Austrian Journal of Statistics
July 2021, Volume 50, 101–114.
http://www.ajs.or.at/
doi:10.17713/ajs.v50i5.1195

Biclustering Analysis Using Plaid Model on Gene Expression Data of Colon Cancer

Titin Siswantining Universitas Indonesia
A. Eriza Aminanto Universitas Indonesia
Devvi Sarwinda Universitas Indonesia
Olivia Swasti Universitas Indonesia

Abstract

Unlike other typical clustering analysis, which considers column only, biclustering analysis processes a matrix into sub-matrices based on rows and columns simultaneously. One method of bicluster analysis uses the probabilistic model, like the plaid model, that provides overlapping bicluster. The plaid model calculates the value of an element given from a particular sub-matrix for each cell; thus, the value can be seen as the number of contributions of a particular bicluster. The algorithm begins with preparing the input data as a matrix, then an initial model is assessed and makes a residual matrix from the model. After that, we determine bicluster candidates, which are evaluated for its effect parameters and bicluster membership parameters. Finally, the bicluster candidate is pruned to give the optimal bicluster. We implemented the algorithm on gene expression dataset of colon cancer, where the rows and columns contain observations and types of genes, respectively. We carried out in six distinct scenarios in which each scenario uses different model parameters and threshold values. We measured the results using Jaccard index and coherence variance. Our experiments show that biclustering analysis on a model with mean, row, and column effects of colon cancer data output low coherence variance.

Keywords: biclustering, expression gene dataset, overlapping bicluster, plaid model.

1. Introduction

Data clustering analysis aims to group variables in the data matrix based on specific global patterns (concerning all variables), meaning that patterns formed in either rows or columns are considered. In contrast to clustering analysis, biclustering analysis aims to find local patterns in a big data matrix. Local patterns are patterns found in the data of a particular set of rows and columns simultaneously (Kasim, Shkedy, Kaiser, Hochreiter, and Talloen 2016).

Biclustering methods can be divided into four main classes: correlation maximization methods, correlation minimization methods, two-way clustering methods, and probabilistic or generative methods (Denitto, Bicego, Farinelli, and Figueiredo 2017). A probabilistic model is a model created using statistical analysis to describe data based on probability theory. Parameter identification is made by using the given statistical model by minimizing specific criteria. One of the probabilistic biclustering methods is the plaid model. The biclustering approach is generally based on a sum or multiplication model, which evaluates each bicluster's con-

tribution separately without considering the interaction between bicluster. The plaid model considers the value of an element that is the value of a submatrix given in the matrix. The value of these elements can be seen as the value of contributions from different bicluster. The plaid model method can also consider the effects of rows and columns to obtain the right model. One strength of the plaid model is the model's ability to model bicluster that can be overlapping so that it can get a suitable model. Figure 1 shows the overlapping bicluster visualization by Kasim, Shkedy, Kaiser, Hochreiter, and Talloen (2016).

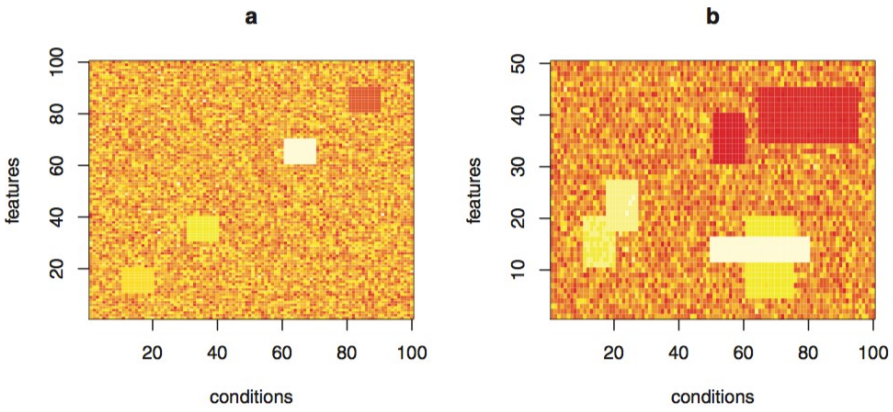

Figure 1: Overlapping bicluster visualization (a) No overlapping biclusters, (b) Overlapping biclusters are existed (Kasim *et al.* 2016)

The biclustering method was initially proposed by Cheng and Church (2000) and used to conduct microarray-shaped gene expression data in bioinformatics research. The microarray data is in the form of a matrix in which each row states the research observations or individual samples, and the column states the characters in the microarray data are in the form of genes, while the entries state the level of gene expression. Microarray data analysis is widely used to find information about gene function. Gene expression research is carried out with thousands of genes, measuring the level of gene expression in several observations, possibly representing a series of different experimental or class conditions. A group of genes with the same expression patterns in a group of observations can be co-regulated, which can indicate a general function of genetics (Turner, Bailey, and Krzanowski 2005). Likewise, the observation group with a group of genes with the same expression value has the same attributes; for example, they might be observations from patients with the same disease. The purpose of bicluster analysis varies, one of which is early detection of a disease that attacks genetics such as Alzheimer's, cancer, and tumors. Ardaneswari, Bustamam, and Siswantining (2017) and Latief, Siswantining, Bustamam, and Sarwinda (2019) analyzed gene expression data of carcinoma tumors and hepatocellular carcinoma, respectively. Ardaneswari, Bustamam, and Siswantining (2017) leveraged a parallel k-means algorithm for two-phase method biclustering, while Latief, Siswantining, Bustamam, and Sarwinda (2019) classified the gene expression data using random forest feature selection. In this research, the implementation of the plaid model method was carried out in gene expression data of colon cancer. The disease dataset was obtained from the Kaggle website (https://www.kaggle.com/masudur/colon-cancer-gene-expression-data).

