Continuous Ranked Probability Score Validation Methods in Mixture Bayesian Model for Microarray Data in Indonesia

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052012 doi:10.1088/1757-899X/546/5/052012

Continuous Ranked Probability Score Validation Methods in Mixture Bayesian Model for Microarray Data in Indonesia

Ani Budi Astuti

Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Brawijaya Malang, Jl. Veteran Malang 65145 Indonesia

E-mail: ani_budi@ub.ac.id

Abstract. Validation in statistical modeling becomes a very important part to get information on how well the model has been built. Algorithm of Continuous Ranked Probability Score (CRPS) is a validation method of goodness of fit model in statistical modeling. A model that has a small CRPS value and has a small statistical significance, then the model is declared fit for data. Conversely, if a model has a large CRPS value, then the model is declared not fit for data. Several applications of the CRPS Algorithm have been developed for unimodal distribution models. Bayesian mixture is a modeling with Bayesian approach where data has a multimodal distribution. Characteristics of multimodal distribution are owned by microarray data in Indonesia, namely data on gene expression differences for several gene IDs from Chickpea plants in Indonesia. The purpose of this study was to obtain a performance from the Continuous Ranked Probability Score (CRPS) Algorithm as a goodness of fit model method in Bayesian Mixture Model (BMM) modeling for microarray data in Indonesia in a series of activities to find new varieties of Chiekpea plants that are resistant to attack by pathogenic fungal diseases Ascochyta Rabiei. The results of this study have succeeded in establishing the Algorithm of Continuous Ranked Probability Score (CRPS) for the distribution of normal mixture for data on gene expression differences of Chickpea plants in Indonesia as a result of microarray experiments with Bayesian approaches. BMM modeling on microarray data is declared fit because it has a small average value of CRPS, which is 0.0412 to 0.385.

Keywords: Continuous Ranked Probability Score, Bayesian, Mixture Model, Data Microarray-Indonesia

1. Introduction

In statistical modeling, the model validation that has been built is an equally important part to do. The model validation is a series of goodness of fit model activities, namely identifying conformity between models that have been built with the model from the original data ([1], [2]). A model is said to be fit when the model built does not have a significant difference with the model from the original data, both from its characteristics and behavior [1]. Model validation can be done with several statistical test tools, one of them is the distribution suitability test. In this test, the concept of suitability can be measured through the largest vertical distance between the cumulative distribution of the model constructed (the predicted model) and the cumulative distribution of the original data model (empirical model). This concept is known as goodness of fit Kolmogorov-Smirnov (KS) ([2], [3]). In addition, there is another

concept to measure the suitability of distribution, that is through the area of the difference in the cumulative distribution square of the model constructed (estimated model) and the cumulative distribution of the original data model (empirical model). This concept is known as Continuous Ranked Probability Score (CRPS) goodness of fit ([4], [5] and [6]). The Kolmogorov-Smirnov test is widely applied in the case of data with unimodal (non mixture) distribution, while the Continuous Ranked Probability Score test is not widely used and even if it is used it is still limited to cases of data with unimodal (non-mixture) distribution.

Chickpea, which is a producer of Chickpea Beans, which in Indonesia is called Arab Beans, is a very beneficial plant, both its roots and its fruit. The peanut products have a good taste and have high nutritional value. But the price of Chickpea beans in Indonesia is quite expensive as a result of the decline in peanut production. Chickpeas are susceptible to pathogenic fungal diseases, Ascochyta Rabiei, which results from attacks of this disease can reduce the production of Chickpea Plants [7]. Therefore we need a study of groups of gene IDs from Chickpea Plants that are characterized as Up-regulated (resistant to the attack of pathogenic fungal diseases Ascochyta Rabiei), so that they can obtain new varieties of superior Chickpea plants that can increase the production of Chickpea Beans. The challenge of data on gene expression differences in Chickpeas in Indonesia as a result of a microarray experiment, in order to classify gene IDs is that the sample sizes available are very small for each gene ID [7], but the distribution of the data is complex (mixture distribution) ([8], [9], [10], [11], [12], [13], [14] and [15]). This requires special handling to get the right model, so that the appropriate Up-regulated gene IDs grouping can be obtained. In addition, data on differences in gene expression have specific meanings, so special scenarios are needed to simulate data on differences in gene expression according to the original data pattern [16].

Bayesian analysis is a superior method in providing inference to an unknown parameter based on its posterior distribution while maintaining the data condition as it is through a data driven concept [17]. In Bayesian analysis, it is possible to have different combinations of prior distributions and facilitate iterative updating based on new information, so as to overcome the uncertainty and complexity of the model from the data [18]. In addition, Bayesian analysis is a statistical analysis method that does not consider sample size in data, so this analysis is very flexible to use for data that has small or large sample sizes, especially for small sample sizes. Bayesian analysis is also flexible in the form of data distribution, both for unimodal (nonmixture) distribution and for multimodal (mixture) distribution [19].

