regression coefficients for Days Fever, IL-10, Lymphocytes, and Sex after the number of iterations was 25,000 times (also see Table 7). If we look at the 50% quantile of regression coefficients in Table 7, then we found that the estimates in Table 7 are consistent with the estimates in Table 4 if the scale of the

Table 7 Posterior quantiles of regression coefficients

	2.5% Quantile	50% Quantile	97.5% Quantile
Days Fever	-0.49	1.14	2.70
IL-10	0.86	2.13	3.50
Lymphocytes	− 2.92	− 1.35	0.22
Sex	0.14	1.08	2.09

estimates is standardized. We also wanted to see the relationship between Class and Sex (Males and Females) with each one of the variables (Days Fever, Lymphocytes, and IL-10). Figure 3 displays meaningful results based on Sex. The IL-10 versus Sex plots in Fig. 3 show that IL-10 is the most important variable for classifying DF and DHF and more females develop DHF. Days Fever versus Sex plots in Fig. 3 also show that Days Fever is an important variable to use to classify DF and DHF, and shows that more females develop DHF. Lymphocytes versus Sex plots in Fig. 3 show that Lymphocytes is an important variable for classifying DF and DHF, and indicate that more females develop DHF. In addition, we find that the classification using IL-10 is more efficient than the classifications using Days Fever and Lymphocytes.

3.3 Gaussian Copula Partial Correlation

In this paper, we want to apply the Kim et al. [14] Gaussian copula partial correlation to dengue infection data to see the dependence structure. Given an n-dimensional distribution function F with continuous marginal (cumulative) distributions $F_1, \ldots, F_n$, there exists a unique n-copula $C : [0,1]^n \to [0,1]$ such that

$$F(x_1, \ldots, x_n) = C(F(x_1), \ldots, F(x_n)).$$

Suppose Y and Z are real-valued random variables with conditional distribution functions

$$F_{2|1}(y|x) = p(Y \le y|X = x) \quad \text{and} \quad F_{3|1}(z|x) = p(Z \le z|X = x).$$

Then the basic property of

$$U = F_{2|1}(Y|X) \quad \text{and} \quad V = F_{3|1}(Z|X)$$

is as follows: suppose, for all x, $F_{2|1}(y|x)$ is continuous in y and $F_{3|1}(z|x)$ is continuous in z. Then U and V have uniform marginal distributions. Likewise, if $X_1, \ldots, X_n$ is a vector of n random variables with absolutely continuous multivariate distribution function F, then the n random variables

$$U_1 = F_1(X_1), \quad U_2 = F_{2|1}(X_2|X_1), \ldots, U_n = F_{n|1,2,\ldots,n-1}(X_n|X_1, \ldots, X_{n-1})$$

are *i.i.d.* $U(0, 1)$.

