Implementation of Bayesian Mixture Models in identifying subpopulation of breast cancer patients based on blood test measurements

N Dwimantara1, S Abdullah2,a, A Bustamam3, A Rachman4
1,2,3 Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia
4 Faculty of Medicine, Universitas Indonesia, Depok, 16424, Indonesia

a) E-mail: sarini@sci.ui.ac.id

Abstract. A complete blood test is one of a series of initial examinations of cancer patients that is relatively easy. The use of blood measurement components in analysing patient conditions is commonly used. However, it is not the case for the ratio and inter-ratio components of blood measurements, and this is what is proposed in this study. The built hypothesis is that the ratio and inter-ratio components of blood tests that can explain the condition of cancer patients are better than the blood test’s own components. An analysis will also be conducted to develop a patient profile based on these measurements, and those that can clearly distinguish between patient groups will be identified. The Finite Mixture Model is a method for modelling heterogeneous data that may originate from different subpopulations, where subpopulations represent groups of patients based on a particular latent condition. This model takes the form of a superposition of several distributions, which in this study, a Gaussian distribution is used. The parameter estimation used in the Bayesian method, which determines the prior distribution of the model parameters, and it is combined with the likelihood which will produce a posterior distribution. Then, the Markov Chain Monte Carlo-Gibbs Sampler is used to draw samples on the parameters of the posterior distribution. By using the breast cancer patient blood test data from the Oncology Department of a hospital in Jakarta, with 100,000 iterations as burn-in, and 200,000 iterations for sampling, based on Deviance Information Criterion values, the optimal grouping is two subpopulations using blood ratio and inter-ratio measurements. Two subpopulations were identified, with the first population is characterized by low distribution value and the second subpopulation with the opposite characteristics. The explanatory factors of ratio data are ratio neutrophils to lymphocytes, ratio platelets to lymphocytes, and ratio lymphocytes to monocytes.

1. Introduction
Some tests that are useful for cancer diagnosing are the laboratory tests, image processing, biopsy, and genomic test. However, to get a good result, many kinds of costly check-up are required. The medical industry needs an alternative to facilitate the diagnosis process that easy, cheap, but without loss of accuracy.

A complete blood test is one of the laboratory checks that requires lower cost. In last few years, the medical industry relentlessly did a research about the effect of blood test ratio between platelet, lymphocyte, neutrophil, and monocyte to the cancer diagnosis. Ratio neutrophil to lymphocyte (NLR) [1] and ratio platelet to lymphocyte (PLR) [2] have been a prognostic marker from hematologic
measurements that are interesting, easy, and cost-effective. NLR and PLR reflect breast cancer inflammation, in which a higher value of NLR means that the prognosis of breast cancer worsens [2].

Platelet count, white blood cell count (WBC), platelet volume (MPV), NLR, and PLR as a surrogate marker are compared between a breast cancer patient and healthy people, but it is rarely the research about cancer diagnosis with complete blood test [3]. NLR takes effect significantly to breast cancer stadium, unlike the PLR [4]. A high cut off value of PLR indicates bad overall survival, and disease-free-survival of the breast cancer patient, so a standard value for PLR cut off is needed [5]. It means that the characteristic of the blood test ratio still gives various results.

The significant factors of the blood test will be described by knowing the subpopulation’s characteristics of cancer patients based on the blood test. Clustering is one way to know the characteristics of the data which divides it into several clusters that have the same character each cluster.

Furthermore, this research will define another measurement that is expected to be better me, which contains the ratios between each blood test ratio called the inter-ratio.

2. Data and Methods
This research uses patient blood test data from the Oncology Department of one hospital in Jakarta, with 100,000 iterations to burn, and 200,000 iterations for sampling. The ratio factors are NLR (Ratio Neutrophils to Lymphocytes), PLR (Ratio Platelets to Lymphocytes), LMR (Ratio Lymphocytes to Monocytes), HPR (Ratio Haemoglobin to Platelets), PWR (Ratio Platelets to White Blood Cells), and LWR (Ratio Lymphocytes to White Blood Cells). The Inter-Ratio factors are PLRNLR (Ratio PLR to NLR), PLRLMR (Ratio PLR to LMR), PLRPWR (Ratio PLR to PWR), NLRLMR (Ratio NLR to LMR), NLRPWR (Ratio NLR to PWR), LMRPWR (Ratio LMR to PWR), LMRHPR (Ratio LMR to HPR), LMRLWR (Ratio LMR to LWR), HPRLWR (Ratio HPR to LWR), and HPRPWR (Ratio HPR to PWR). In addition, it will be defined Finite Mixture Model (FMM), Markov Chain – Gibbs Sampler (MCMC-GS), and Deviance Information Criterion will be defined by [6][7][8].

