บทความวิจัย

การคัดเลือกตัวแปรแบบเบส์สำหรับตัวแบบการถดถอยเชิงเส้นที่มีมิติสูงโดยใช้กราฟแบบมีทิศทาง

บุษราคัม ประทานทรัพย์ และ วิฐูรา พึ่งพาพงศ์*
ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

* ผู้นิพนธ์ประสานงาน อีเมล: vitara@cbs.chula.ac.th
DOI: 10.14416/j.kmutnb.2024.10.020
รับเมื่อ 2 กรกฎาคม 2567 แก้ไขเมื่อ 6 กันยายน 2567 ตอบรับเมื่อ 3 ตุลาคม 2567 เผยแพร่ออนไลน์ 29 ตุลาคม 2567

บทคัดย่อ

ในการสร้างตัวแบบการถดถอยที่มีมิติสูง การคัดเลือกตัวแปรอย่างมีประสิทธิภาพเป็นสิ่งสำคัญในการเพิ่มความสามารถในการตีความและความแม่นยำของตัวแบบ บทความนี้นำเสนอวิธีการคัดเลือกตัวแปรสำหรับตัวแบบการถดถอยที่มีมิติสูงด้วยวิธี Iterated Conditional Modes/Medians Algorithm (ICM/M) ซึ่งนำกราฟแบบมีทิศทางเข้ามาใช้ประกอบการคัดเลือกตัวแปรแบบเบส์เพื่อจับความสัมพันธ์ที่มีทิศทางระหว่างตัวแปรต่าง ๆ โดยเรียกวิธีการใหม่ว่า ICM/M_D ในบทความนี้เปรียบเทียบประสิทธิภาพของวิธี ICM/M_D กับวิธีลาสโซ่ วิธี ICM/M แบบไม่พิจารณาความสัมพันธ์ระหว่างตัวแปร และวิธี ICM/M แบบพิจารณาความสัมพันธ์ระหว่างตัวแปรโดยใช้กราฟแบบไม่มีทิศทางผ่านข้อมูลจำลองต่าง ๆ ในบริบทของจีโนม ผลลัพธ์แสดงให้เห็นว่าวิธี ICM/M_D ให้อัตราการเกิดผลบวกเทียมที่ต่ำกว่าอย่างมีนัยสำคัญ ในขณะที่รักษาอัตราการเกิดผลลบเทียมในระดับที่สามารถแข่งขันกับวิธีอื่นได้ โดยเฉพาะในกรณีที่มีบางยีนในเครือข่ายมีความสัมพันธ์กับตัวแปรตามและตัวแปรอิสระมีเป็นจำนวนมาก ความสมดุลของความแม่นยำและความไวในการคัดเลือกตัวแปรนี้ทำให้ตัวแบบมีความน่าเชื่อถือและมีความสามารถในตีความได้ดีขึ้น วิธี ICM/M_D พิสูจน์ได้ว่าเป็นเครื่องมือที่แกร่งและมีคุณค่าสำหรับนักวิจัยซึ่งต้องจัดการกับชุดข้อมูลที่มีมิติสูงที่ซับซ้อน โดยเฉพาะอย่างยิ่งในสาขาพันธุศาสตร์และชีวสารสนเทศศาสตร์ ซึ่งจะได้ผลลัพธ์ที่ถูกต้องมากขึ้นภายใต้โครงสร้างทางชีวภาพหรือเครือข่ายที่ซับซ้อน

คำสำคัญ: ตัวแบบการถดถอยที่มีมิติสูง การคัดเลือกตัวแบบ วิธีแบบเบส์ กราฟแบบมีทิศทาง การวนซ้ำของฐานนิยม/มัธยฐานแบบมีเงื่อนไข

การอ้างอิงบทความ: บุษราคัม ประทานทรัพย์ และ วิฐูรา พึ่งพาพงศ์, "การคัดเลือกตัวแปรแบบเบส์สำหรับตัวแบบการถดถอยเชิงเส้นที่มีมิติสูงโดยใช้กราฟแบบมีทิศทาง," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 35, ฉบับที่ 4, หน้า 1–10, เลขที่บทความ 254-7618, ต.ค.–ธ.ค. 2568.

