MATEMATIKA, Volume 39, Number 3, 227–238
© Penerbit UTM Press. All rights reserved

Biclustering Models Under Collinearity in Simulated Biological Experiments

[1,2]C. N. Nnamani* and [1]N. Ahmad

[1]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia

[2]Department of Statistics, Faculty of Physical Sciences, Ahmadu Bello University
Zaria, Nigeria

*Corresponding author: norhaiza@utm.my

Article history
Received: 5 January 2023
Received in revised form: 15 October 2023
Accepted: 31 October 2023
Published on line: 1 December 2023

Abstract Biclustering models allow simultaneous detection of group observations that are related to variables in a data matrix. Such methods have been applied in biological data for classification. Collinearity is a common feature in biological data as there exist interactions between genes and proteins in their respective pathways. These relationships could seriously reduce the efficiency of biclustering models. In this study, synthetic data are generated to investigate the effect of collinearity on the performance of biclustering models. Specifically, the data are generated and induced with varying degrees of collinearity using Cholesky decomposition, and are implanted with biclusters to produce different sets of synthetic data. The effectiveness of three models namely Biclustering by Cheng and Church (BCCC), Spectral Bicluster (BCSpectral) and Plaid Model in correctly detecting three types of biclusters in the generated data matrix were compared. The results show that all the models investigated are sensitive to changes in the level of collinearity. At low collinearity, all biclustering models detected the implanted biclusters in the data correctly. However, as the level of collinearity in the data increased, the proportion of detected biclusters captured by the models reduced. In particular at high collinearity, BCCC outperformed the other two models with Jaccard coefficients as high as 0.75 and 0.873 for one and two implanted biclusters respectively.

Keywords two-way cluster; collinearity; biological data simulation

Mathematics Subject Classification 62P10, 62H86.

1 Introduction

Advanced technology has enabled the gathering of large volumes of data from throughput biological experiments within a short space of time. The substantial amount of data generated from these experiments, and the amalgamation of different biological sources and systems leads to the creation of biological big data which is more difficult to analyze [1–4]. In order to mine information from such datasets, many multivariate statistical techniques have been developed including biclustering [5, 6].

Biclustering model, also known as two-way clustering, is a clustering method that can concurrently detect groups or layers of observations that are related with respect to certain variables such that the within-group dissimilarity is minimal. These models consist of algorithms that allow the sub-partitioning of rows and columns in a data matrix in the form of submatrix known as bicluster. Certain overlapping biclustering models are able to capture member clusters contained in two or more biclusters in the data matrix. Such methods have been applied to many biological data for the classification and identification of biological entities [7–9].

An inherent feature of biological data is collinearity. Collinearity is a measure of the linear dependency among the predictor variables in a multivariable process and it is an intrinsic property in many multivariable data especially from biological experiments [10]. The presence of collinearity is likely to lead to non-unique and unreliable estimates as a result of the singularity in the covariance matrix [11–14]. Hence, it is important to ensure that the collinearity level in a dataset is within tolerable limits before reliable inferences can be made from any given set of observations.

Different algorithms in the biclustering models cater to different data behaviors. For example, Plaid Model is suitable for additive biclusters while other models such as the Maximum Similarity Biclustering model are appropriate for multiplicative biclusters [15]. Studies have also been conducted to investigate the performance of several biclustering models under different data structures. For instance, [16] investigated the biclustering algorithms with different percentage of missing data; Another study by [17], introduced an approach to assess the importance of local patterns in biclusters with additive, multiplicative, symmetric, order-preserving and plaid coherencies. Other studies include the evaluation of the bicluster's size and recovery score on different biclustering models [18–20]. In addition, [21] tested the performance of the biclustering algorithms for noisy data by embedding non-overlapping constant and additive patterns of columns and rows with Gaussian noise of variance ranging from 0.1 to 0.5; Also, [22] studied the effect on the presence of noise to identify simultaneous clusters of gene expressions. They recommended the control of noise in biclustering algorithms to get reliable results. The presence of noise in gene expression data which may lead to phenotic variability has been particularly highlighted by [23].

