Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 7471516, 9 pages
https://doi.org/10.1155/2021/7471516

Research Article

Bayesian Gene Selection Based on Pathway Information and Network-Constrained Regularization

Ming Cao,[1,2] **Yue Fan**,[1] **and Qinke Peng**[1]

[1]*Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China*
[2]*School of Mathematics and Statistics, Shaanxi Xueqian Normal University, Xi'an 710100, China*

Correspondence should be addressed to Qinke Peng; qkpeng@xjtu.edu.cn

Received 25 April 2021; Revised 5 July 2021; Accepted 23 July 2021; Published 5 August 2021

Academic Editor: Maria N. D.S. Cordeiro

High-throughput data make it possible to study expression levels of thousands of genes simultaneously under a particular condition. However, only few of the genes are discriminatively expressed. How to identify these biomarkers precisely is significant for disease diagnosis, prognosis, and therapy. Many studies utilized pathway information to identify the biomarkers. However, most of these studies only incorporate the group information while the pathway structural information is ignored. In this paper, we proposed a Bayesian gene selection with a network-constrained regularization method, which can incorporate the pathway structural information as priors to perform gene selection. All the priors are conjugated; thus, the parameters can be estimated effectively through Gibbs sampling. We present the application of our method on 6 microarray datasets, comparing with Bayesian Lasso, Bayesian Elastic Net, and Bayesian Fused Lasso. The results show that our method performs better than other Bayesian methods and pathway structural information can improve the result.

1. Introduction

Identifying disease-associated genes, which can be treated as diagnostic biomarkers, can bring a significant effect on disease diagnosis, prognosis, and treatments [1, 2]. With the development of high-throughput technologies in recent years, gene expression profiling has provided a useful way to find biomarkers. Researchers can identify the genes which are differentially expressed between two groups of samples. These genes are regarded as disease-associated genes. However, gene expression data usually contains a large number of genes and a relatively small sample size [3, 4]. And many of the genes are also redundant or irrelevant to the prediction [5, 6]. Furthermore, there are also noises in the experiment procedures which will influence the gene expression values. Therefore, identifying the biomarkers from gene expression data is challenging.

During the last decades, a number of gene selection methods have been developed to tackle this problem. Feature selection and feature extraction are two major methods (we treat gene and feature equally in this paper). On the one hand, the aim of feature selection is to select relevant features and do not change the form of the features. On the other hand, feature extraction will extract the feature from the original data and may alter the form of the features. Here, we focus on the feature selection methods since the results of such methods could be interpreted easily. Feature selection methods can be generally organized into three categories: filter, wrapper, and embedded methods. Both the wrapper and embedded methods are classifier-dependent methods; thus, they are always time consuming and easy to overfitting. However, the filter methods are usually based on statistic approaches [7] such as mRMR [5], PLSRFE [8], lasso [9], and elastic net [10], which are relatively efficient in terms of computation and can derive a score of each of the genes which represents the significance of the gene. Therefore, we focus on the filter methods in this paper.

Although these methods are successful in many applications, they usually obtain suboptimal solutions. Therefore, the prediction accuracies are not satisfied and the

disease-associated genes selected from different methods have few overlaps [11]. This is partly due to the fact that the discriminatory power of many biomarkers is similar. Furthermore, some genes which have low discriminatory powers play important roles in cellular functions. Their combinations are highly discriminative while they are usually ignored [12, 13].

Recently, with a large amount of biological information accumulated, there is an increased interest in gene selection with incorporating information on pathways, which can partially compensate for the lack of reliable expression data [14]. Pathways depict a series of chemical interactions in living cells; genes that interact with one another usually mean that they function together concertedly. Therefore, these genes should be highly correlated and have dependence structures. However, many studies only utilize the information that pathways cluster genes into the natural group; the pathway structural information is neglected. Li and Li have overcome this disadvantage by incorporating pathway structure information through a Laplacian matrix of a global graph [15, 16] and combined with lasso penalty to perform network-constrained penalty which can select subgroups of correlated features in the network. This penalty is based on the assumption that genes belonging to the same pathway have similar functions and therefore smoothed regression coefficients. And this penalty has been successfully applied in many studies [17–19].

The Bayesian approach has three major advantages over Bayesian selection methods [20]. Firstly, hyperparameters can be estimated automatically through fulfilling stochastic draws; thus, 10-fold cross-validation for estimating penalized parameters is not required. Secondly, the Bayesian framework can utilize the pathway information naturally by integrating it in the model as prior knowledge. Finally, the Bayesian estimation with the posterior distributions can provide credible intervals for the regression coefficients, which is a great advantage over frequentist methods.