วารสารวิชาการพระจอมเกล้าพระนครเหนือ ปีที่ 35 ฉบับที่ 4 ต.ค.–ธ.ค. 2568
The Journal of KMUTNB., Vol. 35, No. 4, Oct.–Dec. 2025

Research Article

Incorporating a Directed Graph in Bayesian Variable Selection for a High-dimensional Regression Model

Busarakam Pratansup and Vitara Pungpapong*
Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, Bangkok, Thailand

* Corresponding Author, E–mail: vitara@cbs.chula.ac.th DOI: 10.14416/j.kmutnb.2024.10.020
Received 2 July 2024; Revised 6 September 2024; Accepted 3 October 2024; Published online: 29 October 2024

Abstract

In high-dimensional regression models, effective variable selection is critical for enhancing model interpretability and accuracy. This paper introduces a novel method, ICM/M_D, which incorporates directed graphs into the Bayesian variable selection framework to capture directional relationships among variables. We compare the performance of ICM/M_D with Lasso, ICM/M without considering a network, and ICM/M with undirected graph incorporation methods across various simulation scenarios in a genomic context. The results demonstrate that ICM/M_D achieves significantly lower false positive rates while maintaining competitive false negative rates, especially in cases where not all genes in the network are related to the response and the number of predictors is large. This balance of precision and recall ensures more reliable and interpretable models. The ICM/M_D method proves to be a robust and valuable tool for researchers dealing with complex high-dimensional datasets, particularly in genomics and bioinformatics, by providing a more accurate representation of underlying biological or network structures.

Keywords: High-dimensional Regression Model, Variable Selection, Bayesian Method, Directed Graph, Iterative Conditional Modes/Median

Please cite this article as: B. Pratansup and V. Pungpapong, "Incorporating a Directed Graph in Bayesian Variable Selection for a High-dimensional Regression Model," *The Journal of KMUTNB*, vol. 35, no. 4, pp. 1–10, ID. 254-7618, Oct.–Dec. 2025 .

1. Introduction

High-dimensional datasets refer to datasets where the number of variables can far exceed the number of observations. This scenario is common in biomedical science research, where researchers often measure a vast array of variables, such as genetic markers or protein levels, relative to a smaller number of samples, like patients or tissue samples. Variable selection is crucial in high-dimensional data analysis, especially in biomedical research, as it enhances model interpretability by identifying the most relevant predictors. By focusing on the most informative variables, researchers can gain deeper insights into the underlying mechanisms of diseases and improve patient outcomes [1].

When the outcome is a quantitative variable, a regression model is typically employed to find associations between predictors and outcomes. However, the standard regression model with the Ordinary Least Squares (OLS) method suffers from several limitations, including potential multicollinearity among predictors, and reduced effectiveness in model estimation and interpretation in high-dimensional settings [2].

Biological pathways, available in many databases, are series of interactions among molecules within a cell that lead to specific products or changes, playing critical roles in processes like metabolism and signal transduction. Graphs, where nodes represent genes and edges represent interactions, are commonly used to represent interconnections among genes in biological pathways. Gene networks help in understanding how genes work together to regulate various cellular processes. Incorporating these pathways into biomedical research models can help in understanding complex biological mechanisms and provide a more holistic view of biological systems.

Pungpapong *et al.* [3] proposed a Bayesian framework that incorporates biological pathways, represented by an undirected graph, into variable selection in high-dimensional settings. This could be done by using an Ising prior. For fast and easy implementation, the Iterated Conditional Modes/ Median (ICM/M) algorithm was introduced [3]. However, since the interactions among genes in the network are directional, with one gene product regulating or influencing another, a directed graph should be used instead of an undirected graph.

In this paper, we aim to modify the Ising prior to capture directional relationships among genes in the network. The performance of our proposed method is assessed based on its effectiveness in variable selection.

2. Methods

In this section, we first review an empirical Bayes variable selection method incorporating undirected graphs, as proposed by Pungpapong *et al.* [3]. Then, we demonstrate how to modify this method to accommodate directed graphs.

2.1 Incorporating Undirected Graphs for Variable Selection in High-Dimensional Regression

2.1.1 Bayesian Framework

Consider a high-dimensional regression model as shown in Equation (1).

