ENTHUSIASTIC
INTERNATIONAL JOURNAL OF APPLIED STATISTICS AND DATA SCIENCE
Volume 5, Issue 1, April 2025, pp. 101-106

Genetic Cluster Analysis of Insulin Resistance Using KNN Imputation and FABIA-CCA Biclustering

Ditoprasetyo Rusharsono Soemarso [a,1], Titin Siswantining [a,2,*], Setia Pramana [b,3]

[a] Universitas Indonesia, Jl. Margonda Raya, Pondok Cina, Kecamatan Beji, Depok, 16424, Indonesia
[b] Politeknik Statistika STIS, Jl. Otto Iskandardinata No. 64C 1, Jakarta 13330, Indonesia
[1] diditsoemarso@gmail.com; [2] titin@sci.ui.ac.id*; [3] setia.pramana@stis.ac.id
* Corresponding author

ARTICLE INFO

Keywords
Biclustering
Gene Expression
Mean Squared Error
Insulin Resistance
Missing Value Imputation

ABSTRACT

Type 2 diabetes mellitus (T2DM) is a metabolic disorder primarily driven by insulin resistance, involving complex genetic regulation. Understanding the molecular mechanisms underlying insulin resistance is crucial for identifying therapeutic targets. This study compared the performance of two biclustering algorithms, factor analysis for bicluster acquisition (FABIA) and the Cheng and Church algorithm (CCA), in analyzing gene expression data associated with insulin resistance. Using the GSE19420 dataset, simulated missing values were introduced to evaluate the robustness of both methods. Results showed that CCA consistently achieved lower mean squared error (MSE) in reconstructing gene expression patterns, suggesting higher accuracy in capturing co-expression structures. Nevertheless, FABIA effectively detected sparse, biologically relevant clusters. Notably, key genes such as MYO5B, DLG2, AXIN2, and PTK7 were identified within the biclusters, supporting their involvement in insulin signaling and metabolic regulation. These findings underscore the need to select biclustering methods that align with specific analytical goals and offer insights into gene networks involved in insulin resistance.

1. Introduction

Type 2 diabetes mellitus (T2DM) is one of the most common metabolic disorders, characterized by insulin resistance that disrupts blood glucose regulation. If not properly managed, T2DM can lead to severe complications such as cardiovascular disease and nephropathy [1]. Genetic factors play a crucial role in insulin resistance, yet the molecular mechanisms underlying this condition remain largely unclear, making it challenging to develop effective therapies [2].

Gene expression studies are essential for identifying genetic patterns associated with insulin metabolism. However, analyzing microarray datasets presents challenges such as high dimensionality and variability in expression levels [3]. One common issue in gene expression analysis is missing data, which may arise from experimental limitations or technical errors [4]. Although the dataset used in this study is originally complete, simulated missing values were introduced at levels ranging from 5% to 50% (in increments of 5%) to evaluate the robustness of biclustering methods under varying degrees of data loss [5], [6].

p-ISSN 2798-253X
e-ISSN 2798-3153

Biclustering methods are useful for identifying coherent submatrices in gene expression data—groups of genes co-expressed under a subset of conditions. Among various approaches, factor analysis for bicluster acquisition (FABIA) [7], [8] and the Cheng and Church algorithm (CCA) [9] are two widely used techniques. FABIA applies sparse factor models and has shown strong performance in detecting rare or weakly expressed gene patterns [9], [10]. However, its sensitivity to noise and computational complexity can be limiting in large or incomplete datasets. CCA, in contrast, uses the mean squared residue (MSR) to identify coherent patterns and is known for its efficiency and robustness in low-variance scenarios [9], [10].

Previous studies have highlighted strengths and weaknesses in various biclustering algorithms [11]–[13]. For instance, sparse models like FABIA may reveal biologically important but infrequent signals, while CCA excels in datasets where consistent co-expression dominates. Other reviews and comparative analyses further support the complementary nature of these methods in biomedical applications [10], [13], [14].

In this study, the performance of FABIA and CCA in analyzing the GSE19420 gene expression dataset was compared, focusing on their ability to detect biologically meaningful patterns under varying levels of simulated missingness. Missing values were addressed using k-nearest neighbors (KNN) imputation to preserve local gene relationships and minimize data distortion [15]–[17]. The mean squared error (MSE) was used as the primary evaluation metric to quantify reconstruction accuracy and algorithm robustness [18]. Previous studies have also demonstrated the application of biclustering to diabetic nephropathy and retinopathy microarray data, highlighting its effectiveness in diabetes-related gene expression analysis [14].