In this paper, we work with a Bayesian framework to perform gene selection through network-constrained regularization. Similar to the Bayesian Lasso [21], Bayesian Elastic Net [22], and Bayesian Fused Lasso [23], we use shrinkage priors to perform regularization. We show that all the conditional posteriors of the proposed model are available in closed form and proper. Thus, parameter estimation can be performed through Gibbs sampling easily. The pathway information is obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [24], which is the most popular pathway public database, especially pathways associated with several types of cancer could be obtained in the model. Furthermore, following Held and Holmes [25], we extend the regression model to binary regression which can perform binary classification through an auxiliary variable. This method is assessed by applying it to several microarray datasets.

2. Method

2.1. The Bayesian Network-Constrained Model for Gene Selection. Considering an $N \times P$ matrix X, where P is the number of genes and N is the number of the samples, with a response vector $y = (y_1, \cdots, y_n)^T$, we normalize the values of each feature as the tradition in variable selection; thus, the mean and standard deviation of each feature are 0 and 1. We assume the likelihood function of the continuous response is Gaussian function:

$$Y|X, \beta, \sigma^2 \sim N_n\left(X\beta, \sigma^2 I_n\right), \tag{1}$$

which can be also expressed as

$$y = X\beta + \varepsilon, \varepsilon \sim N_n\left(0, \sigma^2 I_n\right). \tag{2}$$

Following Li and Li's work [16], we incorporate the pathway information through its normalized Laplacian matrix. Consider an undirected graph $G = (V, E, W)$. In this graph, genes are represented by a set of nodes V, and the interactions between genes are represented by a set of edges $E = \{u \sim v\}$, and W is the weights of the edges, where $w(u, v)$ represents the weight of edge $e = (u \sim v)$ which indicates the uncertainty of the edge between the vertices u and v. The degree of each vertex is defined as $d_v = \sum_{u \sim v} w(u, v)$. Then, the normalized Laplacian matrix L for graph G with the uth and vth elements can be defined by

$$L(u, v) = \begin{cases} 1 - \dfrac{w(u, v)}{d_u}, & \text{if } u = v \text{ and } d_u \neq 0, \\ -\dfrac{w(u, v)}{\sqrt{d_u d_v}}, & \text{if } u \text{ and } v \text{ are adjacent}, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here, we let $w(u, v) = 1$ if there exists an interaction between gene u and v, and $w(u, v) = 0$, otherwise.

To form the network-constrained regularization, we assign the prior distribution for β as follows:

$$\beta \sim N_p\left(0, \frac{\sigma^2}{r} \Lambda^{-1}\right), \tag{4}$$

where Λ is taking the form:

$$\begin{aligned} \Lambda &= \operatorname{diag}\left(\tau_1^{-1}, \tau_2^{-1}, \cdots, \tau_p^{-1}\right) + L \\ &= \begin{Bmatrix} 1 + \tau_1^{-1} & L(1,2) & \cdots & L(1,p) \\ L(2,1) & 1 + \tau_2^{-1} & \cdots & L(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ L(p,1) & L(p,2) & \cdots & 1 + \tau_p^{-1} \end{Bmatrix}. \end{aligned} \tag{5}$$

Note that Λ only contains hyperparameter τ.

To eliminate the $|\Lambda|^{1/2}$ in the prior distribution of β, we assign the prior distribution for τ as follows:

$$p(\tau^2 \mid \lambda) = C_\tau |\Lambda|^{-1/2} \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau_j^2\right), \tag{6}$$

where C_τ is the normalizing constant.

The prior distribution defined in (6) is proper, due to the following analysis:

Let $A = \Lambda - I_n$, and A is a symmetric and positive semidefinite matrix.

Let $D_A = \text{diag}(a_1, \cdots, a_p)$, where $a_1, \cdots, a_p$ are eigenvalues of A and $0 \le a_1 \le \cdots \le a_p$.

Since A is the symmetric and positive semidefinite, there exists an orthonormal matrix Q. Hence, the eigendecomposition of matrix A can be written as $A = QD_AQ^T$.

Because of $\Lambda = A + I_n = QD_AQ^T + QQ^T = Q(D_A + I_n)Q^T$, so $|\Lambda| = \prod_{i=1}^{n}(a_i + 1) \ge 1$.