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is a vector of size n of response variable, X is a $n \times p$ matrix of covariates, β is a p-dimensional vector representing the regression coefficient, and ε is a n-dimensional vector of error term and assume that $\varepsilon \sim N(0,\sigma^2 I_n)$, where I_n is an $n \times n$ identity metrix. In other words, errors are independent and identically distributed with a constant variance of σ^2.

By further assuming that the response is centered and the covariates are standardized, it can be shown that, for $j = 1, \ldots, p$, the sufficient statistic for β_j given known value of all other coefficients $(\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ is shown in Equation (2).

$$\frac{1}{n-1} X_j^t \tilde{Y}_j \sim N\left(\beta_j, \frac{1}{n-1}\sigma^2\right) \tag{2}$$

where X_j is the j-th column of X and $\tilde{Y}_j = Y - X\beta + X_j\beta_j$.

To perform variable selection as well as integrating undirected graph among predictors, the mixture prior in Equation (3) is put on each of β_j independently and the Ising prior [4] in Equation (4) is employed to capture relationships among predictors.

$$\beta_j \mid \tau_j \sim (1-\tau_j)\delta_0(\beta_j) + \tau_j\gamma(\beta_j) \tag{3}$$

$$P(\tau) \propto \exp\left\{a\sum_j \tau_j + b\sum_{\langle j,k\rangle \in E} \tau_j\tau_k\right\} \tag{4}$$

where $\delta_0(\cdot)$ is a dirac delta function (i.e., mass at 0), $\gamma(\cdot)$ is a Laplace density with scale parameter α, $\tau = (\tau_1, \ldots, \tau_p)$ where τ_j is an indicator variable defined as $\tau_j = 1\{\beta_j \neq 0\}$. From (3), when $\tau_j = 0$, β_j is equal to 0 and when $\tau_j = 1$, the prior of β_j follows the Laplace distribution. The Laplace density with scale parameter α is given by Equation (5).

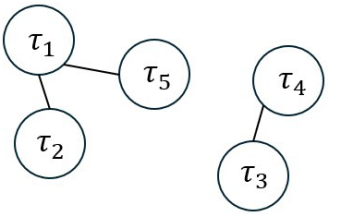

Figure 1 Example of an undirected graph.

$$\gamma(\beta_j) = \frac{\alpha\sqrt{n-1}}{2\sigma}\exp\left\{-\frac{\alpha\sqrt{n-1}}{\sigma}|\beta_j|\right\} \tag{5}$$

The Ising prior in Equation (4) has two hyperparameters a and b where a influences the overall tendency of nodes to be in particular state (i.e., $\tau_j = 1$ or $\tau_j = 0$) and b represents the interaction strength when nodes in undirected graphs are connected. The latter summation in Equation (4) indicates that we sum all over all edges in undirected graphs. The edge set E contains all edges in undirected graphs. For example, the edge set is $E = \{\langle 1,2\rangle, \langle 2,1\rangle, \langle 1,5\rangle, \langle 5,1\rangle, \langle 3,4\rangle, \langle 4,3\rangle\}$ in Figure 1. For σ, the Jeffreys' prior [5] is used, meaning it does not favor any particular values of σ. The Jeffreys' prior takes the form $P(\sigma) \propto 1/\sigma$.

2.1.2 The iterative Conditional Modes/Median Algorithm

Pungpapong *et al.* [3] introduced the iterated conditional modes/medians (ICM/M) algorithm for efficient empirical Bayes variable selection. This method determines optimal hyperparameters and auxiliary parameters as the modes of their full conditional distributions, while each regression coefficient is derived as the median of its full conditional distribution. The Bayesian framework allows simultaneous variable selection and estimation using conditional medians.

With initial estimates of all parameters denoted as $\hat{\beta}$ and $\hat{\sigma}^2$, the ICM/M algorithm can be summarized as follows.

1) Compute indicator variable $\hat{\tau} = (\hat{\tau}_1,\ldots,\hat{\tau}_p)$ where $\hat{\tau}_j = 1\{\hat{\beta}_j \neq 0\}$.