2. Method

2.1. Dataset

The dataset used in this study was the GSE19420, obtained from the Gene Expression Omnibus (GEO), a public repository maintained by the National Center for Biotechnology Information (NCBI) [1]. It contains gene expression profiles related to T2DM and is structured as a matrix in which each row represents a gene, and each column represents a sample. Key characteristics of the dataset include:

- A total of 54,675 genes, covering a broad range of expression signals.
- Forty-two samples (conditions), providing sufficient biological variability.

A variance analysis was conducted to assess the variability in gene expression. The observed variance ranged from 0.0029 to 0.2403, with a median value of approximately 0.1136. This suggests that while many genes show stable expression, some exhibit notable fluctuations. Although the original dataset contains no missing values, simulated missing values were introduced at levels ranging from 5% to 50%, increasing in 5% increments. These missing entries were imputed using the KNN method [12], [13], as explained in (4), prior to biclustering analysis. Given its high dimensionality and biological relevance, this dataset is well-suited for biclustering. The focus of this study is to compare the performance of FABIA and CCA in identifying gene–sample co-expression patterns under various levels of missing data.

2.2. Analysis Methodology

This study applied biclustering techniques to identify gene–sample subsets associated with insulin resistance in T2DM. Two algorithms were used: FABIA and the CCA. Their performance was evaluated based on MSE, which measures the difference between the original and reconstructed gene expression matrices [11]. Due to the large size of the dataset (54,675 genes × 42 samples), random sampling of 10% of the genes was used in certain runs to reduce computational complexity while maintaining biological representativeness. The full dataset was also analyzed where feasible to confirm result consistency.

The analysis began with data preparation. The GSE19420 dataset was downloaded from GEO, and a variance analysis was conducted to assess gene expression stability. Simulated missing values

p-ISSN 2798-253X
e-ISSN 2798-3153

were introduced at levels from 5% to 50%, followed by KNN imputation to restore the missing entries. Next, the FABIA algorithm was applied to identify gene–sample biclusters using a sparse factor model. The number of biclusters was determined based on the dataset characteristics. Following the biclustering, the gene expression matrix was reconstructed using the FABIA model, and the MSE was calculated. The same process was repeated using the CCA algorithm, identifying biclusters by minimizing the mean squared residue (MSR). The reconstructed matrix from CCA was also compared to the original using the MSE metric. Finally, the performance of FABIA and CCA was compared across all levels of simulated missing data. This comparison provided insights into the robustness and accuracy of each algorithm in reconstructing biologically meaningful co-expression patterns under incomplete data conditions.

2.3. Factor Analysis for Bicluster Acquisition (FABIA)

FABIA is a biclustering method that utilizes a sparse factor analysis framework to identify subsets of genes and conditions that exhibit coherent expression patterns [6], [7]. It models the gene expression matrix (X) as a linear combination of latent factors (Z), bicluster weights (W), and a noise matrix (E):

$$X = ZW + E \tag{1}$$

In the FABIA model, X denotes the gene expression matrix, which contains the observed expression values across genes and conditions. The latent factors, represented by Z, capture the underlying gene activity patterns that explain the observed data. The contribution of these latent factors to each gene or condition is quantified by the weight matrix W. Finally, E represents the residual error or noise, accounting for variations in the data that are not explained by the model.

FABIA applies sparsity constraints to retain only the most relevant genes and conditions, which makes it well-suited for high-dimensional gene expression data [9]. Its strength lies in detecting weak, rare, or overlapping signals that may be missed by traditional clustering. However, its performance can degrade in the presence of high noise or incomplete data [11].

2.4. Cheng and Church Algorithm (CCA)

The CCA identifies biclusters by minimizing the MSR, which quantifies coherence within a gene-condition submatrix [5]. For a bicluster with gene set (I) and condition set (J), the MSR is defined as in (2).

$$MSR(I,J) = \frac{1}{|I|\cdot|J|}\sum_{i\in I}\sum_{j\in J}\left(A_{ij} - \bar{A}_{iJ} - \bar{A}_{Ij} + \bar{A}_{IJ}\right)^2. \tag{2}$$

In the context of the CCA, A_{ij} represents the expression value of gene i under condition j. The term $\bar{A}_{ij}$ denotes the average expression of gene i across all conditions in cluster J, while $\bar{A}_{ij}$ refers to the average expression of all genes in set I under condition j. Lastly, $\bar{A}_{ij}$ indicates the overall average expression within the bicluster (I, J).

CCA iteratively removes the gene or condition that contributes the most to MSR, refining the bicluster until the residue falls below a user-defined threshold. Its simplicity and efficiency make it a popular choice for expression data analysis [9], [11]. However, CCA may fail to detect sparse or noisy signals that are biologically meaningful [18]. Comparative studies [8], [11] have shown that while CCA excels in finding coherent, low-variance clusters, FABIA is better at identifying biologically relevant but sparse gene sets. These complementary strengths justify comparing their performance under different missing data scenarios.