Then,

$$\begin{aligned}\int_0^\infty C_\tau |\Lambda|^{-1/2} \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau_j^2\right) d\tau^2 \\ \le C_\tau \int_0^\infty \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau_j^2\right) d\tau^2 < \infty,\end{aligned} \tag{7}$$

where the integrand is kernels of the gamma density that indicates the integral is finite. Therefore, the prior distribution is proper.

Since

$$\begin{aligned}\beta^T\Lambda\beta &= \beta^T D^{-1}\beta + \sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}}\right)^2 \\ &\ge 0, \quad D = \text{diag}\left(\tau_1^2, \tau_2^2, \cdots, \tau_p^2\right),\end{aligned} \tag{8}$$

Λ is positive semidefinite.

The joint posterior distribution can be written as

$$\begin{aligned}p(\beta, \lambda, \sigma^2, \tau^2, r \mid X, Y) &\propto \left(\sigma^2\right)^{-n/2} \exp \\ &\cdot \left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right)\left(\sigma^2\right)^{-p/2} r^{p/2}|\Lambda|^{1/2} \exp \\ &\cdot \left(-\frac{r\beta D^{-1}\beta + r\beta^T L\beta}{2\sigma^2}\right)|\Lambda|^{-1/2}\frac{\lambda^2}{2} \exp \\ &\cdot \left(-\frac{\lambda^2}{2}\tau^2\right)p(r)p(\sigma^2)p(\lambda).\end{aligned} \tag{9}$$

Integrating out τ^2, we have

$$\begin{aligned}p(\beta, \lambda, \sigma^2, r \mid X, Y) &= \int p(\beta, \lambda, \sigma^2, \tau^2, r \mid X, Y)p(\tau^2)d\tau^2 \\ &\propto \int_0^\infty \left(\sigma^2\right)^{-n/3} \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right) \\ &\cdot \left(\sigma^2\right)^{-p/2} r^{p/2}|\Lambda|^{1/2} \exp \\ &\cdot \left(-\frac{r\beta D^{-1}\beta + r\beta^T L\beta}{2\sigma^2}\right)|\Lambda|^{-2/2}\frac{\lambda^2}{2} \exp \\ &\cdot \left(-\frac{\lambda^2}{2}\tau^2\right)p(r)p(\sigma^2)p(\lambda)d\tau^2 \propto \int_0^\infty \exp \\ &\cdot \left(-\frac{\|Y - X\beta\|^2 + r\beta^T L\beta}{2\sigma^2}\right) \exp \\ &\cdot \left(-\frac{r\beta D^{-1}\beta}{2\sigma^2}\right)\frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau^2\right)d\tau^2.\end{aligned} \tag{10}$$

Applying the fact as follows to the above equation:

$$\frac{a}{2}\exp(-a \mid z \mid) = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2} \exp\left(-\frac{a^2 s}{2}\right)ds, \quad a > 0, \tag{11}$$

we have

$$\begin{aligned}p(\beta, \lambda, \sigma^2, r \mid X, Y) &\propto \int_0^\infty \exp \\ &\cdot \left(-\frac{\|Y - X\beta\|^2 + r\beta^T L\beta}{2\sigma^2}\right) \exp \\ &\cdot \left(-\frac{r\beta D^{-1}\beta}{2\sigma^2}\right)\frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\tau^2\right)d\tau^2 \\ &= \exp\left(-\frac{\|Y - X\beta\|^2 + r\lambda \mid \beta \mid + r\beta^T L\beta}{2\sigma^2}\right).\end{aligned} \tag{12}$$

Thus, maximizing the posterior distribution is equivalent to minimizing the following equation:

$$L(r, \lambda, \beta) = (y - X\beta)^T(y - X\beta) + r\lambda|\beta|_1 + \lambda\beta^T L\beta, \tag{13}$$

which has the same regularization term as the method proposed in [19].