In this study, the performance of Continuous Ranked Probability Score will be examined in validating the mixture model through simulation data on the gene expression differences of Chickpeas in Indonesia in order to determine the group gene IDs of Up-regulated in a series of activities to find new varieties of Chickpea plants that are resistant to attack by pathogenic fungal diseases Ascochyta Rabiei.

2. Material and Methods

The simulation data used in this study are based on original data from the research results of [7] on gene expression differences as a result of a microarray experiment of Chickpea plants in Indonesia in healthy conditions and diseased conditions as a result of the attack of pathogenic fungi Ascochyta Rabiei. There are two groups of gene IDs function in Chickpeas used in research, namely the defence function and energy function. In the defence function gene IDs group, which has a mixture distribution of 10 ID genes and in the energy function gene IDs group, which has a mixture distribution of 9 ID genes ([8], [9], [10], [11], [12], [13], [14] and [15]). Each gene ID used in this study has a sample size $n = 3$. The simulation data generation uses a specific scenario to obtain simulation data that is similar to the original data as data from [7] as has been found in the study of [16], given data on gene expression differences have a specific meaning. Data generated as much as 1,000 times with each data generation takes a sample size of 3. The Bayesian mixture model approach was used to model data on gene expression differences of Chickpeas in Indonesia on each gene ID used in this study and validation of the model that have been built using the Continuous Ranked Probability Score (CRPS) concept. The fit Bayesian mixture model is indicated by a small CRPS value. Software R and R2WinBUGS are used as a tool for analyzing data.

The results of exploratory data on gene expression differences in healthy and diseased conditions of Chickpea plants in Indonesia as a result of [7] research for group IDs gene defence function and group IDs gene energy function are shown in Figure 1.

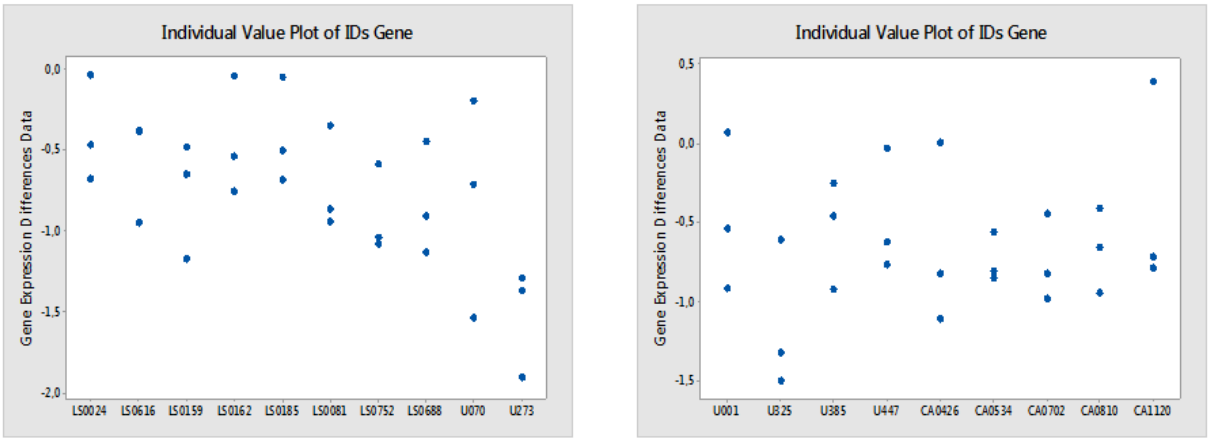

Figure 1. Data Exploration of Gene Expression Differences to Defence Function (a) and Energy Function (b) as a Result of [7] Research.

3. Simulation Modeling Concept, Bayesian Mixture Analysis and Continuous Ranked Probability Score

In this sub-section, the theories related to the research will be elaborated. These theories are simulation modeling concept that are used as a reference for generating simulation data, Bayesian Analysis, Mixture Model, Bayesian Mixture Model as a model of research data and Continuous Ranked Probability Score as a validation of the Bayesian mixture model.

3.1. Simulation Modeling Concept

Analytical evaluation of the system can be done easily but often very difficult because of the complexity of the system being studied. If the system can be learned easily then experiments can be carried out directly on the system, but if the system is very difficult to learn, then experiments must be done through the model of the system. In this case the simulation scenario must be done to get the right model that is able to describe a complex system [1]. Simulation activities to imitate the original system pattern require complete information about the system to be imitated and some specific assumptions and require good computing technology. The assumptions in question are generally about the relationship of mathematical models or logical relationships to the original system, so that the resulting imitation of a system that is able to properly describe the original system.

An experiment is carried out through learning the model of the system because the complexity of the model can be done in two ways, namely through physical models and through mathematical models. If the Mathematical model is used to study the system, there are two ways to complete the Mathematical model, namely through analytic solutions and simulation solutions. A complete illustration of the system completion map can be seen in Figure 2.