2.5. K-Nearest Neighbors (KNN) Imputation

KNN imputation is a widely used method for estimating missing values in gene expression data. It assumes that genes with similar expression profiles are likely to behave similarly across biological conditions. Thus, missing values can be inferred using values from genes with similar patterns [12], [13], [16]. Missing data is a frequent issue in microarray-based experiments and, if left unaddressed, can introduce bias into downstream analysis [15].

The distance between a gene with missing values and other genes with complete data is typically calculated using Euclidean distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \quad (3)$$

In the KNN imputation method, $d(x,y)$ represents the distance between genes x and y. The terms x_i and y_i denote the expression values of genes x and y, respectively, at position i. The variable n indicates the total number of conditions or samples used in the analysis.

After identifying the (k) most similar genes (neighbors), the missing value is imputed by averaging the corresponding values of these neighbors:

$$\hat{x}_{ij} = \frac{1}{k}\sum_{p=1}^{k} x_{pj}. \quad (4)$$

In the KNN imputation process, $\hat{x}_{ij}$ represents the imputed value for gene i under condition j. The term x_{pj} denotes the expression value of the pth nearest neighbor under the same condition, while k refers to the number of nearest neighbors used for the imputation. This method preserves local data structure while reducing the impact of noise and outliers. It has been shown to improve the quality of downstream analyses such as clustering and biclustering [16], [17].

2.6. Mean Squared Error (MSE)

MSE is used to evaluate the accuracy of matrix reconstruction after biclustering. A lower MSE indicates better agreement between the original and reconstructed gene expression data, and thus higher accuracy in identifying meaningful co-expression patterns [11]. The MSE is defined as in (5).

$$MSE = \frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - \widehat{X_{ij}}\right)^2 \quad (5)$$

where X_{ij} represents the original expression value for gene i under condition j, while $(\hat{X}_{ij})$ denotes the corresponding value in the reconstructed matrix. The variable m indicates the total number of genes (rows), and n refers to the total number of conditions or samples (columns).

MSE is particularly useful for comparing the performance of different biclustering algorithms and understanding how imputation affects reconstruction. In this study, MSE served as the primary metric to assess the impact of missing data on the performance of FABIA and CCA.

3. Results and Discussion

This section presents the results of the biclustering analysis on the GSE19420 gene expression dataset using two algorithms: FABIA and CCA. The analysis was conducted under varying levels of simulated missing data, ranging from 5% to 50%, in 5% increments. The performance of each method was evaluated based on the MSE, with lower values indicating better reconstruction accuracy.

Table 1 summarizes the MSE scores for both methods across all levels of missing data. The results showed that CCA consistently outperformed FABIA, achieving lower MSE values throughout the simulation. For example, at 5% missing data, the MSE of FABIA is 31.8051, while CCA records a significantly lower value of 22.3334. This trend remained consistent across all tested proportions.

Table 1. Comparison of MSE Score for FABIA and CCA Biclustering

No.	Percentage	Biclustering Method	MSE Score
1.	5%	FABIA	31.80509036
2.	5%	CCA	22.33344089
3.	10%	FABIA	31.67558645
4.	10%	CCA	22.33324523
5.	15%	FABIA	31.59508037
6.	15%	CCA	22.27471595
7.	20%	FABIA	34.34777162
8.	20%	CCA	22.24609422
9.	25%	FABIA	31.55570215
10.	25%	CCA	22.20938414

No.	Percentage	Biclustering Method	MSE Score
11.	30%	FABIA	31.55786547
12.	30%	CCA	22.13535348
13.	35%	FABIA	31.32743781
14.	35%	CCA	21.93532216
15.	40%	FABIA	31.78240904
16.	40%	CCA	21.927171
17.	45%	FABIA	31.30728998
18.	45%	CCA	21.77362267
19.	50%	FABIA	30.44305725
20.	50%	CCA	21.05085613

Both methods exhibited an increase in MSE values as the proportion of missing data increased. However, the increase was more pronounced in FABIA. These findings suggest that CCA is more robust to missing data, likely due to its use of the MSR criterion, which minimizes local variance and enhances stability under incomplete conditions.

Notably, both methods successfully identified genes such as MYO5B, DLG2, AXIN2, and PTK7. These genes are associated with insulin signaling and metabolic regulation, suggesting the biological relevance of the biclusters detected. Despite its higher reconstruction error, FABIA's ability to detect such genes highlights its strength in identifying low prevalence but significant expression patterns.