In this study, a different type of data behaviour was investigated, focusing on collinear and invariant features in the data. Specifically, the reliability of several biclustering models in the presence of collinearity, featuring at least one bicluster whose entries are constant, constant in the rows, and constant in the columns respectively. The three biclustering models examined in this study are Biclustering by Cheng and Church (BCCC), Plaid Model, and Spectral Bicluster (BCSpectral). These models are chosen since they are amongst the common methods used in bioinformatics. In order to evaluate their performances, the level of collinearity in a dataset was progressively increased and several biclusters were implanted at different levels of collinearity: low, moderate, and high. Then, the ability of the models to detect the implanted bicluster at that level of collinearity was examined. At each level of collinearity, one and two known biclusters are added. This is to verify whether the number of biclusters in the data matrix has any effect on the ability of the investigated models to resist collinearity. Also examined was whether the models are sensitive to the type of bicluster. This was done by varying the types of biclusters planted into the datasets, and carrying out the analysis using the respective models

The rest of the paper is organized as follows: Section 2 presents the biclustering techniques used in the study and Section 3 explains the procedure for simulating the generated datasets used in the comparison of biclustering models. In Section 4, the results and findings on the model performance comparison correctly detecting biclusters in the presence of collinearity in the simulated data were discussed. Section 5 concludes the findings of the study.

2 Biclustering Models

A biological dataset can be presented in the form of a rectangular data matrix, Y, with elements y_{ij} such that $i = 1, \ldots, n; j = 1, \ldots, m$; i and j are the row index for n samples and column index for m features respectively. A bicluster is a subset of rows that show similar behaviors across a subset of columns, and vice-versa, displayed as a submatrix of Y. The models to be investigated in this study are Biclustering by Cheng and Church (BCCC), Plaid Model (PM) and Spectral Bicluster (BCSpectral). Each algorithm is applied to the generated datasets in the form of data matrix Y and the resulting biclusters are observed. these algorithms are presented in this section. Also presented is a discussion of some applications of biclustering.

2.1 Biclustering by Cheng and Church (BCCC)

The BCCC is a biclustering algorithm that extracts clusters that have constant values throughout, or constant values in either the rows or in the columns [24, 25]. It is also called the δ-biclustering algorithm because it seeks for biclusters with a mean squared residual score less than a given threshold (δ). The threshold can be determined by a greedy iterative search method as recommended by [26]. Given a data matrix and its submatrices, the mean squares residual score is given by the following,

$$\psi(Y) = \frac{1}{nm} \sum_{i=1}^{n} (y_{ij} - \alpha_i - \beta_j - \mu)^2 \tag{1}$$

where μ is the mean for the overall values, while α_i and β_j are the row and column means respectively, as defined in equations (2) - (4).

$$\mu = \frac{1}{nm} \sum_{i=1}^{n} \sum_{i=1}^{m} y_{ij} \tag{2}$$

$$\alpha_i = \frac{1}{m} \sum_{j=1}^{m} y_{ij} \tag{3}$$

$$\beta_j = \frac{1}{n} \sum_{i=1}^{n} y_{ij} \tag{4}$$

2.2 Plaid Model

The Plaid model expresses a data matrix, as a sum of additive-biclusters [27–29]. Every object in the bicluster is expressed within, and only within, those objects in bicluster k. Algebraically, each entry in the data matrix corresponds to the following expression:

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} \mu_k \rho_{ik} \kappa_{jk} \tag{5}$$

where μ_0 is the overall mean of the data, μ_k is the mean in bicluster k, ρ_{ik} is 1 if observation i is in the k th bicluster (zero otherwise), and ρ_{jk} is 1 if variable j is in the k th bicluster (zero otherwise).

In order for every object and every variable to be in exactly one bicluster, the stipulated conditions are that $\sum k\rho_{ik} = 1$ for all i, and $\sum k\kappa_{jk} = 1$ for all j, respectively. Thus, for overlapping clusters to be detected, it is allowed that $\sum k\rho_{ik} > 1$ for some i, or $\sum k\kappa_{jk} > 1$ for some j. However, to allow these conditions to exist, there are likely to be some objects or variables that do not fit well into any of the k th-biclusters. These are called ragbag biclusters, and the scenario is modeled by $\sum k\rho_{ik} = 0$ for some i, or $\sum k\kappa_{jk} = 0$ for some j.