2.1. Finite Mixture Model (FMM)
Let \( y = (y_1, ..., y_n) \) is a sample of the data which has \( n \)-size of observation. The likelihood of a single \( y_i \) for a Finite Mixture Model with \( K \) components is given by,

\[
p(y_i|\eta, \theta) = \sum_{k=1}^{K} \eta_k p(y_i|\theta_k), \quad i = 1, 2, ..., n \tag{1}
\]

and, for any \( y = (y_1, ..., y_n) \)

\[
p(y|\eta, \theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \eta_k p(y_i|\theta_k) \tag{2}
\]

\( \eta_k \) represent the proportion of data that is explained by subgroup \( k \), known as the component weights. Each \( \eta_k \) ranges between 0 and 1 and \( \sum_{k=1}^{K} \eta_k = 1 \). The component-specific parameter \( \theta_k \) defines \( p(y_i|\theta_k) \), whose distribution also describes each component.

The membership of each observation to one of these subgroups is the main interest. To indicate this for \( y_i \), we define \( z_i \), a discrete single-valued latent variable introduced. Taking values from \( k = 1, ..., K \), combining the latent variable into the model does not alter the likelihood of \( y_i \)

\[
p(y_i, z_i = k|\eta, \theta) = \eta_k p(y_i|\theta_k) \tag{3}
\]

\[
p(y_i|\eta, \theta) = \sum_{k=1}^{K} p(y_i, z_i = k | \eta, \theta) \tag{4}
\]
2.2. **MCMC-Gibbs sampler**

The Markov chain Monte Carlo (MCMC) is a technique that employed to correct and generate the estimates of unknown parameters $\theta$ in order to have a better estimate of the desired posterior distribution. When MCMC is processed, we need to check whether the algorithm converges to desired the posterior distribution [9]. Markov chain Monte Carlo (MCMC) offers a way to produce samples from the posterior when analytical normalization constants cannot be implemented, as is often the case in very simple models. The general idea of the MCMC method is to provide a large number of samples or iterations of the posterior estimates, the order of the sample values of each parameter provides estimates for the exact posterior distribution, with the expectation that each unknown value is given from the Monte Carlo estimate. The most commonly implemented MCMC technique is the Gibbs sampler, mainly useful for model interface under conjugate priors.

The Gibbs decomposes the posterior, which is considered to be multivariate and complex, into a number of simpler and easier samples of the distribution for each unknown parameter. This distribution is known as a full posterior. Let parameters of interest $\theta = \theta_1, ..., \theta_J$ be the full conditional of each $\theta_j$ depending on parameter $J - 1$ unknown parameters take the general form

$$p(\theta_j | \theta_k, y) \propto p(y | \theta_1, ..., \theta_j)p(\theta_j)$$  \hfill (5)

where $k \neq j$ and each full condition is familiar parametric form. For each $\theta_j$, an initial estimate is provided, then a cyclical sampling scheme is provided by these full conditionals where each $\theta_j$ is processed in order from full conditions given the latest estimates for other parameters. For a single iteration $t$, a generalization of this scheme is given by the equation

$$\theta_1^{(t)} \sim p(\theta_1 | \theta_2^{(t-1)}, ..., \theta_J^{(t-1)}, y)$$
$$\theta_2^{(t)} \sim p(\theta_2 | \theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_J^{(t-1)}, y)$$
$$\vdots$$
$$\vdots$$
$$\theta_j^{(t)} \sim p(\theta_j | \theta_1^{(t)}, ..., \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, ..., \theta_J^{(t-1)}, y)$$  \hfill (6)