We assign the prior distribution for σ^2 as follows:

$$\sigma^2 \sim \text{Inverse Gamma}(a, b). \tag{14}$$

And we assign the following prior for the hyperparameters r and λ:

$$r \sim \text{Gamma}(c, d),$$
$$\lambda \sim \text{Gamma}(e, f). \tag{15}$$

Then, the hierarchical Bayesian model is

$$\begin{aligned} Y|X, \beta, \sigma^2 &\sim N_n\left(X\beta, \sigma^2 I_n\right), \\ \beta|\sigma^2, \tau^2, r &\sim N_p\left(0, \frac{\sigma^2}{r}\Lambda^{-1}\right), \\ \tau^2|\lambda &\sim |\Lambda|^{-1/2}\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau^2\right), \\ \sigma^2 &\sim \text{Inverse Gamma}(a, b), \\ r &\sim \text{Gamma}(c, d), \\ \lambda &\sim \text{Gamma}(e, f). \end{aligned} \tag{16}$$

2.2. Gibbs Sampling Method. The likelihood is

$$p\left(y \mid X, \beta, \sigma^2\right) \propto \left(\sigma^2\right)^{-n/2}\exp\left(-\frac{(Y - X\beta)^T(Y - X\beta)}{2\sigma^2}\right). \tag{17}$$

According to the above hierarchical model and the likelihood, the joint posterior distribution on data is

$$\begin{aligned} p\left(\beta, \sigma^2, \tau^2, \lambda^2, r \mid Y, X\right) &\propto \left(\sigma^2\right)^{-n/2}\exp \\ &\cdot\left(-\frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2}\right)\left(\sigma^2\right)^{-p/2}r^{p/2}|\Lambda|^{1/2}\exp \\ &\cdot\left(-\frac{r\beta^T\Lambda\beta}{2\sigma^2}\right)|\Lambda|^{-1/2}\prod_{j=1}^{p}\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau_j^2\right)\left(\sigma^2\right)^{-a}\exp \\ &\cdot\left(-\frac{b}{\sigma^2}\right)r^{-c}\exp\left(-dr\right)\left(\lambda^2\right)^{-e}\exp\left(-f\lambda^2\right). \end{aligned} \tag{18}$$

Due to the fact that all the prior distributions are conjugated, the full conditional posterior distributions for the parameters have closed forms.

$$\begin{aligned} p\left(\beta, |\sigma^2, \tau^2, r, Y, X\right) &\propto \exp\left(-\frac{(Y - X\beta)^T(Y - X\beta)}{2\sigma^2}\right)\exp \\ &\cdot\left(-\frac{r\beta\Lambda\beta}{2\sigma^2}\right) \propto \exp\left(-\frac{\left(X'X + r\Lambda\right)\beta^2 - 2YX\beta}{2\sigma^2}\right). \end{aligned} \tag{19}$$

Let $\mu = \left(X'X + r\Lambda\right)^{-1}X'Y, \Sigma = \sigma^2\left(X'X + r\Lambda\right)^{-1}$, we have

$$\beta|\sigma^2, \tau^2, r, X, Y \sim N_p(\mu, \Sigma), \tag{20}$$

$$\begin{aligned} p\left(\sigma^2 \mid \beta, \tau^2, r, Y, X\right) &\propto \left(\sigma^2\right)^{-n/2}\exp\left(-\frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2}\right)\left(\sigma^2\right)^{-p/2}\exp \\ &\cdot\left(-\frac{r\beta^T\Lambda\beta}{2\sigma^2}\right)\left(\sigma^2\right)^{-a}\exp\left(-\frac{b}{\sigma^2}\right) \propto \left(\sigma^2\right)^{-n+p/2-a}\exp \\ &\cdot\left(\left(-\frac{(y - X\beta)^T(y - X\beta) + r\beta^T\Lambda\beta}{2} + b\right)\frac{1}{\sigma^2}\right), \end{aligned} \tag{21}$$

$$\sigma^2|\beta, \tau^2, r, Y, X \sim \text{Inverse Gamma} \cdot\left(\frac{n + p}{2} + a, \frac{(y - X\beta)^T(y - X\beta) + r\beta^T\Lambda\beta}{2} + b\right), \tag{22}$$

$$p\left(\tau^2 \mid \beta, \sigma^2, \lambda^2, r\right) \propto \exp\left(-\frac{r\beta^T\Lambda\beta}{2\sigma^2}\right)\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau^2\right). \tag{23}$$