2.3 Spectral Bicluster (BCSpectral)

The BCSpectral constructs a checkerboard structure with the given dataset [30,31]. The algorithm reorders the data and computes a singular value decomposition to get eigenvalues and eigenvectors. It then detects biclusters starting with the largest or second largest eigenvalue. The bicluster detection is guided by the normalisation method chosen. The three normalisation methods are Independent Rescaling of Rows and Columns (IRRC), in which the rows and columns are rescaled independently; bistochastization, in which simultaneous recalling of rows and columns are carried out; and log transformation is applied to the given data matrix. Biclusters are reported if they satisfy the conditions of minimum number of rows, minimum number of columns and maximum variation allowed within each bicluster.

3 Simulation Procedures and Model Evaluation

3.1 Simulated Data

In order to effectively evaluate the performance of the biclustering models, the effects of collinearity on sets of simulated data were used. The use of these generated synthetic data offers the opportunity to investigate the effectiveness of the models in detecting implanted biclusters on various characteristics of data. A total of 540 datasets were generated and examined. Specifically, each dataset is composed of 100 row observations and 50 column variables that form a 100× 50 data matrix Y, with its base elements sampled from an i.i.d standard normal distribution. This distribution is selected since biological expression data are typically generated in multiples of thousands, and would tends toward normality asymptotically [32,33].

Each generated data matrix contains one or two implanted biclusters. Each bicluster consists of four rows and four columns for each one of three types i.e. constant biclusters, constant row biclusters, and constant column biclusters. These biclusters are planted in the dataset for the purpose of detection by the algorithms. Figure 1 illustrates three types of implanted biclusters: Figure 1(a) shows an example when all entries in the implanted matrix contain the same values, indicating constant bicluster; Figure 1(b) shows an example when certain row-wise entries contain similar values, indicating constant row biclusters; and Figure 1(c) shows an example when certain column-wise entries contain similar values, reflecting constant column biclusters. Thus, each element y_{ij} in the simulated data Y can be expressed as in the equation $y_{ij} = x_{ij} + \varepsilon_{ij}$ where x_{ij} is the element in the implanted bicluster and ε_{ij} is the background element sampled from a standard normal distribution.

In addition, each dataset was perturbed with different levels of collinearity using Cholesky decomposition. Cholesky decomposition is a method of matrix decomposition that splits a positive definite square matrix into its lower triangular matrix and its transpose. The Cholesky decomposition has been shown to simplify calculations involving correlations and reduce the errors of approximation [34, 35].

$$(a) \begin{pmatrix} 20 & 20 & 20 & 20 \\ 20 & 20 & 20 & 20 \\ 20 & 20 & 20 & 20 \\ 20 & 20 & 20 & 20 \end{pmatrix} \quad (b) \begin{pmatrix} 10 & 10 & 10 & 10 \\ 25 & 25 & 25 & 25 \\ 20 & 20 & 20 & 20 \\ 15 & 15 & 15 & 15 \end{pmatrix} \quad (c) \begin{pmatrix} 10 & 25 & 20 & 15 \\ 10 & 25 & 20 & 15 \\ 10 & 25 & 20 & 15 \\ 10 & 25 & 20 & 15 \end{pmatrix}$$

Figure 1: Ilustration of implanted biclusters (a) constant biclusters, (b) constant row biclusters, and (c) constant column biclusters.

3.2 Performance and Evaluation

There are several methods that can be used to ascertain the degree of collinearity in a dataset, such as Variance Inflation Factor, Condition Index, Condition Number, and Tolerance [36, 37]. In this study, Condition Number (CN) was used as the test statistic to gauge the level of collinearity. Let a matrix Y consist of m-dependent variables, and $\lambda_1, \lambda_2, \ldots, \lambda_m$, are the eigenvalues of Y^TY. The CN is given by $\dfrac{\lambda_i}{\lambda_{max}}$. Thus, with respect to the maximum eigenvalue of λ_{max}, if there are small $\lambda_i's$, then there exists multicolinearity in the data matrix Y. As a guideline, CN <100 indicates low collinearity; 100$\leq CN \leq$1000 indicates moderate collinearity while CN >1000 implies that there exists a strong linear relationship among some of the variables [38]. The strength of each biclustering model is tested by its ability to detect true biclusters on data with different degrees of collinearity. The variation in the settings will indicate the extent of failure or success of the models as a result of the induced collinearity. A measure called the Jaccard coefficient index is used to measure the model's success by comparing the implanted bicluster with the bicluster produced from the biclustering model [39, 40]. The Jaccard coefficient (JC) index for two sets of biclusters, b_1 and b_2, is given by

$$JC = S(b_1, b_2) = \left| \frac{b_1 \cap b_2}{b_1 \cup b_2} \right| \in [0, 1] \tag{6}$$

Table 1: Model performance using Jaccard coefficient indices based on data containing three levels of collinearity and one implanted bicluster with different types.