2.3. **Deviance information criterion (DIC)**

DIC strongly agrees with Bayesian analysis because it formally takes into account uncertainties in the model parameters. Similar with other information criteria, DIC is the sum of the model fit term and penalty conditions. Given the parameter $\psi$, the general form of the DIC is given by,

$$DIC = D(\psi^*) + 2p_D.$$  \hfill (7)

Here $D(\psi^*)$ represents the deviance of the model given an estimate of $\psi$,

$$D(\psi^*) = -2 \log p(y | \psi^*) + 2 \log f(y).$$  \hfill (8)

The inclusion of $f(y)$ acts as a standardising term and is a function of $y$ only. For example, on a general linear model, this may correspond to the likelihood of the saturated model, resulting in $D(\psi^*)$ representing the scaled residual deviance. For model comparison, it is generally accepted that $f(y)$ is set to one. To motivate the derivation of the penalty term, in this case, given by $p_D$,

$$p_D = -2 \log p(y | \psi) + 2 \log p(y | \psi^*)$$  \hfill (9)
Within Bayesian paradigm, since the true value of \( \psi \) is unknown, this discrepancy may be approximated by its posterior expectation, with respect to the posterior of the unknown parameters, \( p(\psi|y) \), resulting in the definition,

\[
p_D = E_{\psi|y}\{-2\log p(y|\psi)\} + 2\log p(y|\psi^*)
\]

\[
p_D = \bar{D}(\psi) - D(\psi^*)
\]

In other words, \( p_D \) measures the difference between the average deviance and the given estimated parameters of the model. Thus, it can be interpreted as reducing the uncertainty of the model resulting from parameter estimation. In practice, \( \bar{D}(\psi) \) is easy to calculate, because it is solely the average deviance in all MCMC iterations. Therefore, it can be seen that DIC takes into account uncertainty in estimating parameters through \( \bar{D}(\psi) \), Referring to Equation (7), given expressions for \( D(\psi^*) \) and \( p_D \), the DIC may be rewritten as follows

\[
DIC = D(\psi^*) + 2p_D
\]

\[
= D(\psi^*) + 2(\bar{D}(\psi) - D(\psi^*))
\]

\[
= 2\bar{D}(\psi) - D(\psi^*)
\]

3. Results

By using the ratio and inter-ratio data, FMM can create the subpopulation of the data. We apply the algorithm with 200000 iterations to sample the parameters posterior distribution. For all tables, \( P[] \) informs the weight of subpopulations. Deviance is a value that will be used to calculate DIC in determining the best number of subpopulations. Lambda[] represents the mean value in the subpopulation.

### Table 1. Ratio data with 4-subpopulation

| Node | Mean  | Std. Dev. | MC error | 2.5% | Median | 97.5% |
|------|-------|-----------|----------|------|--------|-------|
| P[1] | 0.3354| 0.2216    | 0.003949 | 0.02627 | 0.2982 | 0.8174 |
| P[2] | 0.2019| 0.1709    | 0.002522 | 0.005825 | 0.1558 | 0.6313 |
| P[3] | 0.1941| 0.1702    | 0.002476 | 0.004981 | 0.1454 | 0.6244 |
| P[4] | 0.2686| 0.2043    | 0.003623 | 0.01733 | 0.2194 | 0.7482 |
| Deviance | 7389.0 | 4.016 | 0.02875 | 7382.0 | 7389.0 | 7389.0 |
| Lambda[1] | 34.45 | 10.04 | 0.1424 | 8.516 | 36.43 | 47.53 |
| Lambda[2] | 42.7 | 7.25 | 0.08358 | 27.37 | 43.01 | 56.04 |
| Lambda[3] | 50.09 | 8.979 | 0.1032 | 36.79 | 48.89 | 71.53 |
| Lambda[4] | 61.22 | 16.02 | 0.1972 | 43.99 | 56.87 | 106.9 |
| Sigma | 94.76 | 2.715 | 0.00609 | 89.63 | 94.7 | 100.3 |