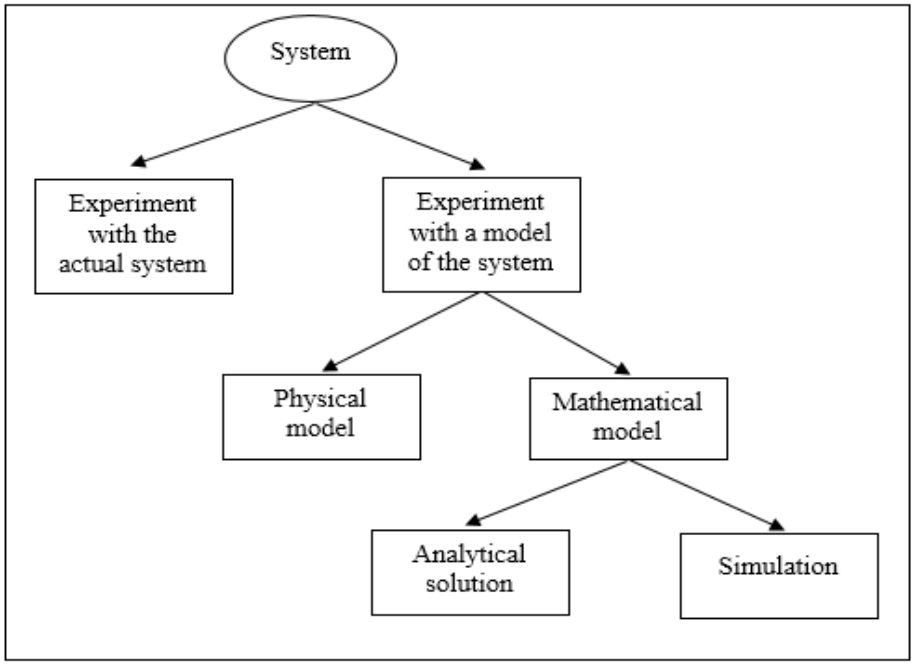

Figure 2. Maps out Different Ways in Which a System Might be Studied [1].

3.2. Bayesian Analysis

Bayesian analysis is a statistical analysis method based on the posterior probability distribution model with a structure as a combination of two information, namely prior information about parameter model and likelihood function ([2], [20], [21], [22] and [23]). In Bayesian analysis, model parameters $\boldsymbol{\theta}$ seen as a random variable in the parameter space $\boldsymbol{\Omega}$. In the Bayesian analysis concept, suppose there are observational data $\mathbf{x}$ which has a likelihood function $f(\mathbf{x}|\boldsymbol{\theta})$, then the information about the parameter $\boldsymbol{\theta}$ which is known before observation, is called prior $\boldsymbol{\theta}$, that is $p(\boldsymbol{\theta})$. Posterior probability distribution of the $\boldsymbol{\theta}$, that is $p(\boldsymbol{\theta}|\mathbf{x})$ can be known based on equation 1.1 ([19], [22]).

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{f(\mathbf{x})} \tag{1.1}$$

where:

$$f(\mathbf{x}) = \begin{cases} \int\limits_{\boldsymbol{\theta}\in R} f(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})\,d\boldsymbol{\theta}, & \text{when } \boldsymbol{\theta} \text{ is continuous,} \\ \sum\limits_{\boldsymbol{\theta}\in B} f(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}), & \text{when } \boldsymbol{\theta} \text{ is discrete,} \end{cases}$$

According to [22], $f(\mathbf{x})$ in equation 1.1 applies as a normalized contant, so equation 1.1 can be written in proportional form as in equation 1.2.

$$p(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}). \tag{1.2}$$
$$\text{posterior} \propto (\text{likelihood function}) \times (\text{prior})$$

3.2.1. Mixture Model

The mixture model is a special model for data that has multimodal properties, namely data that has a composition of sub-populations or groups, where each sub-population is a element component of the mixture model that has different proportions The mixture model is called a special model because this model is able to combine data while maintaining the characteristics of the original data [22], [24], [25-26]. According to [22], [24], and [27], the mixture probability function which is a probabilistic model of an observation made, i.e. $\mathbf{x}=(x_1, x_2, ..., x_n)$ taken from a number of k subpopulations as shown in equation 1.3.

$$f(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{w}) = w_1 g_1(\mathbf{x} \mid \theta_1) + ... + w_k g_k(\mathbf{x} \mid \theta_k) = \sum_{j=1}^{k} w_j g_j(\mathbf{x} \mid \theta_j), \quad (1.3)$$

where:

$f(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{w})$ = mixture probability function of the data $\mathbf{x}$ with the model parameter vector $\boldsymbol{\theta}$ and weighting vector $\mathbf{w}$,

$g_j(\mathbf{x} \mid \theta_j)$ = j[th] probability function, $j = 1, 2, ..., k$ with parameter θ_j which is a parameter vector whose characteristics depend on the form of distribution g_j each component in the mixture model, and

$\mathbf{w}$ = weighting parameter vector of the mixture model with elements $w_1, w_2, ..., w_j$, where $0 < w_j < 1, \forall j$ with $j = 1, 2, ..., k$ and $\sum_{j=1}^{k} w_j = 1$.