The contributions of this research are two-fold as follows.

- We analyze the leverage of biclustering analysis using the plaid model and implement the method in the case of a matrix of gene expression data of colon cancer.
- We examine and prove empirically that the lowest coherence variance can be achieved from six different scenarios of models and threshold values.

The remainder of this paper is organized as follows: Section 2 describes the background theories needed for this study. Experiment results and discussion are provided in Section 3. Section 4 concludes this paper with an overview of future work.

2. Methodology

This section explains K-means clustering and plaid model in biclustering analysis. The K-means clustering is necessary for initialization step. We describe the plaid model in biclustering analysis with the construction of initial model and residual matrix, determination of initial bicluster candidates, parameter estimation and pruning the bicluster. All these theories are given as follows.

2.1. K-means clustering

K-means clustering analysis is one of the methods of clustering analysis that aims to obtain a cluster by minimizing the objective function (cost function), which generally uses the Euclidean distance to the center of a group. Given a collection of observations X as many as $n(x_1, x_2, ..., x_n)$ so the k-means clustering algorithm creates a partition on X observations of several k clusters. The minimized cost function in the K-Means Clustering analysis will be as Equation (1) below:

$$\xi = \sum_{i=1}^{n} \min(D_E(x_i - c_j)|j = 1, 2, ..., k) \tag{1}$$

Where x_i is the i-th observation, c_j is the j-th centroid cluster,$D_E(y_i - c_j)$ is the Euclidean distance between x_i and c_j and $\min(D_E(y_i - c_j)|j = 1, 2, ...k)$ is the Euclidean distance x_i and c_j which produces the smallest value. The Euclidean distance with the variable G obtained with Equation (2)

$$D_E(x_i - c_j) = \sqrt{\sum_{g=1}^{G}(x_{ig} - c_{jg})^2} \tag{2}$$

K-Means Clustering analysis carried out using the following algorithm:

1. Determination of the number of clusters (K) to be sought.
2. Determine the number of K centroids (cluster center points) starting, randomized or random.
3. Calculate the distance of each observation to the centroid of each cluster. The calculation done using Euclidean distance, then enter each observation into the cluster with the closest distance
4. Calculate new centroids with equations:

$$c_{ig} = \frac{1}{n_k} \sum_{i} x_{ig};\ i = 1, ..., n;\ i \in \text{cluster } k;\ j = 1..., k;\ g = 1..., G,$$

where c_{ig} is i-th centroid cluster and g-th dimension. n_k is the number of observations in the k-cluster. x_{ig} is the i-th observation on the g-th dimension.

5. Perform stages 3 and 4 continuously until the convergent result is achieved, i.e., the members of a cluster have not changed anymore (Rokach and Maimon 2010).