		Degree of Collinearity		
Model	Bicluster Type	Low	Moderate	High
BCCC	Constant	1.000	0.875	0.750
	Constant Row	1.000	0.739	0.499
	Constant Column	1.000	0.750	0.750
BCSpectral	Constant	1.000	0.590	0.390
	Constant Row	1.000	0.380	0.106
	Constant Column	1.000	0.044	0.245
Plaid	Constant	1.000	0.333	0.289
	Constant Row	1.000	0.750	0.750
	Constant Column	1.000	0.333	0.282

where $b_1 \cap b_2$ is the number of cells in their intersection, and $b_1 \cup b_2$ is the number of cells in their union. The Jaccard coefficient is the percentage of cells shared by both biclusters. A maximum score of 1 indicates that both algorithms identified exactly the same biclusters. A score of 0 indicates that both models identified completely different biclusters.

4 Results and Finding

In order to ensure that the effects of the biclustering models are properly investigated, different numbers of implanted biclusters are conditioned to the generated data. This section presents the results following the execution of the biclustering models on the simulated data. The strength of each biclustering model is tested by its ability to detect true biclusters based on the degree of collinearity between the variables in the data, the number of implanted biclusters, and the type of bicluster. The numerical results of the Jaccard Coefficient (JC) indices obtained from the three models for three levels of collinearity at one and two implanted biclusters are shown in Table 1 and Table 2 respectively. The implanted biclusters also contain one of three behaviours of constant bicluster, constant row bicluster and constant column bicluster as described in section 3.1. Each entry in the table represents the JC index results based on the average of 100 runs of simulations. A maximum JC index value of 1 indicates that the model was able to detect exactly the implanted biclusters in the data. The lowest JC index value of 0 indicates that the model was not able to detect any of the implanted biclusters.

It can be observed that BCCC, BCSpectral and Plaid are sensitive to changes in the level of collinearity and bicluster types, regardless of the number of biclusters that exist in the data. As shown in Table 1 and Table 2, all models indicate JC index values of one at low level of collinearity when detecting one or two implanted biclusters respectively. As the collinearity increases from moderate to high, the JC index values for all models appear to reduce with varying results. These patterns can be seen in the downward trend of the graphs in Figure 2. Figure 2 shows the performance of the models in detecting the implanted biclusters with respect to collinearity. The different colours in the graph shows the different bicluster types of

Table 2: Model performance using Jaccard coefficient indices based on data containing three levels of collinearity and two implanted biclusters with different types.

		Degree of Collinearity		
Model	Bicluster Type	Low	Moderate	High
BCCC	Constant	1.000	0.936	0.873
	Constant Row	1.000	0.746	0.746
	Constant Column	1.000	0.873	0.866
BCSpectral	Constant	1.000	0.000	0.140
	Constant Row	1.000	0.000	0.456
	Constant Column	1.000	0.000	0.000
Plaid	Constant	1.000	0.414	0.406
	Constant Row	1.000	0.512	0.481
	Constant Column	1.000	0.000	0.000

constant, constant row and constant column implanted in the data.

It appears that BCCC seem to perform comparatively better than the other models since the degree of reduction of JC index values is less prominent, as collinearity in the data increased for all bicluster types. Specifically, it shows a maximum JC index value of 0.875 and a minimum JC index value of 0.499 (Table 1) for one-bicluster experiment; and JC index values ranging 0.936 to 0.74 for a two bicluster experiment (Table 2). BCSpectral appear to detect less than 50% of the implanted biclusters at high collinearity for one-bicluster and two-bicluster experiments respectively. At one-bicluster experiment, Plaid Model performed well for the constant row experiments at all levels of the collinearity, detecting about 75% of the implanted biclusters.