Overall, the findings emphasize the importance of selecting biclustering methods based on analytical objectives. CCA is preferable for general pattern recognition in noisy or incomplete datasets, whereas FABIA is more suitable for uncovering rare gene expression signatures. Further improvements, such as enhanced imputation strategies or hybrid models, may help optimize performance across both approaches.

4. Conclusion

This study compared the performance of FABIA and CCA biclustering algorithms on gene expression data with simulated missing values. MSE as the evaluation metric, CCA consistently achieved lower MSE scores across all missing data levels, indicating better reconstruction accuracy and robustness. Although FABIA was more sensitive to data incompleteness, it successfully identified biologically relevant, sparse expression patterns.

These results underscore the importance of aligning biclustering method selection with research goals and data conditions. CCA is more suitable for general-purpose analysis under incomplete data, while FABIA remains valuable for detecting rare gene expression signals. Future work may explore hybrid models or improved imputation strategies to enhance analysis of high-dimensional genomic data.

References

[1] American Diabetes Association, "Standards of medical care in diabetes—2022 abridged for primary care providers," *Clin. Diabetes*, vol. 40, no. 1, pp. 10–38, 2022, doi: 10.2337/cd22-as01.

[2] M.O. Goodarzi et al., "Classification of type 2 diabetes genetic variants and a novel genetic risk score association with insulin clearance," *J. Clin. Endocrinol. Metabolism*, vol. 105, no. 4, pp. 1251–1260, Apr. 2020, doi: 10.1210/clinem/dgz198.

[3] A. Mahmoud and A. Mohammed, "A survey on deep learning for time-series forecasting," in *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*, A.E. Hassanien and A. Darwish, Eds., Cham, Switzerland: Springer, 2021, pp. 365–392, doi: 10.1007/978-3-030-59338-4_19.

[4] S. Hochreiter et al., "FABIA: factor analysis for bicluster acquisition," *Bioinformatics*, vol. 26, no. 12, pp. 1520–1527, Apr. 2010, doi: 10.1093/bioinformatics/btq227.

[5] Y. Cheng and G.M. Church, "Biclustering of expression data," in *Proc. 8th Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 93–103.

[6] B. Pontes, R. Giráldez, and J.S. Aguilar-Ruiz, "Biclustering on expression data: A review," *J. Biomed. Inform.*, vol. 57, pp. 163–180, Oct. 2015, doi: 10.1016/j.jbi.2015.07.003.

[7] Breast Cancer Association Consortium, "Breast cancer risk genes—association analysis in more than 113,000 women," *New Engl. J. Med.*, vol. 384, no. 5, pp. 428–439, Feb. 2021, doi: 10.1056/NEJMoa1913948.

[8] M.I. Love, A.M. Bush, L.H. Chen, S.K. Patel, A.J. Cutler, and J.D. Cooper, "Large-scale genomic analyses reveal insights into pleiotropy across traits," *Nat. Commun*., vol. 13, Jun. 2022, Art. no. 3428, doi: 10.1038/s41467-022-30678-w.

[9] S.C. Madeira and A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 1, no. 1, pp. 24–45, Mar.–Jun. 2004, doi: 10.1109/TCBB.2004.2.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2019.

[11] A. Prelić 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.

[12] M.G. Rahman and M.Z. Islam, "Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques," Knowl.-Based Syst., vol. 53, pp. 51–65, 2013, doi: 10.1016/j.knosys.2013.08.023.

[13] T. Siswantining, A.E. Aminanto, D. Sarwinda, and O. Swasti, "Biclustering analysis using plaid model on gene expression data of colon cancer," *Austrian J. Stat.*, vol. 50, no. 5, pp. 101–114, Aug. 2021, doi: 10.17713/ajs.v50i5.1195

[14] T. Siswantining, D. Rahmawati, S. 'Uyun, and A.Z. Arifin, "Biclustering of diabetic nephropathy and diabetic retinopathy microarray data using a similarity-based biclustering algorithm," *Int. J. Bioinform. Res. Appl.*, vol. 17, no. 4, pp. 343–362, 2021, doi: 10.1504/IJBRA.2021.117934.

[15] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, Jun. 2001, doi: 10.1093/bioinformatics/17.6.520.

[16] I. Bitan-Roch, D. Levin, and D. Mahgereftekhari, "Imputation of missing $PM_{2.5}$ observations in a network of air quality monitoring stations by a new k-NN method," *Atmosphere*, vol. 13, no. 11, Nov. 2022, Art. no.1934, doi: 10.3390/atmos13111934.

[17] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, PA, USA: SIAM, 2007.

[18] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra, "Minimum sum-squared residue co-clustering of gene expression data," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 114–125, doi: 10.1137/1.9781611972740.11.