This implies that τ^2 follows a generalized inverse Gaussian distribution:

$$\tau_j^2|\beta, r, \sigma^2, \lambda^2 \sim \text{GIG}\left(\frac{1}{2}, \lambda^2, \frac{r\beta_j^2}{\sigma^2}\right), \quad j = 1, 2, \cdots, p, \tag{24}$$

$$\begin{aligned} p\left(r \mid \beta, \sigma^2, \tau^2\right) &\propto r^{p/2}\exp\left(-\frac{r\beta^T\Lambda\beta}{2\sigma^2}\right)r^c\exp\left(-dr\right) \\ &\propto r^{p/2+c}\exp\left(-\left(\frac{\beta^T\Lambda\beta}{2\sigma^2} + d\right)r\right), \end{aligned} \tag{25}$$

$$r|\sigma^2, \beta, \tau^2 \sim \text{Gamma}\left(\frac{p}{2} + c, \frac{\beta^T\Lambda\beta}{2\sigma^2} + b\right), \tag{26}$$

$$\begin{aligned} p\left(\lambda^2 \mid \tau^2\right) &\propto \prod_{j=1}^{p}\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau_j^2\right)\left(\lambda^2\right)^e\exp\left(-f\lambda^2\right) \\ &\propto \left(\lambda^2\right)^{p+e}\exp\left(-\left(\frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + f\right)\lambda^2\right), \end{aligned} \tag{27}$$

$$\lambda^2|\tau^2 \sim \text{Gamma}\left(p + e, \frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + f\right). \tag{28}$$

The Gibbs sampling scheme iterates as follows:

(1) Update β by sampling from (20)

(2) Update σ^2 by sampling form (22)

(3) Update τ^2 by sampling from (24)

(4) Update r by sampling from (26)

(5) Update λ by sampling from (28)

2.3. The Binary Response Case. Binary data such as absence or presence or different types of a disease are often used as response variables in gene selection problems. To perform binary classification, we use probit regression using auxiliary variables. Then, the model can be represented as follows:

$$P(y_i = 1) = X_i\beta, \tag{29}$$

where X_i is the ith sample and $P(y_i = 1)$ is the probability of $y_i = 1$. Here, latent variables $Z = (z_1, z_2, \cdots, z_n)$ are defined as

$$z_i = X_i\beta + \varepsilon, \varepsilon \sim N_n\left(0, \sigma^2 I_n\right). \tag{30}$$

Then, the full conditional posterior distribution for each z_i is truncated normal:

$$z_i|\beta, X_i, y_i \sim \begin{cases} N\left(X_i\beta, \sigma^2\right)I(z_i > 0), & y_i = 1, \\ N\left(X_i\beta, \sigma^2\right)I(z_i \le 0), & \text{otherwise.} \end{cases} \tag{31}$$

And Z follows a multivariate truncated normal distribution:

$$p\left(Z \mid \beta, \sigma^2, X, Y\right) \propto N_n\left(X\beta, \sigma^2 I_n\right)\prod_{i=1}^{n} I(A_i), \tag{32}$$

$$A_i = \begin{cases} \{Z_i \mid Z_i > 0\}, (Y_i = 1), \\ \{Z_i \mid Z_i \le 0\}, (Y_i = 0). \end{cases} \tag{33}$$

Sampling from this distribution directly is difficult. We use the method proposed in [26] to sample this latent variable.

Then, the hierarchical Bayesian model is

$$\begin{aligned} Z|X, Y, \beta, \sigma^2 &\sim N_n\left(X\beta, \sigma^2 I_n\right)\prod_{i=1}^{n} I(A_i), \\ \beta|\sigma^2, \tau^2, r &\sim N_p\left(0, \frac{\sigma^2}{r}\Lambda^{-1}\right), \\ \tau^2|\lambda &\sim |\Lambda|^{-1/2}\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau^2\right), \\ \sigma^2 &\sim \text{Inverse Gamma}(a, b), \\ r &\sim \text{Gamma}(c, d), \\ \lambda &\sim \text{Gamma}(e, f). \end{aligned} \tag{34}$$

To derive the Gibbs sampling scheme, we only need to replace Y with Z in the Gibbs sampling scheme defined in Section 2.2. And the latent variables Z are sampled from (32).

Table 1: Binary classification microarray datasets used.

Dataset name	No. genes	Samples	P/N	References
Leukemia	1883	72	47/25	[28]
DLBCL	2427	77	58/24	[29]
Prostate	3238	102	50/52	[30]
GSE412	3234	108	60/48	[31]
GSE4922	4476	204	70/134	[26]

3. Results

3.1. Datasets and Preprocessing. To demonstrate the effectiveness of our methods, a regression microarray dataset and 5 real-life binary classification microarray datasets were tested in this paper, which are described as follows. The pathway information was obtained from the KEGG database.