2.2. Plaid model in biclustering analysis

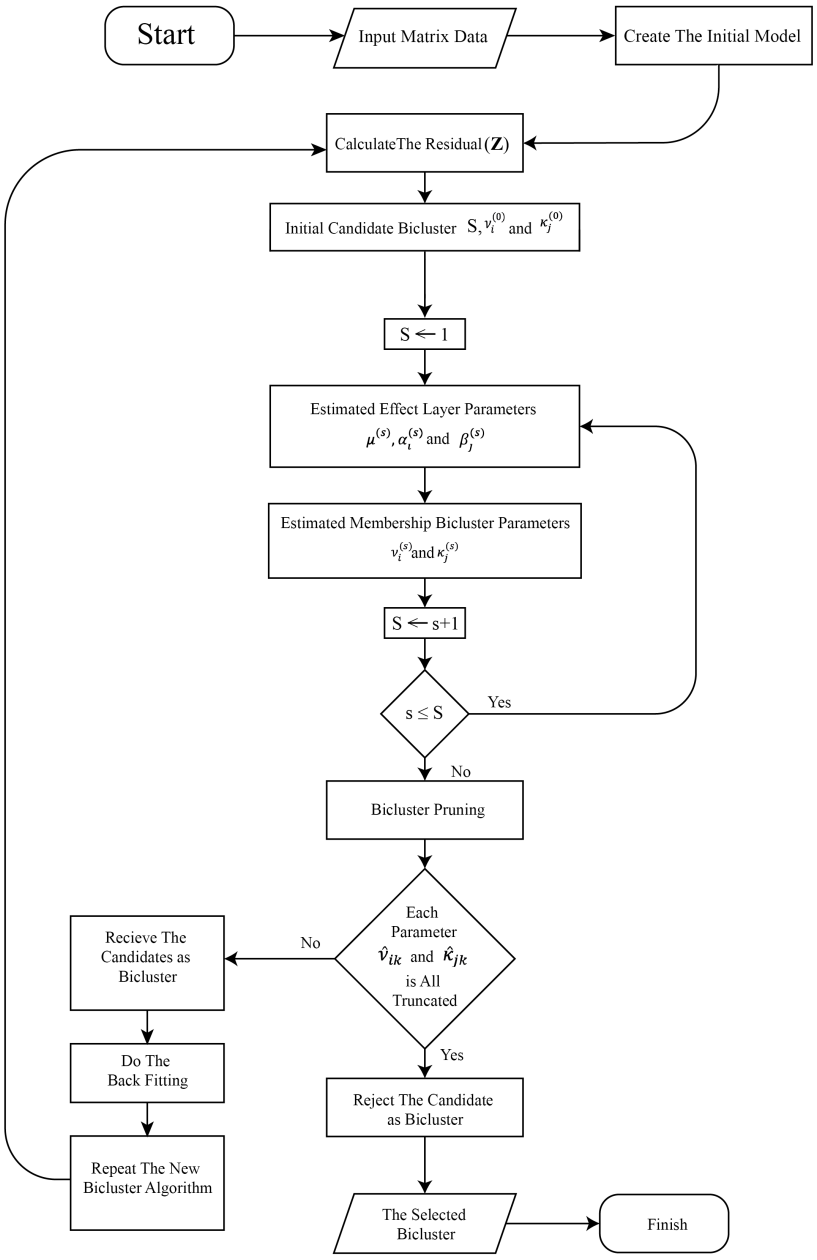

Figure 2: Flowchart of biclustering analysis using the plaid model

The Plaid model takes into account the value of an element that is the effective value of a submatrix given to the matrix. In biclustering analysis, these elements seen as the effect of different biclusters on the value of data entry. The model is additive, which contains the impact of the overall matrix and the impact of a bicluster. Each k-layer or bicluster has an effect on the model. The model was first proposed by Lazzeroni and Owen (2000). In the plaid model, the data entry value or expression level is Y_{ij}, which $i = 1, 2, ..., N, j = 1, 2, ..., M$ in the i-th sample/observation and j-th type genes / characteristic features have a model like

the following in Equation (3):

$$Y_{ij} = \theta_{ij0} + \sum_{k=1}^{K} \theta_{ijk}\nu_{ik}\kappa_{jk} + \varepsilon_{ij} = \mu_0 + \alpha_{0i} + \beta_{0j} + \sum_{k=1}^{K} \theta_{ijk}\nu_{ik}\kappa_{jk} + \varepsilon_{ij}, \tag{3}$$

where:

- The k letter notation is the k-layer or bicluster index.
- θ_{ij0} which is the background effect written as a sum of μ_0, α_{0i} and β_{0j}. μ_0 is a notation of the background effect or grand mean.α_{0i} is the effect of the i-th row. β_{0j} is the effect of the j-th row
- θ_{ij0} is the sum of the mean effect (grand mean), row effect and column effect on k-bicluster. In a k-bicluster, there are four possible forms of the average expression of genes (θ_{ij0}). The condition while $\theta_{ijk} = \mu_k$ shows the bicluster effect which only consists of the grand mean on k-bicluster. $\theta_{ijk} = \mu_k + \alpha_{ik}$ and $\theta_{ijk} = \mu_k + \beta_{jk}$ shows the effect of biclusters consisting of the grand mean effect and the effect of rows or columns on a bicluster. On $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$ implies a bicluster effect consisting of grand mean, row effect and column effect of a bicluster.

$$\theta_{ij0} = \begin{cases} \mu_k \\ \mu_k + \alpha_{ik} \\ \mu_k + \beta_{jk} \\ \mu_k + \alpha_{ik} + \beta_{jk} \end{cases}$$

- ν_{ik} is the k- bicluster affiliation parameter in the i-row which has a binary value ($\nu_{ik}\epsilon\{0,1\}$).ν_{ik} is one of the observations, or the i-sample is in k-bicluster, and zero if it is not k-bicluster. κ_{jk} is also a bicluster affiliation variable which indicates bicluster affiliation for the j-th characteristic or gene in k-bicluster($\kappa_{jk}\epsilon\{0,1\}$), and ε_{ij} is an error. Plaid models can take into account the effects of an overlapping bicluster, which can be seen as more than one bicluster, that gives effect to the model. In addition to parameters ν or κ for every Y_{ij} applicable,

$$\sum_{k=1}^{K} \nu_{ik} = \begin{cases} 1 \\ \geq 2 \\ 0 \end{cases}$$

Output 1 when the i-th observation is only a member of one bicluster. The output ≥ 2 when having a bicluster affiliation of more than one. Output 0 when the i-th observation is not a member of any bicluster. This equation also applies to the κ parameter, as is

$$\sum_{k=1}^{K} \kappa_{jk} = \begin{cases} 1 \\ \geq 2 \\ 0 \end{cases}$$

Output 1 when the j-th gene type is only a member of one bicluster. Output ≥ 2 when having more than one bicluster affiliation. Output 0 when the j-type gene is not a member of any bicluster. Figure 2 describes biclustering analysis using the plaid model.

Construction of initial model and residual matrix (Z)

The first stage in the bicluster search algorithm using the plaid model is the search for residual matrices (Z). In the first search process, we do not have a single bicluster, so the first model

has only contained the effect of the background or the entire matrix of data that owned. The initial model can take several forms. The complete initial model provides the grand mean ($\hat{\mu_0}$), row effect ($\hat{\alpha}_{0i}$) and column effect ($\hat{\beta}_{0j}$). The initial model can also be a combination of the sum of the grand mean effects, row effects, and column effects. The simplest initial model contains only a grand mean. These effects can be searched by Equation (4).

$$\hat{\mu}_0 = \bar{Z}_{..};\ \hat{\alpha}_{0i} = \bar{Z}_{i.} - \bar{Z}_{..};\ \hat{\beta}_{0j} = \bar{Z}_{.j} - \bar{Z}_{..}. \tag{4}$$

The next step is to look for a residual matrix (Z) of size NxM which contains the residual values of Z_{ij}. Z_{ij} value can be searched by reducing the value of data inputted with the effects of the obtained model, as explained in Equation (5).

$$Z_{ij} = Y_{ij} - (\hat{\mu}_0 + \hat{\alpha}_{0i} + \hat{\beta}_{0j}) \tag{5}$$