3.2.2. Mixture Bayesian Model

Bayesian analysis in the mixture model views that all parameters in the model are random variables that have certain prior distributions, so this analysis requires prior distribution specifications for each parameter in the model [28]. According to [29], in arrange the mixture model consider that each observation x_i will be a member of one of the unknown subpopulations. If the allocation of each observation in each subpopulation of the mixture model in equation 1.3 is denoted by $\mathbf{z}$, then the proportion of allocation for each observation z_i determined from the distribution in equation 1.4.

$$p(z_i = j) = w_j, \; i = 1, 2, 3, ..., n \text{ and } j = 1, 2, 3, ..., k. \quad (1.4)$$

Based on equation 1.4, if given a value z_i then according to equation 1.3, observation data x_i derived from the subpopulation which is distributed in equation 1.5.

$$x_i \mid z_i \sim g_j(\mathbf{x} \mid \boldsymbol{\theta}_{z_i}), \; i = 1, 2, \cdots, n. \quad (1.5)$$

Thus, the combined posterior distribution produced by the mixture model will take the form of a combined distribution of all variables in the mixture model as described in equation 1.6.

$$p(k, \mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{x}) = p(k)\, p(\mathbf{w} \mid k)\, p(\mathbf{z} \mid \mathbf{w}, k)\, p(\boldsymbol{\theta} \mid \mathbf{z}, \mathbf{w}, k)\, p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, k), \quad (1.6)$$

where:

$p(k, \mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{x})$ = the combined posterior distribution of the mixture model with the number of mixture components k from data $\mathbf{x}$ with parameter model $\boldsymbol{\theta}$, weighting $\mathbf{w}$ and allocation parameter $\mathbf{z}$,

$p(k)$ = prior distribution of parameters for the number of mixture components as much as k,

$p(\mathbf{w}|k)$ = prior distribution of weighting parameters $\mathbf{w}$ with certain value of k,

$p(\mathbf{z}|\mathbf{w},k)$ = prior distribution of allocation parameters $\mathbf{z}$ conditional $\mathbf{w}$ and certain value of k,

$p(\boldsymbol{\theta}|\mathbf{z},\mathbf{w},k)$ = prior distribution of the model parameters $\boldsymbol{\theta}$ conditional $\mathbf{z}$, $\mathbf{w}$ and certain value of k, and

$p(\mathbf{x}|\boldsymbol{\theta},\mathbf{z},\mathbf{w},k)$ = likelihood function of data $\mathbf{x}$ known $\boldsymbol{\theta}$, $\mathbf{z}$, $\mathbf{w}$ and certain value of k.

Posterior distribution in Bayesian analysis is often very complicated to do and requires a complex integration process in determining marginal posterior of the model parameter, so a numerical Markov Chain Monte Carlo (MCMC) approach method is needed ([2], [30] and [31]). The MCMC algorithm approach to the Gibbs Sampler method can facilitate identification of the posterior distribution in Bayesian modeling ([31], [32], [33] and [34]).

3.2.3. The MCMC Algorithm with the Gibbs Sampler Approach

A numerical Markov Chain Monte Carlo (MCMC) approach is needed to obtain the posterior distribution in Bayesian analysis which is often very complicated to do and requires a complex integration process in determining posterior marginal parameters of a model [30]. Through the MCMC Algorithm with the Gibbs Sampler approach it will make it easy for complex modeling so that this method is considered a breakthrough in the use of Bayesian analysis. The MCMC method is a simulation method that combines Monte Carlo with Markov Chain properties to obtain sample data based on certain sampling scenarios ([22], [23] and [30]). The following is presented in full the MCMC algorithm with the Gibbs Sampler approach. If the model parameters are denoted by $\boldsymbol{\theta}$, then the posterior distribution can be found through three steps from these MCMC algorithm with the Gibbs Sampler approach ([15], [30],[32] and [33]) as shown in the Algorithm 1.

Algorithm 1. MCMC Algorithm with the Gibbs Sampler Approach

Step 1. We will provide initial value of the parameter $\boldsymbol{\theta}$.

$$\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)},...,\theta_r^{(0)}\right)$$

Step 2. We will do the sampling of the parameter $\boldsymbol{\theta}$.

Generate the value of θ_j, $j = 1,...,r$ from their conditional distribution as follows:

Step 2.1. Sampling $\theta_1^{(k+1)}$ from $p\left(\theta_1 | \mathbf{x}, \theta_2^{(k)},...,\theta_r^{(k)}\right)$

Step 2.2. Sampling $\theta_2^{(k+1)}$ from $p\left(\theta_2 | \mathbf{x}, \theta_1^{(k+1)}, \theta_3^{(k)},...,\theta_r^{(k)}\right)$

.
.
.