2) Obtain the estimate of hyperparameters $(\hat{a},\hat{b})$ as the mode of its pseudo-likelihood function as shown in Equation (6).

$$(\hat{a},\hat{b}) = mode\left\{\prod_{j=1}^{p} P(\hat{\tau}_j \mid \hat{\tau}_k : \langle j,k\rangle \in E)\right\} \quad (6)$$

3) For $j = 1, \ldots, p$, update $\hat{\beta}_j$ as the posterior median as shown in Equation (7).

$$\hat{\beta}_j = median\{P(\beta_j \mid Y, X, \hat{\tau}, \hat{a}, \hat{b}, \hat{\beta}_{-j})\} \quad (7)$$

where $\hat{\beta}_{-j}$ is $(p-1)$-dimensional vector of current estimates of the regression coefficients except β_j.

4) Update $\hat{\sigma}^2$ as the mode of its full conditional function as shown in Equation (8).

$$\hat{\sigma}^2 = mode\{P(\sigma^2 \mid Y, X, \hat{\tau}, \hat{a}, \hat{b}, \hat{\beta})\} \quad (8)$$

5) Iterate steps 1–4 until convergence in $\hat{\beta}$ That is, in the k-th iteration, the algorithm stops when

$$\frac{\sum_{j=1}^{p}\left(\widehat{\beta}_j^{(k)} - \widehat{\beta}_j^{(k-1)}\right)^2}{\sum_{j=1}^{p}\left[\widehat{\beta}_j^{(k)}\right]^2} < 10^{-6}$$

2.2 Incorporating Directed Graphs for Variable Selection in High-Dimensional Regression

We can see that the Ising prior in Equation (4) can be used to capture the interactions among nodes in undirected graphs. In undirected graphs, the edge set is symmetrical, meaning if there is an edge between nodes i and j, it is represented equally from i to j and from j to i. In contrast, in directed graphs, the edge set is not symmetrical because the direction of the edges matters. The edge set of a directed graph is a collection of ordered pairs of vertices, where each pair represents a directed edge. Consider an example in Figure 2, the edge set is $E = \{\langle 1,2\rangle,\langle 1,5\rangle,\langle 4,3\rangle\}$. We further define pa_j to be the set of parents of node j. For example, in Figure 2, $pa_5 = \{1\}$ since node 1 is the parent of node 5 (i.e, there exists a directed edge 1→5) and pa_1 is an empty set.

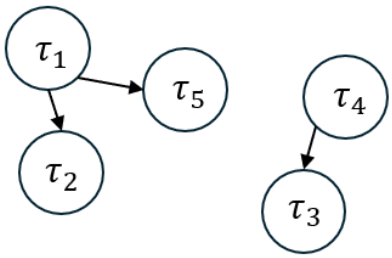

Figure 2 Example of a directed graph.

In order to incorporate directed graphs into the model, we propose modifying the Ising prior in Equation (4) as shown in Equation (9).

$$P(\tau) \propto \exp\left\{a\sum_j \tau_j + b\sum_{k:k\in pa_j} \tau_j\tau_k\right\} \quad (9)$$

The ICM/M algorithm can still be employed for implementation. The only modification is in Step 2, where the hyperparameters $(\hat{a},\hat{b})$ are estimated using the mode of their pseudo-likelihood function, as described in Equation (10).

$$(\hat{a},\hat{b}) = mode\left\{\prod_{j=1}^{p} P(\hat{\tau}_j \mid \hat{\tau}_k : k\in pa_j)\right\} \quad (10)$$

2.3 Model Evaluation Metrics

To assess the effectiveness of our proposed method for variable selection, we employ the following evaluation metrics.

2.3.1 Precision

Precision is the proportion of selected variables that are truly relevant defined as in Equation (11).

$$Precision = \frac{\sum_{j=1}^{p} 1\left\{\beta_j \neq 0, \widehat{\beta}_j \neq 0\right\}}{\sum_{j=1}^{p} 1\left\{\widehat{\beta}_j \neq 0\right\}} \quad (11)$$

2.3.2 Recall (or Sensitivity)

Recall is the proportion of truly relevant variables that are correctly selected defined in Equation (12).