Bicluster search algorithm using the plaid model method is an algorithm that searches bicluster results one by one, meaning that when it gets bicluster results, it will repeat the algorithm to get a new bicluster. There is a slight difference in the search for residual matrices when bicluster has obtained. In that case, the residual value obtained by reducing the value of the data inputted with background effects, i.e., θ_{ij0} or $(\hat{\mu}_0 + \hat{\alpha}_{0i} + \hat{\beta}_{0j})$, and the effects of bicluster that have had, the equation is written in Equation (6).(Busygin, Prokopyev, and Pardalos 2008):

$$Z_{ij} = Y_{ij} - (\hat{\mu}_0 + \hat{\alpha}_{0i} + \hat{\beta}_{0j} + \sum_{k=1}^{K} \theta_{ijk}\nu_{ik}\kappa_{jk}) \tag{6}$$

Determination of initial bicluster candidates

The bicluster algorithm with the plaid model begins with the initial value of the parameter $\hat{\nu}_i$ and $\hat{\kappa}_j$. There are many ways to determine this value. Chaturvedi and Carroll (1994) proposes the use of random values to estimate these parameters. However, the data in the form of gene expression is usually extensive in dimensions, so more sophisticated methods are needed to handle it. Bicluster affiliation parameters only divide genes and observations into groups that are in bicluster and those that are not. Busygin, Prokopyev, and Pardalos (2008) found the initial value for the bicluster affiliation parameter (bicluster candidate) by looking at it as a two-way clustering analysis problem. The k-means algorithm with $k = 2$ used to classify gene types and individual observations. Then take clusters that have fewer members and pair row clusters with column clusters to form the initial bicluster. Smaller clusters used because minority groups have different gene expressions in the context of the data provided.

Parameter estimation

Suppose that $l - 1$ layer has been found, then the residue of the model is calculated as Equation (7).

$$\hat{Z}_{ij} = Y_{ij} - \hat{\theta}_{ij0} - \sum_{k=1}^{l-2} \hat{\theta}_{ijk}\hat{\nu}_{ik}\hat{\kappa}_{jk} \tag{7}$$

where $\hat{Z}_{ij}$ is the entry of residual matrix $(\hat{Z})$. $(\hat{Z})$ as an input of the matrix data, which is then estimated parameters. Layer subscript (k) removed because the bicluster search was done one by one, meaning that there was only one k value in the estimated parameter. The purpose of parameter estimation is to obtain the bicluster effect estimation μ, α_i and β_j, and affiliation parameters ν_i and κ_j, this can be done by minimizing the sum of squared residuals

on the k-bicluster in Equation 8.

$$Q = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}(\hat{Z}_{ij} - \theta_{ij}\nu_i\kappa_j)^2 \tag{8}$$

the algorithm starts with determining the initial value for $\hat{\nu}_i$ and $\hat{\kappa}_j$. The parameters of layer effects and layer affiliation will continue to be updated until they reach convergence or reach a certain iteration number. In this study, the number of iterations will be denoted in S as determined by the researcher.

The estimation step of the layer effect parameter explains the calculation of the effect given by a bicluster on the model. Suppose that $\hat{\nu}_i$ and $\hat{\kappa}_j$ is a parameter of affiliation parameter. $\mathbf{Z}^*$ is a submatrix of $\mathbf{Z}$, which is a member of the bicluster. $\mathbf{Z}^*$ have the entry of observations is $\hat{\nu}_i = 1$, and the entry of genes is $\hat{\kappa}_j = i$. The estimation of layer effects is similar to the principle of searching for row and column effects in a two-way analysis of variance (ANOVA) because the Q value in Equation 8 has the same shape as the residue squared number of two-way ANOVA with one observation per cell.

In the principle of searching for row and column effects in the two-way ANOVA model, τ_i and β_j occur because of treatments A and B. In contrast, in biclustering analysis using plaid models, both parameters are effects derived from rows and columns. So by using the plaid model can be written in Equation 9.

$$\hat{\mu}_k = \bar{Z}^*_{..};\ \hat{\alpha}_{ik} = \bar{Z}^*_{i.} - \bar{Z}^*_{..};\ \hat{\beta}_{jk} = \bar{Z}^*_{.j} - \bar{Z}^*_{..} \tag{9}$$

Z^* is a submatrix, which is also a bicluster candidate. For each observation and gene outside the bicluster candidate, no effects included in the model. The estimated effect is continuously updated for S iterations.

The effect parameters of the bicluster used to minimize Equations 10 and 11 to estimate the membership parameters. Estimation of membership parameters done one by one using binary least-square. The estimation is done by trying out the possible values of membership parameters (1 and 0) in Equations 10 and 11 and using the value of affiliation parameters that result in smaller values of Equations 10 and 11. By using the parameters $\hat{\mu}_k, \hat{\alpha}_{ik}$ and $\hat{\beta}_{jk}$ that have obtained, the parameter $\hat{\nu}_{ik}$ estimated using the minimum value from Equations 10. Equations 10 is the number of residual squares like Equations 8, but i is not a notation that moves on the summation, because i is the same notation at $\hat{\nu}_{ik}$.

$$\sum_{j=1}^{J}(\hat{Z}_{ijk} - \hat{\nu}_{ik}[\hat{\theta}_{ijk}\hat{\kappa}_{jk}])^2 \tag{10}$$

$\hat{\theta}_{ijk}$ has the entry is $\hat{\mu}_k, \hat{\alpha}_{ik} and \hat{\beta}_{jk}$. Same as before, we suggest that the parameters $\hat{\mu}_k, \hat{\alpha}_{ik}, \hat{\beta}_{jk}$ and $\hat{\nu}_{ik}$ is known. The parameters of $\hat{\kappa}_{jk}$ can also estimate using the minimum value from Equation 9.