The robustness of the three models also vary as the number of implanted biclusters with different types are implanted in the data. The BCSPectral and Plaid Model appear to be severely affected as the collinearity is steadily increased. When two biclusters of type constant column were implanted, these models were unable to detect any of the implanted biclusters at moderate and high levels of collinearity. On the other hand, BCCC performed even better as the number of implanted biclusters are added. For this model, increasing the number of biclusters leads to an increase in the total number of objects in the search space. This leads to higher success in the detection when compared to fewer biclusters in the dataset.

5 Conclusion

Biclustering provides an approach to extract two-way characterisation of behaviour from a biological dataset by detecting biclusters. Biological datasets are often rich in collinear features, reflecting the interconnectedness of biological processes. However, collinearity can confound biclustering models, making it challenging to uncover distinct and biologically meaningful patterns. In this study, challenges posed by collinear biological data when applying biclustering models are explored based on simulated studies. Other features including the presence of multiple biclusters and different bicluster types are also considered in the generated data. It can be seen that although that the three models: BCCC, BCSpectral and Plaid Models have the ability to identify meaningful subsets of data that exhibit coherent block structure patterns, they behaved differently as the level of collinearity is increased. Also, if there are more than one cluster with similar behaviour in the data, the robustness of the three models would vary as well. Based on the simulated data, BCCC outperformed the other models for moderate to high collinearity, for at least two implanted biclusters. To address the challenges posed by collinear biological data, several recommendations could be considered prior to applying the biclustering models. These include using data-prepocessing techniques such as feature selection, dimensionality reduction, and data transformation to help prepare the data for more effective biclustering. In addition, as these biclustering models use an optimization approach, fine-tuning the algorithm parameters should be considered, particularly when dealing with collinear biological data. Researchers should seek to strike a balance between capturing biologically relevant biclusters and avoiding overfitting.

In conclusion, the challenges associated with biclustering models when confronted with collinear biological data and multiple biclusters highlight the need for a thoughtful and holistic approach. By implementing the recommended strategies, researchers can overcome these limitations and advance their understanding of complex biological systems, ultimately contributing to advancements in biomedical research.

Acknowledgments

The authors would like to thank Malaysia Ministry of Higher Education and Universiti Teknologi Malaysia for the financial funding via the grant $PY/2019/02100$ (UTM-FR). $Q.J130000.2554.21H70$.

References

[1] G. Florimbi, E. Torti, S. Masoli, E. D'Angelo, G. Danese, and F. Leporati. Exploiting multi-core and many-core architectures fo *J. Parallel Distrib. Comput.* vol. 126, pp. 48-66, Apr. 2019, doi: 10.1016/J.JPDC.2018.12.004.

[2] R. Duchesne, A. Guillemin, F. Crauste and O. Gandrillon Calibration, Selection and Identifiability Analysis of a Mathematical Model of the in vitro Erythropoiesis in Normal and Perturbed Contexts. *In Silico Biol.* vol. 12, pp. 1-15, Apr. 2019, doi: 10.3233/ISB-190471.

[3] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudne4, C. Evangelista, et al. NCBI GEO: mining tens of millions of expression profiles-database and tools update. *Nucleic Acids Res..* vol. 35, pp. 760-765, Nov.2006, doi: 10.1093/nar/gkl887.

[4] M. Trovati, R. Hill, A. Anjum and Z. Y. Shao, L. Liu. Big-Data Analytics and Cloud Computing. *Theory and Algorithms in Big-Data Analytics and Cloud Computing.* Switzerland: Springer International Publishing, 2015.

[5] D. S. Rodriguez-Baena, A. J. Perez-Pulido and J. S. Aguilar-Ruiz A biclustering algorithm for extracting bit-patterns from binary datasets. *Bioinformatics.* vol. 27, no. 19, pp. 2738–2745, Aug. 2011, doi: 10.1093/bioinformatics/btr464.