$$Recall = \frac{\sum_{j=1}^{p} 1\left\{\beta_j \neq 0, \widehat{\beta}_j \neq 0\right\}}{\sum_{j=1}^{p} 1\left\{\beta_j \neq 0\right\}} \quad (12)$$

2.3.3 Specificity

Specificity is the proportion of truly irrelevant variables that are correctly not selected as shown in Equation (13).

$$Specificity = \frac{\sum_{j=1}^{p} 1\left\{\beta_j = 0, \widehat{\beta}_j = 0\right\}}{\sum_{j=1}^{p} 1\left\{\beta_j = 0\right\}} \quad (13)$$

2.3.4 False Positive Rate

False Positive Rate (FPR), defined in Equation (14), is the proportion of truly irrelevant variables that are incorrectly selected.

$$FPR = \frac{\sum_{j=1}^{p} 1\left\{\beta_j = 0, \widehat{\beta}_j \neq 0\right\}}{\sum_{j=1}^{p} 1\left\{\beta_j = 0\right\}} \quad (14)$$

2.3.5 False Negative Rate

False Negative Rate (FNR) is the proportion of truly relevant variables that are incorrectly not selected as shown in Equation (15).

$$FNR = \frac{\sum_{j=1}^{p} 1\left\{\beta_j \neq 0, \widehat{\beta}_j = 0\right\}}{\sum_{j=1}^{p} 1\left\{\beta_j \neq 0\right\}} \quad (15)$$

3. Results

Simulation studies were conducted to mimic real microarray data using gene expression data from Schmidt *et al.* [6] as covariates X. This dataset includes 22,283 genes from 200 samples. Genes were sampled in scenarios with $p = 2{,}000$ and $p = 5{,}000$ genes while keeping the sample size at $n = 200$. From the list of selected genes, directed gene pathways were extracted from the BioGrid database [7]. True coefficients β were generated with most being zero, except for nonzero coefficients in 10 networks. The scenarios included all genes in 10 networks with nonzero coefficients and 50% of genes in each of 10 networks with nonzero coefficients. For nonzero coefficients, β were drawn from a uniform distribution from 0.5 and 2 ($U(0.5,2)$). This choice of nonzero coefficients reflects both relatively small and larger effect sizes while evenly balancing the proportion between small and large effect sizes. The response variable is generated under the model in Equation (1) with two levels of Signal-to-noise Ratio (SNR), 1 and 3, corresponding to cases where the noise is strong and moderate relative to the signal, respectively. Table 1 summarizesthe 8 simulation cases. All simulations were run 100 times for each scenario using the R programming environment [8].

Table 1 Summary of simulation scenarios.

Case	p	Nonzero β in 10 Networks	SNR
1	2,000	All	1
2	2,000	All	3
3	2,000	Some	1
4	2,000	Some	3
5	5,000	All	1
6	5,000	All	3
7	5,000	Some	1
8	5,000	Some	3

To evaluate the performance of our proposed method, we compared it with three existing methods. Therefore, we ran 4 algorithms as follows:

1) Lasso [9]

2) ICM/M without prior network information (ICM/M)

3) ICM/M incorporating undirected graphs (ICM/M_U)

4) ICM/M incorporating directed graphs (ICM/M_D) - our proposed method.

All ICM/M algorithms required initial regression coefficient values, for which we used Lasso estimates with ten-fold cross-validation to select regularized parameter computed via the *glmnet* R package [10], [11]. For the ICM/M algorithm, we utilized the *icmm* R package [12], using a fixed scaled parameter value $\alpha = 0.5$ in the Laplace density as recommended by Johnstone and Silverman [13].

Table 2 shows the average precision, recall, and specificity among 100 simulations in 8 scenarios. All ICM/M algorithms outperformed Lasso in terms of precision and specificity. However, Lasso performed the best in terms of recall.

Table 2 Average precision, recall, and specificity among 100 simulations in 8 scenarios. Values in bold represent the best method.