$$\sum_{i=1}^{I}(\hat{Z}_{ijk} - \hat{\kappa}_{jk}[\hat{\theta}_{ijk}\hat{\nu}_{ik}])^2 \tag{11}$$

The next step is minimizing Equations 10 and 11 separately for each $\hat{\nu}_i$ or $\hat{\kappa}_j$ value of 1 and 0. In each parameter, there are only two possibilities, namely 0 and 1; this process done by trial and error. Trial and error did by comparing the values of Equations 10 and 11, which have the smallest values. In the calculation, $\hat{\nu}_i$ and $\hat{\kappa}_j$ are updated in parallel. In other words, when $\hat{\nu}_i$ is refreshed, the $\hat{\kappa}_j$ parameter used is derived from the previous iteration and vice versa in the $\hat{\kappa}_j$ parameter. Therefore, any parameter ($\hat{\nu}_i$ or $\hat{\kappa}_j$) that was calculated first will

give the same result. Estimation of these parameters carried out S times, as determined by the researcher.

Pruning the bicluster

The plaid model algorithm has phases for pruning genes and bicluster observations that do not match based on self-determined threshold values (τ_1 and τ_2). The value of bicluster affiliation parameters based on the bicluster pruning process is as follows:

$$\nu_i = \begin{cases} 1, & if\ \nu_i = 1\ and\ \sum_{j:\hat{\kappa}_j=1}(\hat{Z}_{ij} - \hat{\theta}_{ij})^2 < (1-\tau_1)\sum_{j:\hat{\kappa}_j=1}(\hat{Z}_{ij})^2 \\ 0, & others \end{cases} \tag{12}$$

$$\kappa_i = \begin{cases} 1, & if\ \kappa_i = 1\ and\ \sum_{i:\hat{\nu}_i=1}(\hat{Z}_{ij} - \hat{\theta}_{ij})^2 < (1-\tau_2)\sum_{i:\hat{\nu}_i=1}(\hat{Z}_{ij})^2 \\ 0, & others \end{cases} \tag{13}$$

where τ_1 and τ_2 is between 0 and 1, or we can write such as: $\tau_1\tau_2\varepsilon(0,1)$. $\hat{Z}_{ij}$ is the entry value in the residual matrix. $\hat{\theta}_{ij}$ is the effect value of the k-bicluster. The value of τ_1 and τ_2 which approaching 1 will give bicluster results the more similar the value (the more coherent) and vice versa. The bicluster candidates who have been pruned accepted as bicluster. After having a new bicluster, then back fitting is done and continued with the search for another new bicluster. However, if all genes or all observations truncated, the bicluster search algorithm is stopped. Using a small value of τ_1 and τ_2 can accept bicluster more quickly, so it can receive many biclusters. Busygin, Prokopyev, and Pardalos (2008) suggest the use of values $\tau_1\tau_2\varepsilon(0,5;0,7)$, because the value is high enough to accept bicluster candidates who have essential effects and sufficiently low to get bicluster candidate effects that are low enough. Figure 3 explains the pruning biclustering.

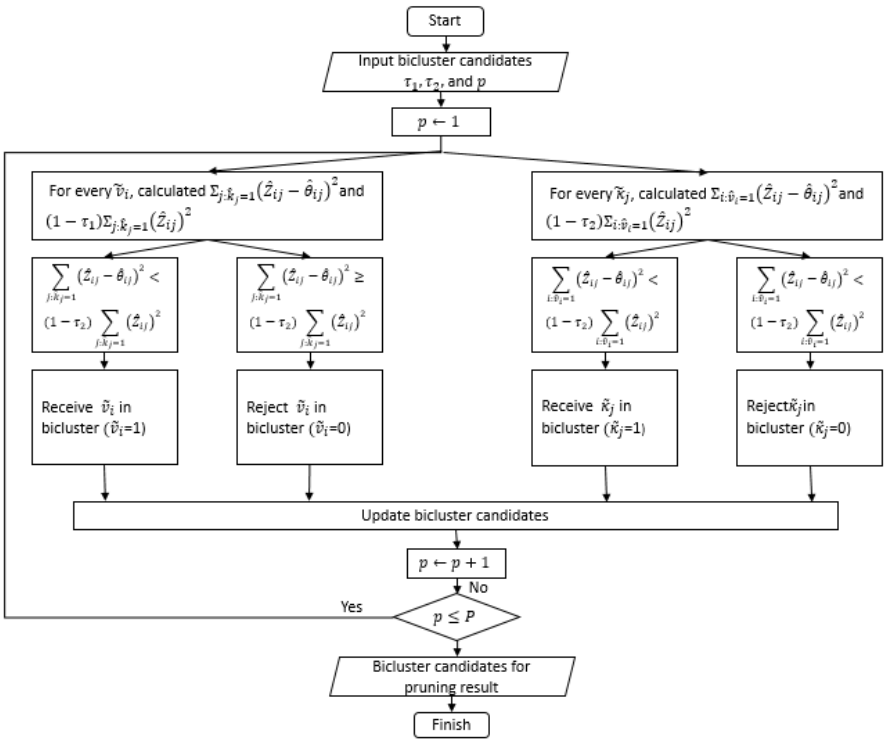

Figure 3: Flowchart of bicluster pruning

3. Results and discussion

3.1. Bicluster output measurement

We can measure the effectiveness of biclustering analysis using several measurements. In this study, we leverage two measurements, namely the variance of coherence and Jaccard index, which are explained as follows.