A breast cancer dataset was used to predict the survival time of patients [27]. We used gene expression profiles of 76 patients. Each patient was measured with 24481 probes. 3592 genes were found in the KEGG database from this dataset. We used the logarithm of survival times of patients as the response variable in this dataset.

The other 5 binary classification microarray datasets are shown in Table 1. No. genes mean the genes we found both existing in the microarray dataset and KEGG pathway database.

Lastly, the gene expression values were normalized; thus, its mean and standard deviation are 0 and 1.

3.2. Parameter Settings. In the procedure of Bayesian network-constrained regularization, we recommend small values for a, b, c, d, e, f in (16) and we set these values to 0.01 in our experiments. The Gibbs sampling iteration was conducted 6000 times, and we chose the second half of the samples to estimate the regression parameters. The posterior estimates of all parameters were obtained through the posterior averages of the chains. For the classification problem, the classifiers were built by a support vector machine (SVM). In this paper, we used the radial basic function as the kernel function in SVM. And the regularization parameter and the kernel width parameter were optimized by a grid search approach. We used Libsvm [32] to model the SVM.

3.3. Results and Analysis. In this section, we will describe the results on 6 microarray gene expression datasets (Table 1) to evaluate the performance of the proposed method. Our method was compared with the other three Bayesian regularized regression methods, including Bayesian Lasso, Bayesian Elastic Net, and Bayesian Fused Lasso. A comprehensive review of these methods can be found in [23]. When $L = I$, which means we know nothing about the pathway structure, the Bayesian network-constrained regularization is equivalent to Bayesian Elastic Net. And when $L = O$, our method is equivalent to Bayesian Lasso. These three methods can also be extended to perform binary classification through an auxiliary variable. We also used Gibbs sampling to perform parameter estimation. Previous review [23] also shows that these three Bayesian methods' performances are similar to

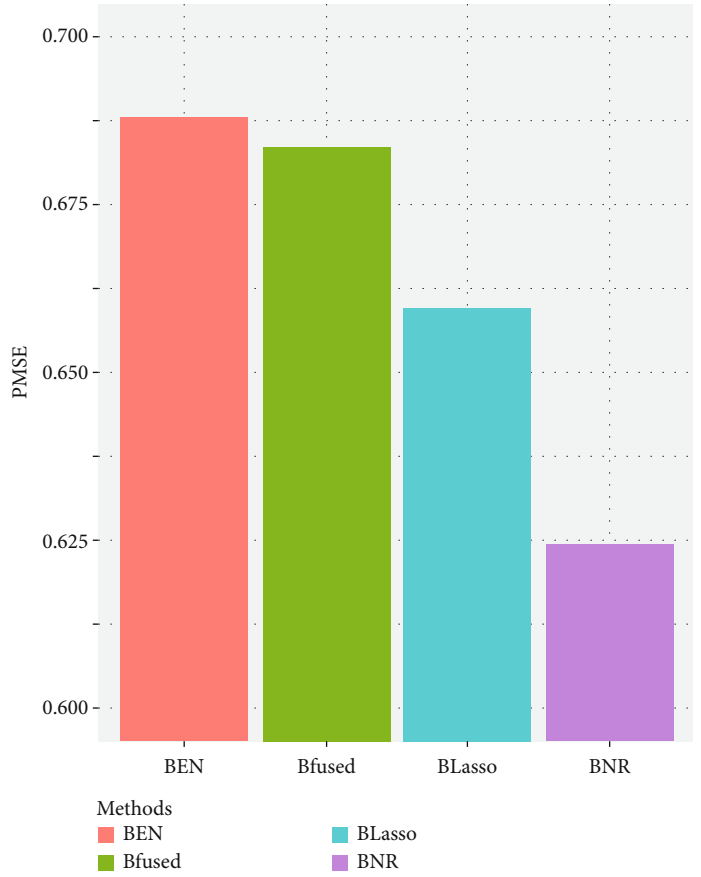

Figure 1: PMSE performance on regression microarray dataset.

and in some cases better than the frequentist methods. Prediction mean square error was used to evaluate the performance on regression problem. Meanwhile, ACC and AUC were used as the evaluation criteria for binary classification problem. According to previous studies, the number of important genes is probably about 50 [28]; thus, we selected the top 50 genes based on the absolute value of their regression coefficient for the binary classification problem.