Precision				
Case	Lasso	ICM/M	ICM/M_U	ICM/M_D
1	0.1346	0.5263	**0.5554**	0.5289
2	0.1611	0.5476	**0.6245**	0.5803
3	0.1413	**0.7086**	0.6871	0.7005
4	0.1492	0.7355	0.7498	**0.7535**
5	0.0932	0.1036	**0.3475**	0.2898
6	0.1120	0.1144	**0.4786**	0.3161
7	0.1021	0.0417	0.0520	**0.1307**
8	0.1013	0.0745	0.0691	**0.1375**
Recall				
Case	Lasso	ICM/M	ICM/M_U	ICM/M_D
1	**0.1669**	0.0546	0.0656	0.0544
2	**0.2905**	0.1126	0.1587	0.1274
3	**0.2771**	0.1233	0.1357	0.1248
4	**0.4110**	0.2495	0.2857	0.2557
5	**0.0831**	0.0053	0.0323	0.0224
6	**0.1924**	0.0153	0.0669	0.0313
7	**0.1645**	0.0061	0.0100	0.0224
8	**0.2527**	0.0200	0.0145	0.0264
Specificity				
Case	Lasso	ICM/M	ICM/M_U	ICM/M_D
1	0.9754	**0.9989**	**0.9989**	**0.9989**
2	0.9670	0.9980	**0.9981**	**0.9981**
3	0.9783	**0.9993**	0.9992	**0.9993**
4	0.9718	0.9990	0.9989	**0.9991**
5	0.9891	0.9993	**0.9994**	0.9993
6	0.9797	0.9983	**0.9993**	0.9992
7	0.9886	**0.9992**	0.9991	0.9991
8	0.9839	0.9984	**0.9990**	**0.9990**

In model variable selection, precision reflects the proportion of true positives among selected variables. Lasso generally showed lower precision, indicating a higher chance of selecting irrelevant

variables. ICM/M improved precision significantly, while ICM/M_U consistently achieved the highest precision, effectively selecting relevant variables. ICM/M_D also performed well, with precision comparable to ICM/M_U. When not all genes in the network were related to the response, thus mimicking a more realistic scenario than assuming all genes in the network were relevant, it became evident that ICM/M_D achieved significantly higher precision than ICM/M_U, particularly when dealing with a large number of predictors (case 7 and 8). Hence, ICM/M_U and ICM/M_D are superior in precise variable selection, making them preferable for applications requiring accurate identification of relevant variables such as in biomedical field.

In terms of recall, Lasso was a clear winner, indicating it was most effective at capturing the majority of relevant variables and demonstrating superior performance in identifying true positive variables compared to the other methods. However, this came with a trade-off with precision; Lasso tended to select too many variables, resulting in a higher number of false positives.

Specificity measures the proportion of irrelevant variables correctly identified as irrelevant by the model. Both ICM/M_U and ICM/M_D consistently achieved the highest specificity, effectively minimizing the selection of irrelevant variables.

The results in Table 3, showing the average false positive rate (FPR) and false negative rate (FNR) among 100 simulations in 8 scenarios, were consistent with those from Table 2. Specifically, it was evident that Lasso consistently exhibited the highest FPR, indicating it often selected irrelevant variables. In contrast, ICM/M, ICM/M_U, and ICM/M_D methods showed significantly lower FPRs, indicating a strong ability to exclude irrelevant variables. Among these, ICM/M_U and ICM/M_D demonstrated competitive FNRs, balancing the precision of selecting relevant variables with minimal false positives. ICM/M showed a slightly higher FNR compared to ICM/M_U and ICM/M_D, indicating it missed more relevant variables. Overall, ICM/M_U and ICM/M_D outperformed Lasso by minimizing FPR and maintaining competitive FNR, making them more reliable for precise and comprehensive variable selection.

Table 3 Average false positive rate and false negative rate among 100 simulations in 8 scenarios. Values in bold represent the best method.

FPR				
Case	Lasso	ICM/M	ICM/M_U	ICM/M_D
1	0.0246	**0.0011**	**0.0011**	**0.0011**
2	0.0330	0.0020	**0.0019**	**0.0019**
3	0.0217	**0.0007**	0.0008	**0.0007**
4	0.0282	0.0010	0.0011	**0.0009**
5	0.0109	0.0007	**0.0006**	0.0007
6	0.0203	0.0017	**0.0007**	0.0008
7	0.0114	**0.0008**	0.0009	0.0009
8	0.0161	0.0016	**0.0010**	**0.0010**
FNR				
Case	Lasso	ICM/M	ICM/M_U	ICM/M_D
1	**0.8331**	0.9454	0.9344	0.9456
2	**0.7095**	0.8874	0.8413	0.8726
3	**0.7299**	0.8767	0.8643	0.8752
4	**0.5890**	0.7505	0.7143	0.7443
5	**0.9169**	0.9947	0.9677	0.9776
6	**0.8076**	0.9847	0.9331	0.9687
7	**0.8355**	0.9939	0.9900	0.9776
8	**0.7473**	0.9800	0.9855	0.9736