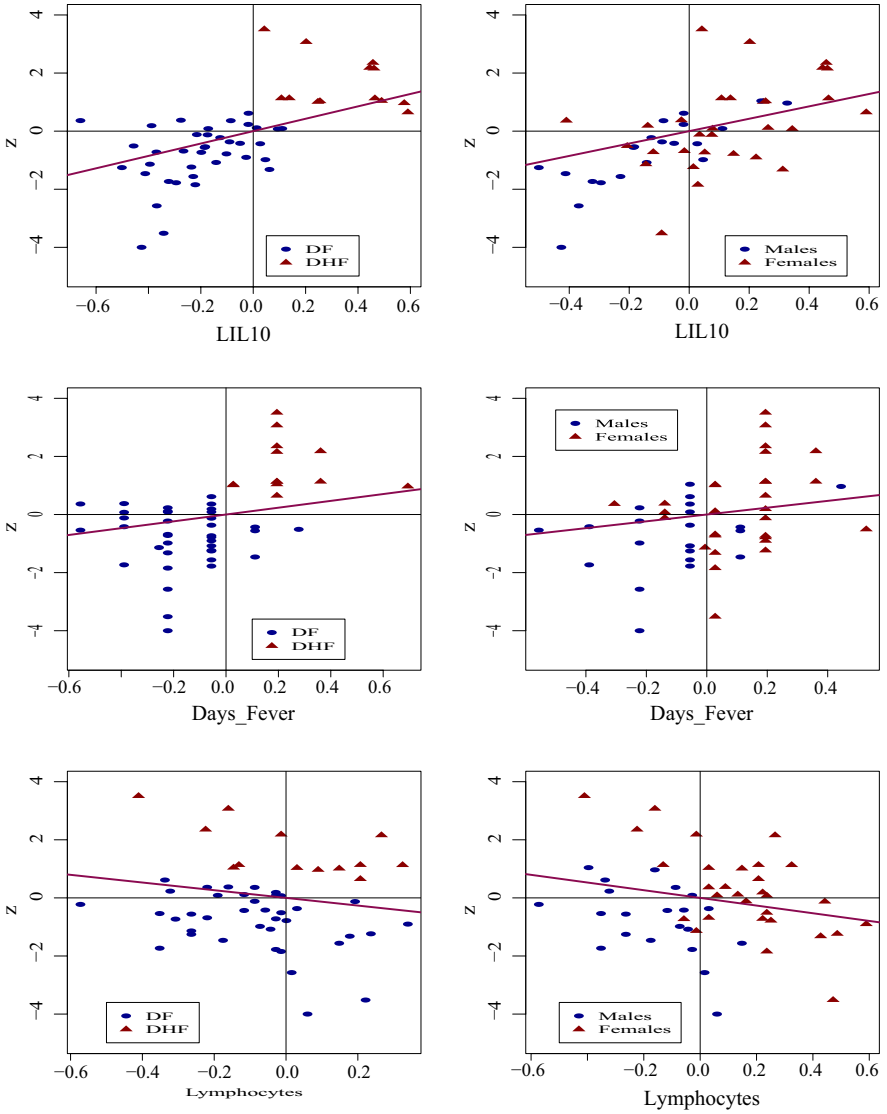

Fig. 3 IL-10, Days Fever and Lymphocytes versus Sex plots

The conditional distribution of $\mathbf{Z_1}$ given $\mathbf{Z_2}$ is also normal with mean vector

$$\mathbf{v_1} = \boldsymbol{\mu_1} + \Sigma_{12}{\Sigma_{22}}^{-1}\left(\mathbf{z_2} - \mu_2\right)$$

and covariance matrix

$$\mathbf{Q_1} = \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21}.$$

It follows that the conditional density function $f_{1|2}(\cdot|\mathbf{z_2})$ of $\mathbf{Z_1}$, when $\mathbf{Z_2} = \mathbf{z_2}$, is specified at the point $\mathbf{z_1}$ by the equation

$$f_{1|2}(\mathbf{z_1}|\mathbf{z_2}) = \frac{f(\mathbf{z_1}, \mathbf{z_2})}{f(\mathbf{z_2})} = \left(\frac{1}{2\pi}\right)^{p/2} \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \exp\left\{ -\frac{(\mathbf{z_1} - \mathbf{v_1})^T \mathbf{Q_1}^{-1} (\mathbf{z_1} - \mathbf{v_1})}{2} \right\}.$$

The cumulative distribution function is

$$F_{1|2}(\mathbf{z_1}|\mathbf{z_2}) = \int_{-\infty}^{z_p} \cdots \int_{-\infty}^{z_1} f_{1|2}(\mathbf{x_1}|\mathbf{z_2}) dx_1 \cdots dx_p \tag{1}$$

where $\mathbf{z_1} = (z_1, \ldots, z_p)$ and $z_1, \ldots, z_p \in \mathbf{R}$.

By using the Eq. (1), we can derive the Gaussian conditional distributions, and then by using the CML method by Genest et al. [8] and the IFM method by Joe [10], we can estimate the Gaussian copula parameter, a n-th order conditional correlation, $\rho_{YX|Z_1,Z_2,\ldots,Z_n}$, using the following:

$$F_{xy|z_1,\ldots,z_n}(Y, X|Z_1, Z_2, \ldots, Z_n) = C^{Ga}(F_{x|z_1,\ldots,z_n}(X|Z_1, Z_2, \ldots, Z_n), F_{y|z_1,\ldots,z_n}(Y|Z_1, Z_2, \ldots, Z_n); \rho_{YX|Z_1,Z_2,\ldots,Z_n}).$$

To use the Gaussian partial correlation to dengue infection data, we needed to transform Class and Sex variables from binary data type to continuous data type. Before obtaining the ranks for each variable, however, notice that the Class and Sex variables are discrete variables valued in a subset of the set of the binary integers, and there are ties in the data. Note that the presence of ties in the data may substantially affect the copula estimation Genest and Nešlehová [7]. In order to fully facilitate the good properties of the copula, one needs to deal with the discreteness of the variable and the presence of ties in an appropriate way.