- Variance of coherence
  One common method to analyze the result of bicluster analysis measures the similarity or coherence of the bicluster. Coherence measurement can be done with coherence variance. Coherence variance obtained as the value of variance from rows and columns simultaneously. The equation wrote in Equation (14). **B** is a bicluster submatrix whose coherence variance will be calculated. The notation i is a row of bicluster **B** where $i = 1, ..., I$; notation j is a column of bicluster **B** where $j = 1, ..., J$. A bicluster expect to have a coherent nature, and then it is expected to have a coherence variance value low (Hartigan 1972).

$$Var(\mathbf{B}) = \frac{1}{|I||J|} \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{..})^2 \tag{14}$$

- Jaccard index
  The Jaccard index measures the similarity between two bicluster results. The index has a range of values from zero to one. A zero value means there is no similarity, while one means that the two biclusters are identical. This value used to compare two bicluster search algorithms. The Jaccard index value obtained in Equation (15). $\mathbf{B}_i$ is an i-th bicluster submatrix, and $\mathbf{B}_j$ is an j-th bicluster submatrix. The Jaccard index calculated by the number of biclusters i and j members, divided by the combined number of bicluster i and j. $n(B_i \cap B_j)$ is the number of entries of i-th bicluster and j-th bicluster results in rows. $n(B_i \cup B_j)$ is the number of data entries from the i-th bicluster and j-th bicluster which are a combined set (Grothaus, Mufti, and Murali 2006).

$$Jac(B_i, B_j) = \frac{n(B_i \cap B_j)}{n(B_i \cup B_j)} \tag{15}$$

3.2. Experimental results

The implementation is carried out on gene expression data, which has the type of genes in the column and observations on the rows. Biclustering algorithm is implemented using the R programming language version 3.6.0 (Team 2013) and the biclust package (Kaiser, Santamaria, Khamiakova, Sill, Theron, Quintales, Leisch, and Troyer 2018). The dataset used in this study was data on the expression of intestinal cancer genes. The data is accessed through the official website of Kaggle (https://www.kaggle.com/masudur/colon-cancer-gene-expression-data) on October 3, 2019, with the title: Colon Cancer Gene Expression Data. The rows in the dataset are different observations or individuals, and there are 62 observations in total. The column in the dataset is a type of gene, and there are 2000 different types of genes as a whole. The entry value in the data is the expression value of a gene in an observation. Each observation has a particular disease condition that is normal or abnormal. Abnormal conditions indicate that the individual has colon cancer. The value of the gene expression is a numeric number.

We apply the biclustering analysis using the plaid model method in the expression of colon cancer gene data using several parameter values. We can determine the values of the parameters based on the best practice (Turner *et al.* 2005). Implementation will be carried out six

times with particular scenarios based on the use of parameters chosen previously. The types of models used are interesting to be implemented in the data. The model used is broadly divided into two, namely a complete model consisting of $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$ (consists of the mean effect, row effect and column effect on bicluster), and the model with $\theta_{ijk} = \mu_k$(only consist of the mean effect). The threshold parameters ($\tau_1 and \tau_2$) also interesting to implement at different values. The threshold parameters ($\tau_1 and \tau_2$) functions as a barrier when bicluster pruning is done as described in Equations 10 and 11. Turner, Bailey, and Krzanowski (2005) suggested to use a threshold value between 0.5 and 0.7, so that we are interested in using a value of 0.5 and 0.7 in the implementation. The research scenarios used are as follow:

1. Model 1: $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$, and the threshold value $\tau_1 = 0.5$ and $\tau_2 = 0.5$
2. Model 2: $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$, and the threshold value $\tau_1 = 0.7$ and $\tau_2 = 0.7$
3. Model 3: $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$, and the threshold value $\tau_1 = 0.5$ and $\tau_2 = 0.7$
4. Model 4: $\theta_{ijk} = \mu_k$, and the threshold value $\tau_1 = 0.5$ and $\tau_2 = 0.5$
5. Model 5: $\theta_{ijk} = \mu_k$, and the threshold value $\tau_1 = 0.7$ and $\tau_2 = 0.7$
6. Model 6: $\theta_{ijk} = \mu_k$, and the threshold value $\tau_1 = 0.5$ and $\tau_2 = 0.7$

Table 1: Bicluster output with size

Scenario	**Bicluster**	**Size**
1	1	2x71
3	1	2x62
4	1	36x54
	2	4x100
	3	1x65
5	1	6x41
6	1	36x35
	2	18x22
	3	1x49
	4	1x69

The results of the bicluster obtained from the implementation can be seen in Table 1. Model 1 output one bicluster with the size of 2 x 71, which means that there are 71 genes from 2 observations in this bicluster. These genes have gene expression values that have characteristic patterns of bowel cancer. Two observations on the bicluster are observations that have an abnormal status. The results of Model 2 did not produce any bicluster, because of the high threshold value, i.e. $\tau_1 = 0.7$ and $\tau_2 = 0.7$. Moreover, the use of the model $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$ compounded it. The model 2 outputs small residual values, thus it difficults to find a bicluster. The results of Model 3 is one bicluster with the size of 2 x 62, which mean that there are 62 genes from 2 observations on this bicluster. These genes have gene expression values that form the characteristic pattern of cancer intestine. Two observations of the bicluster members are observations that have an abnormal status. The results of Model 4 are three biclusters. In the first bicluster, the size of the bicluster is 36 x 54, meaning that there are 54 genes from 36 observations on the bicluster. In the bicluster, 14 observations have normal status, while 22 other observations have abnormal status. Thus, it can be detected that 54 genes included in the first bicluster affect 22 observations that have abnormal status (have colon cancer) and 14 observations that have normal status, so it can be estimated that there is a possibility of genes in observations that have that normal status getting colon cancer. The same interpretation is given to the second and third bicluster on the results of Model 4. The results of Model 5 are one bicluster. In the bicluster, the size of the bicluster is 6 x 41, meaning that there are 41