### Table 2. Ratio with 3-subpopulation

| Node | Mean  | Std. Dev. | MC error | 2.5% | Median | 97.5% |
|------|-------|-----------|----------|------|--------|-------|
| P[1] | 0.4156| 0.2415    | 0.004461 | 0.03855 | 0.3933 | 0.8942 |
| P[2] | 0.2355| 0.1919    | 0.003076 | 0.006581 | 0.1877 | 0.7007 |
| P[3] | 0.349 | 0.2327    | 0.004277 | 0.0246 | 0.3124 | 0.8486 |
| Deviance | 7386.0 | 3.525 | 0.02542 | 7381.0 | 7386.0 | 7394.0 |
| Lambda[1] | 37.43 | 8.82 | 0.1389 | 15.76 | 38.98 | 48.98 |
| Lambda[2] | 46.08 | 7.458 | 0.08239 | 32.85 | 45.79 | 61.36 |
| Lambda[3] | 56.65 | 13.23 | 0.1565 | 42.29 | 53.39 | 93.6 |
| Sigma | 94.74 | 2.708 | 0.006269 | 89.64 | 94.68 | 100.2 |
Table 1 shows the ratio data with four subpopulations. From this table, we can get information that has an overlap for each subpopulation. It appears that the third quantile mean value of the first subpopulation is higher than the first quantile mean value of the second subpopulation, the third quantile mean value of the second subpopulation is higher than the first quantile mean value of the third subpopulation and the third quantile mean value of the third subpopulation is higher than the fourth quantile mean value of the second subpopulation.

Likewise shown in Table 2, in the ratio data with three subpopulations. The third quantile mean value of the first subpopulation is higher than the first quantile mean value of the second subpopulation, and the third quantile mean value of the second subpopulation is higher than the first quantile mean value of the third subpopulation. Furthermore, in Table 3, it appears that in ratio data with two subpopulations seems that the third quantile mean value of the first subpopulation is higher than the first quantile mean value of the second subpopulation.

| Node | Mean | Std. Dev. | MC error | 2.5% | Median | 97.5% |
|------|------|-----------|----------|------|--------|-------|
| P[1] | 0.5179 | 0.253 | 0.004571 | 0.069 | 0.5209 | 0.9511 |
| P[2] | 0.4821 | 0.253 | 0.004571 | 0.04887 | 0.4791 | 0.931 |
| Deviance | 7383.0 | 2.792 | 0.0134 | 7379.0 | 7383.0 | 7390.0 |
| Lambda[1] | 40.75 | 6.495 | 0.06687 | 25.38 | 41.61 | 50.57 |
| Lambda[2] | 51.42 | 9.333 | 0.0977 | 40.39 | 49.69 | 74.71 |
| Sigma | 94.72 | 2.705 | 0.006292 | 89.62 | 94.66 | 100.2 |

Table 4 shows the ratio data with four subpopulations. From the table, we discover that has an overlap for each subpopulation formed. It seems from the third quantile mean value of the first subpopulation is higher than the first quantile mean value of the second subpopulation, the third quantile mean value of the second subpopulation is higher than the first quantile mean value of the third subpopulation, and the third quantile mean value of the third subpopulation is higher than the fourth quantile mean value of the second subpopulation. So also shown in table 5, inter-ratio data with three subpopulations. The third quantile mean value of the first subpopulation is higher than the first quantile mean value of the second subpopulation, the third quantile mean value of the second subpopulation is higher than the first quantile mean value of the third subpopulation is higher than the fourth quantile mean value of the second subpopulation. Furthermore, table 6, inter-ratio data with two subpopulations seems that the third quantile mean value of the first subpopulation is higher than the first quantile mean value of the second subpopulation.

| Node | Mean | Std. Dev. | MC error | 2.5% | Median | 97.5% |
|------|------|-----------|----------|------|--------|-------|
| P[1] | 0.3184 | 0.2108 | 0.003572 | 0.02634 | 0.2814 | 0.7901 |
| P[2] | 0.2044 | 0.1712 | 0.002501 | 0.006072 | 0.1589 | 0.6311 |
| P[3] | 0.1949 | 0.1667 | 0.00245 | 0.005395 | 0.1497 | 0.6128 |
| P[4] | 0.2822 | 0.2012 | 0.003339 | 0.02137 | 0.2392 | 0.749 |
| Deviance | 11580 | 4.089 | 0.03168 | 11570.0 | 11580 | 11590 |
| Lambda[1] | 25.31 | 5.696 | 0.1009 | 11.1 | 26.42 | 32.58 |
| Lambda[2] | 29.88 | 3.915 | 0.04488 | 21.51 | 30.1 | 36.93 |
| Lambda[3] | 33.8 | 4.433 | 0.04679 | 26.77 | 33.3 | 44.13 |
| Lambda[4] | 39.16 | 7.212 | 0.07919 | 30.71 | 37.41 | 59.04 |
| Sigma | 66.44 | 1.469 | 0.003414 | 63.64 | 66.41 | 69.39 |
Table 5. Inter-ratio with 3 sub-population