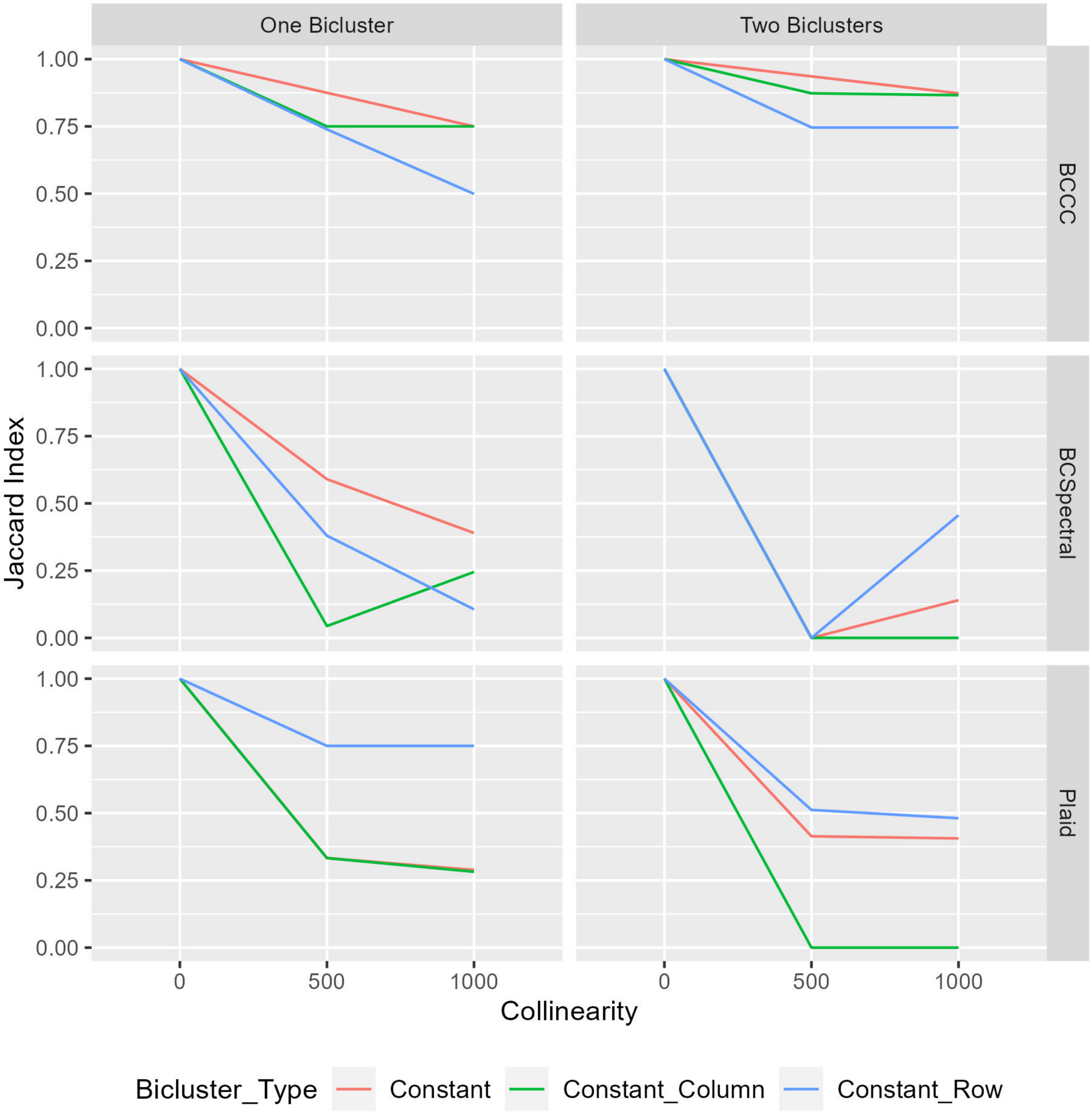

Figure 2: Effect of collinearity and bicluster types on biclustering models

[6] SMM Hossain, S. Ray, TS Tannee, A. Mukhopadhyay, Analyzing Prognosis Characteristics of Hepatitis C using a Biclustering Based Approach, *Procedia Computer Science*. Volume 115, 2017, pp 282-289, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.09.136.