Step 2.r. Sampling $\theta_r^{(k+1)}$ from $p\left(\theta_r | \mathbf{x}, \theta_1^{(k+1)}, \theta_2^{(k+1)},..., \theta_{r-1}^{(k+1)}\right)$

Step 3. We will do iteration of the parameter $\boldsymbol{\theta}$.

Execute step 2 as K times with $K \to \infty$

3.3. Goodness of Fit Continuous Ranked Probability Score (CRPS)

The basic concept of CRPS in validating the model is based on area (integral) of the difference in the square of the predicted cumulative distribution (hypothesis) with the empirical cumulative distribution

(observation). The value of CRPS is the value of the Mean Square Error from the predicted cumulative distribution (hypothesis) [6]. The CRPS formula is as in equation 1.7 ([4], [5]).

$$CRPS = \int_{-\infty}^{+\infty} \left(\hat{F}(x) - F_n(x_i) \right)^2 dx \tag{1.7}$$

where:

$CRPS$: nilai *Continuous Ranked Probability Score*

$\hat{F}(x)$: cummulative density function from predicted model (hypothesis)

$F_n(x_i)$: cummulative density function from empirical model (observation)

The result of CRPS according to equation 1.7 is a constant value that shows the smaller the value of CRPS (close to value 0), the better the predictive value and vice versa if the value of CRPS is greater (far from the value 0), the estimated value gets worse [35]. Illustration of the concept of CRPS theory according to [5] shown in Figure 3 and according to [36] is shown in Figure 4.

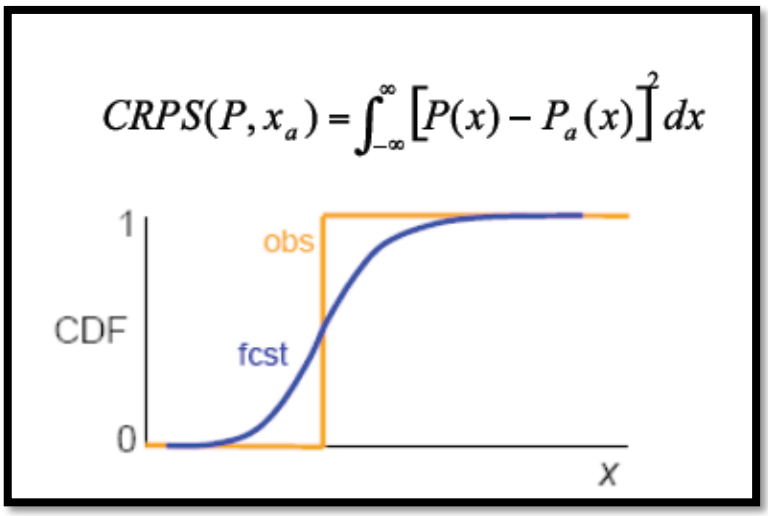

Figure 3. Illustration of the Concept for CRPS Theory According to [5].

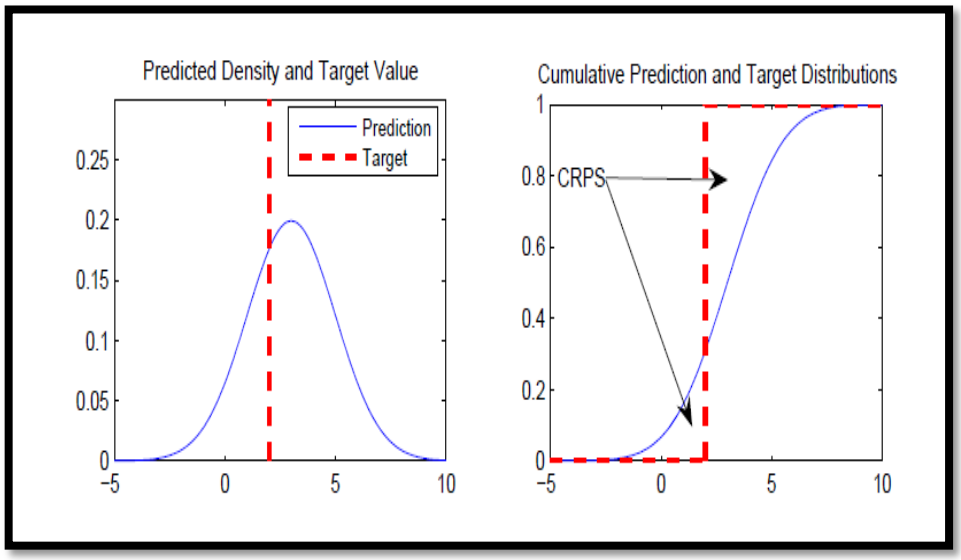

Figure 4. Illustration of the Concept for CRPS Theory According to [36].