Thus, we obtained a continuous extension of the Class and Sex variables by using the randomization technique proposed by Denuit and Lambert [5]: given integer-valued X_i, consider a continuous random variable $X_i^\star = X_i + (U_i - 0.001)$ where U_i is uniform on (0, 0.001) and independent of X_i in Table 8. As shown in Denuit and Lambert [5] and Madsen and Fang [16], the original variable can be recovered

Table 8 Descriptive statistics for Class and Sex from 51 subjects, and the continuous version of Class and Sex

	Class	Continuous version of Class	Sex	Continuous version of Sex
Mean	0.2549	0.2544	0.5882	0.5877
Median	0.0000	− 0.0004	1.0000	0.9991
Standard deviation	0.4401	0.4402	0.4971	0.4970
Skewness	1.0919	1.1248	− 0.3481	− 0.3586

from its continuous extension, and the distribution function of the original variable is exactly the same as that of its continuous extension. Furthermore, this approach randomly breaks the ties in the data. Kojadinovic and Yan [15] verified that the randomization (designed to randomly break the ties) does not change the results for the copula inference. Table 9 shows the Gaussian copula correlation between Class and each of the important variables (Sex, Days Fever, Lymphocytes and IL-10). The value of the Gaussian copula correlation between Class and IL-10 is higher (0.3862) and the value of the Gaussian copula correlation between Class and Lymphocytes is lower (−0.2206) when compared with other relationships, as shown in Table 9.

Table 10 shows the Gaussian copula partial correlation between Class and one of the important variables (Sex, Days Fever, Lymphocytes and IL-10) given the other variable. The value of Gaussian copula partial correlation between Class and Lymphocytes given on the fixed variable, Days Fever, in Table 10 is −0.2426. It is lower than the value of Gaussian copula correlation between Class and Lymphocytes in Table 9. The value of Gaussian copula partial correlation between Class and IL-10 given on the fixed variable (Days Fever or Lymphocytes or Sex) in Table 10 is lower than the value of Gaussian copula correlation between Class and IL-10 in Table 9.

Table 11 shows the Gaussian copula partial correlation between Class and one of the important variables (Sex, Days Fever, Lymphocytes and IL-10) given the other two variables. The value of the Gaussian copula partial correlation between

Table 9 Gaussian copula correlation

	Sex	Days Fever	Lymphocytes	IL-10
Class	0.2615	0.3448	**−0.2206**	**0.3862**

Table 10 Gaussian copula partial correlation conditioning on one variable

	Sex	Days Fever	Lymphocytes	IL-10
Class, Sex		0.2794	0.2738	0.2694
Class, Days Fever	0.3519		0.361	0.2505
Class, Lymphocytes	− 0.2	**− 0.2426**		− 0.1465
Class, IL-10	**0.331**	**0.2625**	**0.3011**	

Bold values indicates Correlation between Class and Variable

Table 11 Gaussian copula partial correlation conditioning on two variables

	(S, D)	(S, L)	(S, IL-10)	(D, L)	(D, IL-10)	(L, IL-10)
Class, Sex				0.3294	0.2989	0.3208
Class, Days Fever		**0.4003**	0.3025			0.3053
Class, Lymphocytes	**− 0.2673**		− 0.1803		− 0.2028	
Class, IL-10	0.2526	0.3077		**0.2118**		

Bold values indicates Correlation between Class and Variable

S Sex, *D* Days Fever, *L* Lymphocytes

Table 12 Gaussian copula partial correlation conditioning on three variables

	(S, D, L)	(S, D, IL-10)	(S, L, IL-10)	(D, L, IL-10)
(Class, Sex)				0.3738
(Class, Days Fever)			0.3783	
(Class, Lymphocytes)		**−0.2476**		
(Class, IL-10)	0.2262			

S Sex, *D* Days Fever, *L* Lymphocytes

Class and Days Fever given the fixed variables, (Sex and Lymphocytes), is the highest (0.4003) when compared with the other relationships shown in subsection. The value of Gaussian copula partial correlation between Class and Lymphocytes given on the fixed variables, (Sex and Days Fever), in Table 11 is the lowest (−0.2673) when compared with the other relationships shown in subsection.

The value of Gaussian copula partial correlation between Class and IL-10 given on the fixed variables (Days Fever and Lymphocytes) in Table 11 is lower than the value of Gaussian copula correlation between Class and IL-10 given on the fixed variable (Days Fever or Lymphocytes or Sex) in Table 10.

Table 12 shows the Gaussian copula partial correlation between Class and one of the important variables (Sex, Days Fever, Lymphocytes and IL-10) given the other three variables. The value of the Gaussian copula partial correlation between Class and Days Fever given on Sex, Lymphocytes and IL-10 is the higher compared with the other relationships in Table 12.