[7] H.T. Turner, C.B. Trevor, J.K. Wojtek and A.H. Cheryl. Biclustering Models for Structured Microarray Data. *Transactions on computational biology and bioinformatics* Vol. 2(4), pp. 316 – 329, 2005

[8] H.M. Chu, J.X. Liu, and K. Zhang. A binary biclustering algorithm based on the adjacency difference matrix for gene expression data analysis. *BMC Bioinformatics* 23, 381, pp. 1 – 16, 2022. https://doi.org/10.1186/s12859-022-04842-4

[9] M. Ramkumar, N. Basker, D. Pradeep, Ramesh Prajapati, N. Yuvaraj, R. Arshath Raja, C. Suresh, Rahul Vignesh, U. Barakkath Nisha, K. Srihari, Assefa Alene. Healthcare Biclustering-Based Prediction on Gene Expression Dataset. *BioMed Research International* Vol. vol. 2022, Article ID 2263194, pp. 1 - 7, 2022, doi: 10.1155/2022/2263194

[10] L. Bobrowski. Biclustering Based on Collinear Patterns. *Bioinformatics and Biomedical Engineering (IWBBIO).* 2017, doi: 10.1007/978-3-319-56148-611.

[11] J. Magidson, New Perspectives in Partial Least Squares and Related Methods, Springer Proceedings in Mathematics & Statistics. *Data Min. Knowl. Discov.*. New York: Springer, Aug. 2013, pp. 65–78.

[12] G. Khalaf and M. Iguernane. Multicollinearity and A Ridge Parameter Estimation Approach. *J. Mod. Appl. Stat. Methods.*. vol. 15, no. 2, pp. 400-410, Jan. 2016, doi:10.22237/jmasm/1478002980.

[13] M. El-Dereny and N. I. Rashwan. Solving Multicollinearity Problem Using Ridge Regression Models. *Int. J. Contemp. Math. Sci.*. vol. 6, no. 2, pp. 585-600, 2011.

[14] Y. Asar, A. KaraibrahimoĞlu and A. Genç. Modified ridge regression parameters: A comparative Monte Carlo study. *J. Math. Stat.* , vol. 43, no. 5, pp. 827-841, Nov. 2014.

[15] V.A. Padilha, R.J.G.B. Campello. A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics* Vol. 18,55, pp. 1 – 25, 2017, doi: 10.1186/s12859-017-1487-1

[16] J. Li, J. Reisner, H. Pham, S. Olafsson, S. Vardeman. Biclustering with missing data, *Information Sciences* Volume 510,2020, pp 304-316, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.09.047.

[17] R. Henriques and S. C. Madeira. BSig: evaluating the statistical significance of biclustering solutions. *Data Min. Knowl. Discov.*. vol. 32, no. 32, pp. 124-161, Jan. 2018.

[18] J. Dale, A. Nishimoto and T. Obafemi-Ajayi Performance Evaluation and Enhancement of Biclustering Algorithms. *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods.* ICPRAM, pp. 202-2013, 2018.

[19] W. Yoo, R. Mayberry, S. Bae, K. Singh and J. W. Lillard A Study of Effects of Multi-Collinearity in the Multivariable Analysis. *Int. J. Appl. Sci. Technol.*. vol. 4, no. 5, pp. 9-19, Oct. 2014.

[20] R. Henriques and S. C. Madeira Calibration, Selection and Identifiability Analysis of a Mathematical Model of the in vitro Erythropoiesis in Normal and Perturbed Contexts. *IIEEE/ACM Transactions on Computational Biology and Bioinformatics.* vol. 12, no. 4, Aug. 2015, doi: 10.1109/TCBB.2014.2388206.

[21] H. Zhao, A. Wee-Chung Liew, D. Z. Wang and H. Yan Biclustering Analysis for Pattern Discovery: Current Techniques, Comparative Studies and Applications. *Curr Bioinform.* vol.7, no.1, pp. 43-55, Mar. 2012, doi: 10.2174/157489312799304413.

[22] M. Fillippone, M. Francesco and R. Stefano. Stability and Performance in Biclustering Algorithms. *Computational Intelligence Methods for Bioinformatics and Biostatistics. 5th International Meeting (CIBB).* vol. 2008, pp. 91 - 101, 2018.

[23] J. M. Raser and E. K. O'shea Noise in Gene Expression: Origins, Consequences, and Control. *Science.* vol. 309, no. 5743, pp. 2010-2013, Sep.2005, doi: 10.1126/science.1105891.

[24] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics..* vol. 22, no. 9, pp. 1122-1129, May 2006, doi: 10.1093/bioinformatics/btl060.

[25] Y. Cheng and G. Church Biclustering of Expression Data. *Proceeding of the Eighth International Conference on Intelligent Systems for Molecular Biology..* 2000, pp. 93–103.