4. Result and Discussion

In this section, the results and discussion of this study will be presented. The results of the research presented are exploration of simulation data obtained, Bayesian mixture models for each gene ID, Continuous Ranked Probability Score Algorithm and Validation of the built Bayesian mixture model.

4.1. Simulation Data Exploration

Exploration of the results of the average generation of simulation data for groups of gene IDs defence function are shown in Table 1 and for groups of energy function gene IDs are shown in Table 2. This simulation data is compatible with the original data as the results of [7] and has been researched by [16]. The values of gene expression differences obtained have a negative sign, this means that 10 gene IDs in the defence function group and 9 gene IDs in the energy function group are healthy (resistant to the attack of pathogenic fungal disease Ascochyta Rabiei). This is indicated by the value of healthy expression is more dominant than the value of deseased expression. The values of gene expression differences are calculated based on the following formula:

$$\ln\left(\frac{\text{value of deseased genes expression}}{\text{value of healthy genes expression}}\right).$$

Table 1. Average Simulation Data from Generation 1,000 Times each Repetition for the Defence Function of Gene IDs

No	Defence Function of Gene IDs	Repetition			Average
		U1	U2	U3	
1	LS0024	-0.6241	-0.4632	-0.0880	-0.3918
2	LS0616	-0.8970	-0.3885	-0.4253	-0.5703
3	LS0159	-1.1179	-0.6544	-0.5298	-0.7674
4	LS0162	-0.6988	-0.5336	-0.0943	-0.4422
5	LS0185	-0.6350	-0.4998	-0.1008	-0.4119
6	LS0081	-0.8875	-0.8623	-0.4046	-0.7181

7	LS0752	-1.0311	-1.0316	-0.6381	-0.6876
8	LS0688	-1.0795	-0.9007	-0.4983	-0.3599
9	U070	-1.4864	-0.7061	-0.2483	-0.4955
10	U273	-1.8522	-1.3727	-1.3413	-1.5221

Table 2. Average Simulation Data from Generation 1,000 Times each Repetition for the Energy Function of Gene IDs

No	Energy Function of Gene IDs	Repetition			Average
		U1	U2	U3	
1	U001	-0.8628	-0.5305	0.0158	-0.4592
2	U325	-1.4493	-1.3215	-0.6623	-0.9236
3	U385	-0.8709	-0.4621	-0.3058	-0.5463
4	U447	-0.7124	-0.6218	-0.0870	-0.4737
5	CA0426	-1.0557	-0.8208	-0.0495	-0.3519
6	CA0534	-0.7974	-0.7974	-0.6152	-0.7367
7	CA0702	-0.9262	-0.8219	-0.4988	-0.7490
8	CA0810	-0.8932	-0.6565	-0.4592	-0.6696
9	CA1120	-0.7348	-0.7129	0.3356	-0.3707

4.2. Mixture Bayesian Model

The normal mixture Bayesian model is formed for each gene ID observed in this study. The formation of the model is based on the results of the Reversible Jump Markov Chain Monte Carlo (RJMCMC) test as a Bayesian approach test method to determine the number of mixture components in the data, where these has been investigated by [17]. The test results for the number of mixture components in each gene ID are shown completely in Table 3 and Table 4. In the defence function group gene IDs, 9 gene IDs were obtained that have a normal mixture Bayesian model with two components mixture and 1 gene ID that has a normal mixture Bayesian model with three components mixture. Whereas in the energy function group gene IDs, 7 gene IDs were obtained that have a normal mixture Bayesian model with two components mixture and 2 gene IDs that have a normal mixture Bayesian model with three components mixture.

4.3. Continuous Ranked Probability Score Algorithm

To facilitate the use of the Continuous Ranked Probability Score (CRPS) theory in validating the model, CRPS Algorithm was developed in this study. There are five steps in the CRPS Algorithm, in full the steps of the CRPS Algorithm are shown in Algorithm 2.

Algorithm 2. CRPS Algorithm

Step 1. Look at the $\mathbf{x}$ data as a random variable with x as the simulation observation data from the Chickpea data with sample size $n = 3$, so the observation data is x_1, x_2, x_3.

Step 2. Set the empirical cumulative distribution function from simulation data, i.e. $F_n(x_i)$.

Step 3. Set the function of the hypothesis cumulative distribution from simulation data, i.e. $\hat{F}(x)$

Step 4. Calculate the value of CRPS by formula $CRPS = \int_{-\infty}^{+\infty} \left(\hat{F}(x) - F_n(x_i)\right)^2 dx$

Step 5. Interpretation of the results of model validation for simulation data. The smaller the value of CRPS (close to the value 0), the more appropriate the model. Conversely, if the value of the CRPS is getting bigger (far from the value 0), then the model is increasingly not suitable.