genes from 6 observations on this bicluster. Four observations on the bicluster are observations that have normal status, while the other two are abnormal. That way, it can be detected that 41 genes included in the bicluster affect two observations that have abnormal status and four observations that have normal status, so the six observations need to be further studied about the status of intestinal cancer because it has a pattern on the value of gene expression. The results of Model 6 are four biclusters, as shown in Table 1. In the first bicluster, the size of the bicluster is 36 x 35, which mean that there are 35 genes from 36 observations in this bicluster. In the bicluster, the observation status is the same as the first bicluster in Model 4, i.e. 14 observations have normal status, while the other 22 observations have abnormal status. Thus, it can be detected that 35 genes included in the first bicluster affect 22 observations that have abnormal statuses and 14 observations that have normal status. Thus, it can be estimated that there is a possibility that genes in observations that have normal status are affected by intestinal cancer with the same interpretation of the results of second, third and fourth bicluster in Model 6. From Table 2, we can see the Jaccard index between models. The results of the coherence variance for each bicluster of each model can be seen in Table 3.

Table 2: Jaccard index of each model

	1	**2**	**3**	**4**	**5**	**6**
1	1	-	0.87	0.01	0	0.02
2	-	-	-	-	-	-
3	0.87	-	1	0.01	0	0.02
4	0.01	-	0.01	1	0.04	0.38
5	0	-	0	0.04	1	0.04
6	0.02	-	0.01	0.38	0.04	1

Table 2 has symmetrical entries, which mean that the values in the i-th row and j-column are the same as the j-row and i-th column entry values. The diagonal entries have value 1 because the Jaccard index counts the same bicluster members per total bicluster member count. The second row and column do not have any value, because, in Model 2, there is no bicluster at all, so the Jaccard index cannot be calculated. Based on Table 2, we can see that the largest index exists between Models 1 and 3 with 0.87, meaning that both scenes have 87% of the same bicluster members. Judging from the results of the two biclusters, both produce only one bicluster and only differ in the bicluster dimensions. The second-highest index value is between Models 4 and 6 with 0.38. The value is indeed not significant compared to index Models 1 and 3. Models 4 and 6 have similarities to the model used, and the value τ_1. The next most substantial index value already has a value below 0.1, meaning that the bicluster members that shared between the two scenarios do not reach 10% of the total number of bicluster members.

The interesting point about the index value is the two top values owned by the pair of models that differ only in the use of threshold values in the trimming process, namely Models 1 & 3, and 4 & 6. The use of threshold values $\tau_1 = 0.5$ and $\tau_2 = 0.5$ excludes results that have closeness based on the Jaccard index to the use of the values $\tau_1 = 0.5$ and $\tau_2 = 0.7$. On the other hand, the Jaccard index between Models with $\tau_1 = 0.5$ and $\tau_2 = 0.7$ has relatively proximity to Models that have $\tau_1 = 0.7$ and $\tau_2 = 0.5$, as seen from the index value which is only worth 0.04 in Models 5 & 6 and not even defined in Models 2 & 3.

Table 3 shows the values of the coherence variance based on the bicluster results for each model. The lower the variance value means that the more similar each data entry in the bicluster. The smallest variance value is owned by Model 1, followed by Model 3, with a variance value that is not much different. The value of variance in Models 4, 5, and 6 have a higher value than Models 1 and 3. Models 1 and 3 only produce one bicluster but have a small variance value. On the other hand, scenarios 4 and 5 have more biclusters, but the variance is quite large compared to Models 1 and 3. The fundamental difference between Models 1 &

Table 3: Variance of coherence of each model

Scenario	Bicluster	Size
1	1	1747.392
3	1	1707.997
4	1	7325.217
	2	1697.639
	3	There is only 1 observation
5	1	2252.233
6	1	5506.432
	2	3000.151
	3	There is only 1 observation
	4	There is only 1 observation

3 and 4 & 6 is the model parameter. Models 1 & 3 have more parameters, which produces a lower residual value. A small residual value makes finding bicluster more difficult, so that Models 1 & 3 provide one bicluster only, but have a low variance. In contrast, the models 4 & 6 use fewer parameters, then the residual value becomes large. Because of the large residual value, the bicluster is easier to obtain, but the biclusters have high variance values. By seeing the lowest coherence variance from Model 3, we claim that the biclustering analysis provides useful insights to analyze the gene expression of colon cancer data.