[26] Effective Biclustering Algorithm for Time-Series Gene Expression Data. In: Wang, X., Pedrycz, W., Chan, P., He, Q. (eds) Machine Learning and Cybernetics. *Communications in Computer and Information Science.Springer, Berlin, Heidelberg..* vol. 481.2014, pp. 93–103. doi.org/10.1007/978-3-662-45652-1-12

[27] L. Lazzeroni and A. Owen Plaid Model for Gene Expression Data. *Statistica Sinica.* vol.12, no. 1, pp. 61-86, Jan. 2002.

[28] A. Kasim, Z. Shkedy, S. Kaiser, S. Hochreiter and W. Talloen Biclustering with Flexible Plaid Models to Unravel Interactions between Biological Processes. *Applied Biclustering Methods for Big and High-Dimensional Data Using R.* Chapman & Hall CRC Biostatistics Series, 2016.

[29] H. A. Majd, S. Shahsavari, A. R. Baghestani, S. M Tabatabaei, N. K Bashi and M. R. Tavirani Evaluation of Plaid Models in Biclustering of Gene Expression Data. *Applied Biclustering Methods for Big and High-Dimensional Data Using R.* Chapman & Hall CRC Biostatistics Series, 2016.

[30] Y. Kluger, R. Basri, J. T. Chang and M. Gerstein. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome Res..* vol. 13, no. 4, pp. 703-716, Apr. 2003, doi: 10.1101/gr.648603.

[31] Yin, L , Liu, YG Ensemble biclustering gene expression data based on the spectral clustering *Neural Computing & Applications* Volume30. Issue8. pp. 2403-2416. 2018. DOI10.1007 s00521-016-2819-1

[32] A. Sugolov, E. Emmenegger, A.D. Paterson et al. Statistical Learning of Large-Scale Genetic Data: How to Run a Genome-Wide Association Study of Gene-Expression Data Using the 1000 Genomes Project Data. *Stat Biosci.* 2023. https://doi.org/10.1007/s12561-023-09375-9

[33] L.J. Cardinal Central tendency and variability in biological systems: Part 2. *J Community Hosp Intern Med Perspect.* Oct 19;5(5):28972., 2015, doi: 10.3402/jchimp.v5.28972. PMID: 26486117; PMCID: PMC4612486

[34] H. Siaby-Serajehlo, M. Rostamy-Malkhalifeh, F. Hosseinzadeh Lotfi and M. H. Behzadi. Usage of Cholesky Decomposition in order to Decrease the Nonlinear Complexities of Some Nonlinear and Diversification Models and Present a Model in Framework of Mean-Semivariance for Portfolio Performance Evaluation. *Adv. Oper. Res.*. vol. 2016, pp. 1-9, Mar. 2016, doi: 10.1155/2016/7828071.

[35] L. Freitag, S. Knecht, C. Angeli and M. Reiher. Multireference Perturbation Theory with Cholesky Decomposition for the Density Matrix Renormalization Group. *J. Chem. Theory Comput.*. vol. 13, no. 2, pp. 451-459, 2017.

[36] N. Adnan, M. H. Ahmad and R. Adnan. A Comparative Study on Some Methods for Handling Multicollinearity Problems. *MATEMATIKA*. vol. 22, 2006.

[37] J. J. Al-Jararha. New Approaches for Choosing the Ridge Parameters. *Hacettepe J. Math. Stat.* vol. 47, no. 6, pp. 1625-1633, 2018.

[38] S. Babichev, V. Osypenko, V. Lytvynenko, M. Voronenko and M. Korobchynskyi. Comparison Analysis of Biclustering Algorithms with the use of Artificial Data and Gene Expression Profiles. *38th International Conference on Electronics and Nanotechnology, ELNANO (IEEE).* 2018, pp. 298–304.

[39] A. K. Gupta and N. Sardana Significance of Clustering Coefficient over Jaccard Index *Eighth International Conference on Contemporary Computing (IC3)* 2015, pp. 463-466, doi: 10.1109 IC3.2015.7346726.

[40] B.S. Haifa and E Mourad A New Study on Biclustering Tools, Biclusters Validation and Evaluation Functions. *International Journal of Computer Sci. and Engn Survey (IJCSES)* Vol.6, No.1, pp. 1 – 13, 2015.