4. Discussion and Conclusion

In this paper, we demonstrate the effectiveness of the proposed ICM/M_D method which incorporates directed graph for variable selection in high-dimensional regression models. Compared to the traditional Lasso method, ICM/M_D shows significantly lower FPR, indicating higher precision in excluding irrelevant variables. This improvement is crucial for high-dimensional datasets, where the inclusion of irrelevant variables can obscure meaningful results and reduce model interpretability. Furthermore, while Lasso has lower FNR, indicating its effectiveness in capturing relevant variables, it does so at the cost of much higher false positives.

ICM/M and ICM/M_U methods also show strong performance, particularly in reducing FPR, but ICM/M_D consistently balances both FPR and FNR more effectively, especially in cases where not all genes in the network are related to the response and the number of predictors is large. This balance is essential for ensuring both precision and recall in variable selection, leading to more reliable and interpretable models. Our results indicate that ICM/M_D can effectively incorporate directional relationships among variables, providing a more accurate representation of underlying biological or network structures.

In conclusion, the ICM/M_D method offers a robust approach to variable selection in high-dimensional settings, outperforming traditional methods by minimizing false positives while maintaining competitive false negative rates. Although the overall performance of ICM/M_U and ICM/M_D is comparable, we observe some scenarios where ICM/M_D is superior to ICM/M_U. This makes ICM/M_D a valuable tool for researchers working with complex genomics and bioinformatics datasets, where understanding the intricate relationships between variables is crucial. Future work could explore the application of ICM/M_D in other domains and further refine the method to enhance its performance and applicability.

References

[1] K. Tadist, S. Najah, N. S. Nikolov, F. Mrabti, and A. Zahi, "Feature selection methods and genomic big data: a systematic review," *Journal of Big Data*, vol. 6, no. 79, 2019.

[2] V. Pungpapong, "A brief review on high-dimensional linear regression," *Thai Science and Technology Journal*, vol. 23, no. 2, 2015. (in Thai)

[3] V. Pungpapong, M. Zhang, and D. Zhang, "Selecting massive variables using an iterated conditional modes/medians algorithm," *Electronic Journal of Statistics*, vol. 9, no. 1, pp. 1243–1266, 2015.

[4] L. Onsager, "Crystal statistics. I. A two-dimensional model with an order-disorder transition," *Physical Review*, vol. 65, pp. 117–149, 1943.

[5] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 196, pp. 453–461, 1946.

[6] M. Schmidt, D. Bohm, C. von Torne, E. Steiner, A. Puhl, H. Pilch, H.-A. Lehr, J. G. Hengstler, H. Kolbl, and M. Gehrmann, "The humoral immune system has a key prognostic impact in node-negative breast cancer," *Cancer*

Research, vol. 68, no. 13, pp. 5405–5413, 2008.

[7] R. Oughtred, J. Rust, C. Chang, B. J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, and F. Zhang, "The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions," *Protein Science*, vol. 30, no. 1, pp. 187–200, 2021.

[8] R Core Team.*R Foundation for Statistical Computing*.(2023),R:A language and environment for statistical computing. [Online]. Available: https://www.R-project.org

[9] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[10] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[11] J. H. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, and J. Qian, *glmnet: Lasso and elastic-net regularized generalized linear models*. (2021). R package version 4.1-1.

[12] V. Pungpapong, M. Zhang, and D. Zhang. *icmm: Empirical Bayes variable selection via ICM/M algorithm*. (2021). R package version 1.2.

[13] I. M. Johnstone and B. W. Silverman, "Empirical Bayes selection of wavelet thresholds," *Annals of Statistics*, vol. 33, no. 4, pp. 1700–1752, Aug. 2005.