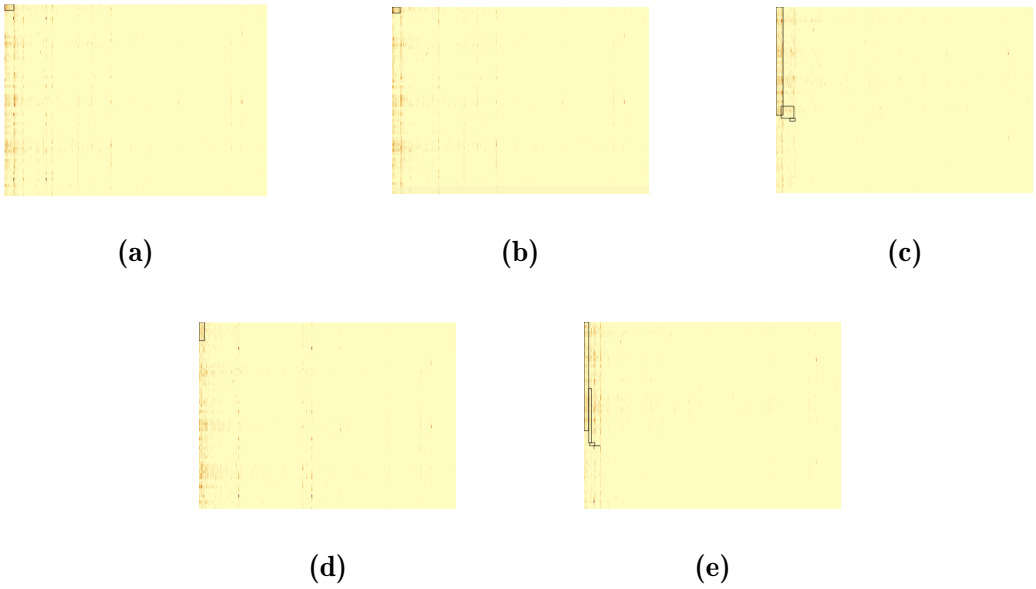

Figure 4: Visualization of biclusters made by: **(a)** Model 1 **(b)** Model 3 **(c)** Model 4 **(d)** Model 5 **(e)** Model 6

Figure 4 shows the visualization all biclusters in Models 1 to 6, except Model 2 because it has no bicluster output. The axes represent the original matrix of data, which x and y axes are gene types and observations, respectively. The highlighted boxes are the biclusters, which is a subset of a matrix. The visualizations are showing the sorted row and column indexes based on bicluster's membership. Therefore, all the boxes are located in the top-left corner.

4. Conclusion and future work

The plaid model in biclustering analysis is the process of partitioning data based on rows and columns simultaneously on a data matrix containing numeric numbers. The model has the

advantage that it can produce overlapping bicluster with the sum of the effects of the biclusters on the model. The model can also properly account for the effects of the grand mean, row effect, and column effect. Based on our empirical study, Model 3 output one bicluster with the lowest coherence variance. Model 3 contains the mean, row and column effects of the bicluster and has a threshold value $\tau_1 = 0.5$ and $\tau_2 = 0.7$. Based on the experiment results and analysis, we can conclude that by using a biclustering analysis with the plaid model, we can analyze the colon cancer data with low coherence variance. In the near future, we plan to conduct further analysis with other biclustering models. Also, it is interesting to see the effectiveness of biclustering analysis with different types of datasets.

Acknowledgment

This research was supported by PUTI research grant with contract number: NKB-1955/-UN2.RST/HKP.05.00/2020. We would like to thank our colleagues from Directorate of Research and Community Engagement Universitas Indonesia who provided insights and expertise to improve this research in innumerable ways.

References

Ardaneswari G, Bustamam A, Siswantining T (2017). "Implementation of Parallel k-means Algorithm for Two-phase Method Biclustering in Carcinoma Tumor Gene Expression Data." *AIP Conference Prosiding*, **1825**(1). URL https://doi.org/10.1063/1.4978973.

Busygin S, Prokopyev O, Pardalos PM (2008). "Biclustering in Data Mining." *Computers & Operations Research*, **35**(9), 2964–2987. URL https://doi.org/10.1016/j.cor.2007.01.005.

Chaturvedi A, Carroll JD (1994). "K-means, K-Medians and K-Modes: Special Cases of Partitioning Multiway Data." *In The Classification Society of North America (CSNA) Meeting Presentation.*

Cheng Y, Church G (2000). "Biclustering of Expression Data." *ISMB*, **8**, 93–103.

Denitto M, Bicego M, Farinelli A, Figueiredo MA (2017). "Spike and Slab Biclustering." *Pattern Recognition*, **72**, 186–195. URL https://doi.org/10.1016/j.patcog.2017.07.021.

Grothaus GA, Mufti A, Murali TM (2006). "Automatic Layout and Visualization of Biclusters." *Algorithms for Molecular Biology*, **1**(1). URL https://doi.org/10.1186/1748-7188-1-15.

Hartigan JA (1972). "Direct Clustering of a Data Matrix." *Journal of The American Statistical Association*, **67**(337), 123–129. URL https://doi.org/10.1080/01621459.1972.10481214.

Kaiser S, Santamaria R, Khamiakova T, Sill M, Theron R, Quintales L, Leisch F, Troyer ED (2018). "biclust: BiCluster Algorithms." *R package version 2.0.1.* URL https://CRAN.R-project.org/package=biclust.

Kasim A, Shkedy Z, Kaiser S, Hochreiter S, Talloen W (2016). *Applied Biclustering Methods for Big and High-dimensional Data Using R.* CRC Press. ISBN 9781482208238, URL https://www.amazon.com/Applied-Biclustering-Methods-High-Dimensional-Biostatistics/dp/1482208237.

Latief MA, Siswantining T, Bustamam A, Sarwinda D (2019). "A Comparative Performance Evaluation of Random Forest Feature Selection on Classification of Hepatocellular Carcinoma Gene Expression Data." In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, pp. 1–6. IEEE.

Lazzeroni L, Owen A (2000). "Plaid Models for Gene Expression Data." *Statistica Sinica*, **12**, 61–86.

Rokach L, Maimon O (2010). *Classification Trees, Data Mining and Knowledge Discovery Handbook*. Springer. ISBN 9780387098227.

Team RC (2013). "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing*. URL http://www.R-project.org/.

Turner H, Bailey T, Krzanowski W (2005). "Improved Biclustering of Microarray Data Demonstrated Through Systematic Performance Tests." *Computational Statistics & Data Analysis*, **48**(2), 235–254. URL https://doi.org/10.1016/j.csda.2004.02.003.

Affiliation:

Titin Siswantining
Department of Mathematics
Faculty of Mathematics and Natural Science
Universitas Indonesia
Depok, West Java, 16424
E-mail: titin@sci.ui.ac.id
URL: https://staff.ui.ac.id/titin

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2020-07-02
Accepted: 2020-09-07