| Node  | Mean   | Std. Dev. | MC error  | 2.5%  | Median | 97.5%  |
|-------|--------|-----------|-----------|-------|--------|--------|
| P[1]  | 0.398  | 0.2355    | 0.004408  | 0.0373| 0.3749 | 0.8715 |
| P[2]  | 0.2333 | 0.1883    | 0.003072  | 0.0070| 0.1859 | 0.6867 |
| P[3]  | 0.3687 | 0.2285    | 0.004031  | 0.0336| 0.337  | 0.8468 |
| Deviance | 11570 | 3.554    | 0.02659   | 11570.0| 11570.0| 11580.0|
| Lambda[1]| 26.91 | 4.99     | 0.07576   | 14.62 | 27.82 | 33.39 |
| Lambda[2]| 31.72 | 3.836    | 0.03973   | 24.54 | 31.63 | 39.44 |
| Lambda[3]| 36.84 | 5.798    | 0.06244   | 29.79 | 35.61 | 52.12 |
| Sigma  | 66.42  | 1.468    | 0.003399  | 63.62 | 66.39 | 69.37 |

Table 6. Inter-ratio data with 2-subpopulation

| Node  | Mean   | Std. Dev. | MC error  | 2.5%  | Median | 97.5%  |
|-------|--------|-----------|-----------|-------|--------|--------|
| P[1]  | 0.5084 | 0.2487    | 0.004567  | 0.0688| 0.5078 | 0.9421 |
| P[2]  | 0.4916 | 0.2487    | 0.004567  | 0.0578| 0.4922 | 0.9311 |
| Deviance | 11570 | 2.825    | 0.01402   | 11570.0| 11570.0| 11580.0|
| Lambda[1]| 28.92 | 3.666    | 0.04184   | 20.18 | 29.43 | 34.36 |
| Lambda[2]| 34.53 | 4.324    | 0.04292   | 28.8  | 33.8  | 45.11 |
| Sigma  | 66.41  | 1.465    | 0.003298  | 63.62 | 66.38 | 69.35 |

Table 7. DIC value

| 2 Subpopulation | 3 Subpopulation | 4 Subpopulation | Ratio DIC | 2 Subpopulation | 3 Subpopulation | 4 Subpopulation | Inter-Ratio DIC |
|-----------------|-----------------|-----------------|----------|-----------------|-----------------|-----------------|-----------------|
| 7525.8          | 7612.2          | 7677.1          | 11822.8  | 11933.7         | 12033.9         |

Each subpopulation deviance will be used to calculate the DIC, in which the lowest DIC value is better. The DIC value is displayed in table 7. Both ratio and inter-ratio data, the best number of subpopulations formed, are two subpopulations with DIC value 7525.8 for ratio data and 11822.8 for inter-ratio data. After that, we will describe all the factors of ratio and inter-ratio in the following figure.
The ratio data has six factors, and three of them, namely PLR, NLR, and LMR do not overlap. Although the first subpopulation has low distribution values, it is centred because the interval of the value is narrow. The second subpopulation has a large distribution value, yet it has a wider interval of the values, meaning that this subpopulation contains various values. A medical analysis is required to analyse the deferent overlap between PWR, HPR, and LWR.