Figure 1 shows the performance of all the four methods on the regression microarray dataset. And the classification performances on the five binary classification microarray datasets are summarized in Table 2. In the binary classification datasets, the first three datasets are usually treated as easy classification datasets, while the other two datasets are relatively hard to classify. From Figure 1, we can see that the PMSE of our method is lower than other Bayesian methods. Table 2 also shows that on the four easy binary classification datasets, our method achieves the highest ACC and AUC. In the other two hard classification datasets, our method achieves the highest ACC and AUC on GSE412. Although the AUC of Bayesian Elastic Net is higher than our method on GSE4922, our method achieves the highest ACC. In general, Bayesian network-constrained regularization shows better prediction and classification ability than other three Bayesian methods, which is similar to the results implied by [15]. Since our method can be transferred to

Table 2: Comparison of results of 4 Bayesian methods.

Dataset	Methods	AUC	ACC
Leukemia	BEN	0.9955	0.9600
	BFused	1	0.9733
	BLasso	1	0.9447
	BNR	1	0.9733
DLBCL	BEN	0.9674	0.9223
	BFused	0.9674	0.9223
	BLasso	0.9485	0.9223
	BNR	0.9958	0.9482
Prostate	BEN	0.9784	0.9414
	BFused	0.9655	0.9314
	BLasso	0.9784	0.9419
	BNR	0.9900	0.9510
GSE412	BEN	0.9428	0.8498
	BFused	0.9046	0.8619
	BLasso	0.9541	0.8792
	BNR	0.9637	0.9074
GSE4922	BEN	0.6274	0.6666
	BFused	0.6028	0.6523
	BLasso	0.6132	0.6860
	BNR	0.6132	0.6860

Table 3: Description of top 18 genes of GSE4922.

Gene symbol	Description	Reference
SYCP3*	Synaptonemal complex protein 3	[35]
CDKN2A*	Cyclin dependent kinase inhibitor 2A	[36]
PLB1*	Phospholipase B1	[37]
CTNNBIP1*	Catenin beta-interacting protein 1	[38]
GBE1*	1,4-Alpha-glucan-branching enzyme 1	[39]
SMURF1*	SMAD-specific E3 ubiquitin protein ligase 1	[40]
NR1H4	Nuclear receptor subfamily 1 group H member 4	/
PDE11A	Phosphodiesterase 11A	/
UGT1A1*	UDP glucuronosyltransferase family 1 member A1	[41]
FGF19*	Fibroblast growth factor 19	[42]
OR51B4*	Olfactory receptor family 51 subfamily B member 4	[43]
RAB7A*	RAB7A, member RAS oncogene family	[44]
SDHD*	Succinate dehydrogenase complex subunit D	[45]
IFNA8	Interferon alpha 8	/
VANGL2*	VANGL planar cell polarity protein 2	[46]
UMPS	Uridine monophosphate synthetase	/
CASP3	Caspase 3	[47]
SUFU	SUFU negative regulator of hedgehog signaling	[48]

*The gene was reported as an oncogene in previous literatures.

Bayesian Lasso or Bayesian Elastic Net when the normalized Laplacian matrix $L = O$ or $L = I$, the results also show that pathway information indeed contributes to the accuracy of the gene selection.

Consistent with previous studies [33, 34], all the Bayesian regularization regression methods could classify Leukemia, DLBCL, Prostate, and GSE412 dataset accurately. However, the performances of all the methods were poor on GSE 4922 dataset. Therefore, we demonstrate the effectiveness of our method by selecting the top 18 genes which make the prediction accuracy to achieve the highest value and most of those genes are associated with breast cancer (Table 3).

4. Conclusion

In this paper, we propose a Bayesian approach to perform gene selection, which can incorporate the pathway information as prior biological knowledge through network-constrained regularization to improve the accuracy of gene selection. All the prior distributions we propose are strictly conjugated; thus, all the conditional posteriors of the model are available in closed form. An auxiliary variable is also introduced to extend the regression model to perform binary classification. An efficient Gibbs sampling method is used to estimate regression coefficients and tune parameters simultaneously, which can perform feature filter feasible for high dimensional microarray datasets. The performance of the proposed method is demonstrated by applying it to a regression microarray dataset and five binary classification microarray datasets. The results show that compared with Bayesian Lasso, Bayesian Elastic Net, and Bayesian Fused Lasso, our method performs better both in prediction and classification. And the pathway information indeed improves the accuracy of gene selection.