4.4. Validation of the Mixture Bayesian Model

Normal mixture Bayesian models that have been formed, then the model validation process is needed to find out the goodness of fit model. In the Table 3 and Table 4, the average CRPS values for each gene ID were observed in this study. The results obtained show that the average value of CRPS has a range between 0.041 to the highest 0.385. Based on these CRPS values indicate that the value of CRPS is close to the value of 0. This means that the normal mixture Bayesian model that has been formed for each gene ID is fit for the original data model. Therefore, a normal mixture Bayesian model with two and three components mixture can well illustrate the distribution form of data on gene expression differences for Chickpeas in Indonesia. Through this model, it can be further utilized to determine gene ID groups that have Up-regulated characters, so that they can finally find new varieties of superior Chickpea Plants in Indonesia.

Table 3. Result of Continuous Ranked Probability Score Goodness of Fit Test for the Simulation Data on Defence Function Gene IDs

No	Defence Function of Gene IDs	*n*	Mixture Bayesian Model for Each Gene IDs	Average Value of CRPS
1	LS0024	3	Two Components Normal Mixture	0.0722
2	LS0616	3	Two Components Normal Mixture	0.3380
3	LS0159	3	Two Components Normal Mixture	0.1650
4	LS0162	3	Two Components Normal Mixture	0.0775
5	LS0185	3	Two Components Normal Mixture	0.0934
6	LS0081	3	Two Components Normal Mixture	0.3400
7	LS0752	3	Two Components Normal Mixture	0.3370
8	LS0688	3	Two Components Normal Mixture	0.0565
9	U070	3	Three Components Normal Mixture	0.0443
10	U273	3	Two Components Normal Mixture	0.3550

Table 4. Result of Continuous Ranked Probability Score Goodness of Fit Test for the Simulation Data on Energy Function Gene IDs

No	Energy Function of Gene IDs	*n*	Mixture Bayesian Model for Each Gene IDs	Average Value of CRPS
1	U001	3	Three Components Normal Mixture	0.0412
2	U325	3	Two Components Normal Mixture	0.3360
3	U385	3	Two Components Normal Mixture	0.1050
4	U447	3	Two Components Normal Mixture	0.3440
5	CA0426	3	Two Components Normal Mixture	0.0578
6	CA0534	3	Two Components Normal Mixture	0.3850

7	CA0702	3	Two Components Normal Mixture	0.1870
8	CA0810	3	Three Components Normal Mixture	0.3730
9	CA1120	3	Two Components Normal Mixture	0.3630

5. Conclusion

In this study, the performance of Continuous Ranked Probability Score as a validation method for normal Bayesian mixture models can be demonstrated for data on differences in gene expression as a result of a microarray experiment from Chickpeas in Indonesia. Normal Bayesian mixture models with two and three mixture components are able to describe the data model well, where the lowest CRPS value is 0.041 and the largest CRPS value is 0.385. The simulation data that was built, through certain scenarios has been compatible with the original data, so that through this simulation data generalization of information about CRPS performance can be done well. The normal mixture Bayesian model with two and three components mixture for each gene ID can be used to find groups of gene IDs that are included in the Up-Regulated characteristic which are then used to find superior Chickpea plant varieties.

Acknowledgements

This paper is part of research Grants the DPP/SPP activities at University of Brawijaya in 2018. We would like the first thank to University of Brawijaya which have financed this research and the second thank to Harijati for allowing use of research data and thank to anonymous reviewer to this paper.

References

[1] Law, A. M. and Kelton, W. D. 1991. Simulation Modeling and Analysis. Second Edition. McGraw-HiII, Inc. New York.
[2] Iriawan, N. 2003a. Simulation Technique. Teaching Module. (In Indonesian). ITS, Surabaya.
[3] Sheskin, D. J. 2007. Handbook of Parametric and Nonparametric Statistical Procedures, 4th Edition. Chapman & Hall/CRC.
[4] Hersbach, H. 2000. Decomposition of the Continous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting. 15. 559-570.
[5] Wilson, L. J. 2009. Verification of Ensemble Forecasts: A Look to the Future. www.ec.gc.ca. Environment, Canada.
[6] Leutbecher, M. 2012. Ensemble Verification II. Training Course.
[7] Harijati, N. 2007. A Study of the Resistance of Chickpea (Cicer Arietinum) to Ascochyta Rabiei and the Effect of Age of Plant Tissue on Disease Development. Ph.D. Thesis. (Australia: La Trobe University).
[8] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2013. Highest Posterior Density for Identifying Differences in Gene Expression Microarray Experiments. Proceeding of the 3nd Annual Basic Science International Conference (BaSIC 2013).Volume 3. UB, Malang. ISSN 2338-0136.
[9] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2014a. Model Components Selection in Bayesian Model Averaging Using Occam's Window for Microarray Data. Natural-A Journal of Scientific Modeling and Computation, 2 (1): 67-74. ISSN: 2303-0135.
[10] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2014b. Kolmogorov-Smirnov and Continuous Ranked Probability Score Validation on the Bayesian Model Averaging for Microarray Data. Applied Mathematical Sciences, 146 (8): 7277-7287. DOI: http://dx.doi.org/10.12988/ams.2014.49760.
[11] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2015b. Occam's Window Selection in Bayesian Model Averaging Modeling for Gene Expression Data from Chickpea Plant. International Journal of Applied Mathematics and Statistics, 53 (4): 160-165. ISSN 0973-1377