4 Results and Discussion

In this study, we found that IL-10, Days Fever, Sex and Lymphocytes associated with disease outcomes are the most important variables by objective Bayesian variable selection method. We employed three different copula models to find the relationship and dependence structure between Class (DHF DHF) and the selected important variables. Our proposed GCMR modeling outperformed the other models by Ju and Brasier [11] tested on the basis of accuracy, sensitivity, specificity, and the area under the receiver operating characteristic (AUC) and predicted ability to generalize. Dependence structure of these four component protein profiles (Days Fever, Lymphocytes, IL-10 and Sex) associated with disease outcomes was discovered by the semi-parametric Bayesian Gaussian copula model and Gaussian partial correlation method. From the results of these copula methods, we found that IL-10 is the main variable to develop DHF and Lymphocytes is the main variable not to develop DHF. These findings suggest optimal approaches for modeling a predictive biomarker panel in human host response to an infectious disease.

Compliance with Ethical Standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Aasa K, Czadob C, Frigessic A, Bakkend H (2009) Pair-copula constructions of multiple dependence. Insur Math Econ 44(2):182–198
2. Ahmad Z (2019) The hyperbolic Sine Rayleigh distribution with application to bladder cancer susceptibility. Ann Data Sci 6:211–222. https://doi.org/10.1007/s40745-018-0165-0
3. Bayarri MJ, Berger JO, Forte A, Garcia-Donato G (2012) Criteria for Bayesian model choice with application to variable selection. Ann Stat 40:1550–1577
4. Brasier AR, Ju H, Garcia J, Spratt HM, Victor SS, Forshey BM, Halsey ES, Comach G, Sierra G, Blair PJ, Rocha C, Morrison AC, Scott TW, Bazan I, Kochel TJ, Venezuelan Dengue Fever Working Group (2012) A Three-Component Biomarker Panel for Prediction of Dengue Hemorrhagic Fever. Am J Trop Med Hyg 86(2):341–348
5. Denuit M, Lambert P (2005) Constraints on concordance measures in bivariate discrete data. J Multivar Anal 93(1):40–57
6. Garcia-Donato G, Forte A (2015) R package BayesVarSel. R Foundation for Statistical Computing, Vienna
7. Genest, C., & Nešlehová, J. (2007). A primer on copulas for count data. ASTIN Bulletin 37(2):475–515. https://doi.org/10.1017/S0515036100014963
8. Genest C, Ghoudi K, Rivest LP (1995) A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82(3):543–552
9. Hoff PD (2007) Extending the rank likelihood for semiparametric copula estimation. Ann Appl Stat 1(1):265–283
10. Joe H (1997) Multivariate models and dependence concepts. Chapman and Hall, London
11. Ju H, Brasier AR (2013) Variable selection methods for developing a biomarker panel for prediction of dengue hemorrhagic fever. BMC Res Notes 6:365
12. Kim D, Kim J-M (2014) Analysis of directional dependence using asymmetric copula-based regression models. J Stat Comput Simul 84(9):1990–2010
13. Kim J-M, Jung Y-S, Sungur EA, Han K, Park C, Sohn I (2008) A copula method for modeling directional dependence of genes. BMC Bioinform 9:225
14. Kim J-M, Jung Y-S, Choi T, Sungur EA (2011) Partial correlation with copula modeling. Comput Stat Data Anal 55(3):1357–1366
15. Kojadinovic I, Yan J (2010) Modeling multivariate distributions with continuous margins using the copula R Package. J Stat Softw 34(9):1–20
16. Madsen L, Fang Y (2011) Joint regression analysis for discrete longitudinal data. Biometrics 67(3):1171–1175
17. Masarotto G, Varin C (2012) Gaussian copula marginal regression. Electron J Stat 6:1517–1549
18. Nelsen R (2006) An introduction to copulas, 2nd edn. Springer, New York
19. Olson D, Shi Y (2007) Introduction to business data mining. McGraw-Hill/Irwin, Boston
20. Shi Y, Tian YJ, Kou G, Peng Y, Li JP (2011) Optimization based data mining: theory and applications. Springer, London
21. Sklar A (1959) Fonctions de repartition a n-dimensions et leurs marges, (French). Publ Inst Stat Univ Paris 8:229–231
22. Song PX-K (2000) Multivariate dispersion models generated from Gaussian copula. Scand J Stat 27:305–320
23. Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Zellner A (ed) In Bayesian inference and decision techniques: essays in Honor of Bruno de Finetti. Edward Elgar Publishing Limited, Cheltenham, pp 389–399