On the other hand, the inter-ratio data with ten factors has eight factors that does not overlap, they are PLRNL, PLRLMR, PLRPWR, NLRLMR, NLRPWR, LMRPWR, LMRLWR, and HPRLWR. Similar to the ratio, Although the first subpopulation that is formed has small values, it is centred as well due to the narrow interval of values. The second subpopulation has a large value but wider interval of values, meaning that this subpopulation contains various values. Therefore, a medical analysis is required to analyse the deferent overlap between LMRPWR, and HPRPWR.
Figure 9. Inter-ratio data distributions with PLRPWR factor

Figure 10. Inter-ratio data distributions with NLRLMR factor

Figure 11. Inter-ratio data distributions with NLRPWR factor

Figure 12. Inter-ratio data distributions with LMRPWR factor

Figure 13. Inter-ratio data distributions with LMRHPR factor

Figure 14. Inter-ratio data distributions with HRPWR factor

Figure 15. Inter-ratio data distributions with LMRHPR factor

Figure 16. Inter-ratio data distributions with HRPWR factor
4. Conclusion

The From the results, FMM can identify to get subpopulations based on distribution of several blood measurements. In ratio data, the subpopulations formed is two subpopulations with DIC score 7525.8, then in inter-ratio data was formed two subpopulations with DIC 11822.8.

Either ratio or inter-ratio data, the best number of subpopulations formed are two subpopulations. The first subpopulation distribution value is low, but it is centred because the interval of the value is narrow. Although the second subpopulation has a large value, the interval of the value is wide and the distribution value is high, it means that this subpopulation contains various values.

The analysis of the result shows that NLR, PLR, and LMR are the significant factors for ratio data. On the other hand, the factors for inter-ratio data that take significant effect for cancer are PLRNLR, PLRLMR, PLRPWR, NLRLMR, NLRPWR, LMRPWR, LMRLWR, and HPRLWR.

5. References

[1] Wei B, Yao M, Xing C, Wang W, Yao J, Hang Y, Liu Y, and Fu P 2016 The neutrophil lymphocyte ratio is associated with breast cancer prognosis: an updated systematic review and meta-analysis Journal of Onco-Targets and Therapy 9 5567-5575
[2] Zhang M, Huang X Z, Song Y X, Gao P, Sun J X, and Wang Z N 2017 High platelet-to-lymphocyte ratio predicts poor prognosis and clinicopathological characteristics in patient with breast cancer: a meta-analysis BioMed Research International 2017 Article ID 9503025, 11 pages
[3] Sun Z, Ju Y, Han F, Sun X, and Wang F 2018 Clinical implications of pretreatment inflammatory biomarkers as independent prognostic indicators in prostate cancer Journal of Clinical Laboratory Analysis 32 (3)
[4] Elyasinia F, Keramati M, Ahmadi F, Rezaei S, Ashouri F, Parsaei R, Yaghoubi M, Elyasinia F, Aboutorabi A, Rezaei S, and Kaviani A 2017 Neutrophil-lymphocyte ratio in different stages of breast cancer Journal of Clinical Laboratory Analys Acta Medica Iraniais 55 (4)
[5] Zhu Y, Si W, Sun Q, Qin B, Zhao W, and Yang J 2017 Platelet-lymphocyte ratio acts as an indicator of poor prognosis in patients with breast cancer Oncotarget 8 (1)
[6] Wichitchan S, Yao W, and Yang G 2019 Hypothesis testing for finite mixture models Computational Statistics and Data Analysis 132 180–189
[7] White N, Johnson H, Silburn P, Mellick G, Dissanayaka N, and Mengersen K 2012 Probabilistic subgroup identification using Bayesian finite mixture modelling: a case study in Parkinsons disease phenotype identification. Statistical Methods in Medical Research 21(6) 563-83
[8] Molitor J, Paphathomas M, Jerrett M, and Richardson S 2010 Bayesian profile regression with an application to the National survey of children’s health Biostatistics 11 (3) 484-498
[9] Ojo O B, Lougue S, Woldegerima W A 2017 Bayesian generalized linear mixed modeling of tuberculosis using informative priors PLoS One 12(3) e0172580.

Acknowledgments

This research is supported by PITTA A Scheme Research Grant (Thesis Indexed International Publication Competitive Grant) with ID: NKB-0446/UN2.R3.1/HKP.05.00/2019, 2019 from DPRM Universitas Indonesia. We thank to all reviewers for the improvement of this article.