(Print), ISSN 0973-7545 (Online).
[12] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2015c. An Algorithm for Determining the Number of Mixture Components on the Bayesian Mixture Model Averaging for Microarray Data. Journal of Mathematics and Statistics, 11 (2): 45-51. DOI: 10.3844/jmssp.2015.45.51.
[13] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2017b. Development of Reversible Jump Markov Chain Monte Carlo Algorithm in the Bayesian Mixture Modeling for Microarray Data in Indonesia. AIP Conference Proceedings 1913, 020033 (2017). https://doi.org/10.1063/1.5016667.
[14] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2017c. Simulation Study Scenario on Chickpea Data in Indonesia for Bayesian Mixture Modeling. International Symposium on Biomathematics (Symomath) 2017. ITB, Bandung.
[15] Astuti, A. B., Iriawan N., Irhamah and Kuswanto H. 2017d. Bayesian Mixture Model Averaging (BMMA) for Identifying the Differences Genes Expression of Chickpea (Cicer Arietinum) Plant Tissue. Communications in Statistics-Theory and Methods, 46 (21): 10564-10581. DOI: http://dx.doi.org/10.1080/03610926.2016.1239112.
[16] Astuti, A. B. and Iriawan N. 2018. Simulation Study Scenario Algorithm on Gene Expression Differences of Chickpea Data in Indonesia to Modeling Series with the Bayesian Approach. IOP Conference Proceedings (Accepted and on Process).
[17] Iriawan, N. 2012. Modeling and Analysis of Data-Driven. (in Indonesian) Vol.1, ITS, Surabaya.
[18] Mengersen, K. 2009. Module 1 Bayesian Analysis, Short Course on Bayesian Modeling. Department of Statistics. ITS, Surabaya.
[19] Gosh, J. K., Delampady, M. and Samanta, T. 2006. An Introduction to Bayesian Analysis Theory and Method. Springer, New York.
[20] Box, G. E. P. and Tiao. 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley Massachusetts, Boston MA.
[21] Zellner, A. 1971. An Introduction to Bayesian Inference in Econometrics. John Wiley, New York.
[22] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. 1995. Bayesian Data Analysis. Chapman & Hall, London.
[23] Congdon, P. 2006. Bayesian Statistical Modelling. 2nd. John Wiley & Sons, USA.
[24] McLachlan, G. J. and Basford, K. E. 1988. Mixture Models Inference and Applications to Clustering. Marcel Dekker, New York.
[25] Escobar, M. D. and West, M. 1995. Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association. 90: 577–588.
[26] McLachlan, G. J., Bean, R. W. and Peel, D. 2002. A mixture Model-Based Approach to the Clustering of Microarray Expression Data. Bioinformatics. 18: 413–422.
[27] Iriawan, N. 2001. Univariable normal mixture model estimation: A Bayesian methods approach with MCMC. (in Indonesian). Proceedings of National Seminar and Konferda VII on Mathematics in DIY & Central Java Region, Yogyakarta, Indonesia, pp:105–110.
[28] Aitkin, M. 2001. Likelihood and Bayesian Analysis of Mixtures. Statistical Modelling. 1: 287-304.
[29] Richardson, S. and Green, P. J. 1997. On Bayesian Analysis with an Unknown Number of Components. Journal of the Royal Statistical Society. B. 59. 4: 731–792.
[30] Iriawan, N. 2000. Computationally Intensive Approaches to Inference in Neo-Normal Linear Models. Ph.D. Thesis. CUT-Australia.
[31] Iriawan, N. 2003b. Modeling Data with MCMC Using WinBUGS 1.4. Teaching Module. (In Indonesian). ITS, Surabaya.
[32] Gamerman, D. 1997. Markov Chain Monte Carlo. Chapman & Hall, London.
[33] Stephens, M. 1997. Bayesian Methods for Mixtures of Normal Distribution. Ph.D. Thesis. Oxford.
[34] Walsh, B. 2004. *Markov Chain Monte Carlo and Gibbs Sampling*. Lecture Notes for EEB 581, version 26.

[35] Gneiting, T. and Raftery, A. E. 2007. Strictly Proper Scoring Rules, Prediction and Estimation. Journal of the American Statistical Association. 102. 359-378.
[36] Camey, M. and Cunningham, P. (2006). Evaluating Density Forecasting Models. https://www.cs.tcd.ie/publications/tech-reports/...06/TCD-CS-2006-21.pdf.