Data Availability

The breast cancer dataset could be obtained from the R package breast cancer NKI. Leukemia, DLBCL, and Prostate datasets are available on the website http://portals.broadinstitute.org/cgi-bin/cancer/. GSE412 and GSE4922 datasets are available in the GEO of NCBI under accession GSE412 and GSE4922.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work is supported in part by the National Science Foundation of China, under Grant 61173111.

References

[1] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, "Microarray and its applications," *Journal of Pharmacy & Bioallied Sciences*, vol. 4, Supplement 2, pp. S310–S312, 2012.

[2] H. Wang, X. Jing, and B. Niu, "A discrete bacterial algorithm for feature selection in classification of microarray gene

expression cancer data," *Knowledge-Based Systems*, vol. 126, pp. 8–19, 2017.

[3] N. Dessì and B. Pes, "Similarity of feature selection methods: an empirical study across data intensive classification tasks," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4632–4642, 2015.

[4] L. Lu, K. A. Townsend, and B. J. Daigle Jr., "GEOlimma: differential expression analysis and feature selection using pre-existing microarray data," *BMC Bioinformatics*, vol. 22, no. 1, p. 44, 2021.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[6] S. Sun, Q. Peng, and X. Zhang, "Global feature selection from microarray data using Lagrange multipliers," *Knowledge-Based Systems*, vol. 110, pp. 267–274, 2016.

[7] K. Kanti Ghosh, S. Begum, A. Sardar et al., "Theoretical and empirical analysis of filter ranking methods: experimental study on benchmark DNA microarray data," *Expert Systems with Applications*, vol. 169, p. 114485, 2021.

[8] W. You, Z. Yang, and G. Ji, "Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1463–1475, 2014.

[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] Z. Hui and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, vol. 67, no. 5, pp. 768–768, 2005.

[11] H. Fröhlich, "Network based consensus gene signatures for biomarker discovery in breast cancer," *PLoS One*, vol. 6, no. 10, article e25364, 2011.

[12] G. Michailidis, "Statistical challenges in biological networks," *Journal of Computational and Graphical Statistics*, vol. 21, no. 4, pp. 840–855, 2012.

[13] M. Y. Wu, X. F. Zhang, D. Q. Dai, L. Ou-Yang, Y. Zhu, and H. Yan, "Regularized logistic regression with network-based pairwise interaction for biomarker identification in breast cancer," *BMC Bioinformatics*, vol. 17, no. 1, p. 108, 2016.

[14] J. Su, B. J. Yoon, and E. R. Dougherty, "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," *PLoS One*, vol. 4, no. 12, article e8161, 2009.

[15] L. H. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[16] C. Li and H. Li, "Variable selection and regression analysis for graph-structured covariates with an application to genomics," *The Annals of Applied Statistics*, vol. 4, no. 3, pp. 1498–1516, 2010.

[17] J. Liu, J. Huang, and S. Ma, "Incorporating network structure in integrative analysis of cancer prognosis data," *Genetic Epidemiology*, vol. 37, no. 2, pp. 173–183, 2013.

[18] W. Zhang, Y. W. Wan, G. I. Allen, K. Pang, M. L. Anderson, and Z. Liu, "Molecular pathway identification using biological network-regularized logistic models," *BMC Genomics*, vol. 14, Supplement 8, p. S7, 2013.

[19] H. Jiang, T. Hong, T. Wang et al., "Gene expression profiling of human bone marrow mesenchymal stem cells during osteogenic differentiation," *Journal of Cellular Physiology*, vol. 234, no. 5, pp. 7070–7077, 2019.

[20] Y. Fan, X. Wang, and Q. Peng, "Inference of gene regulatory networks using Bayesian nonparametric regression and topology information," *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 8307530, 8 pages, 2017.

[21] Z. Hui, *The Bayesian Lasso*, 2008.

[22] Q. Li and N. Lin, "The Bayesian elastic net," *Bayesian Analysis*, vol. 5, no. 1, pp. 151–170, 2010.

[23] G. Casella, M. Ghosh, J. Gill, and M. Kyung, "Penalized regression, standard errors, and Bayesian lassos," *Bayesian Analysis*, vol. 5, no. 2, pp. 369–412, 2010.

[24] M. Kanehisa, M. Araki, S. Goto et al., "KEGG for linking genomes to life and the environment," *Nucleic Acids Research*, vol. 36, Database issue, pp. D480–D484, 2008.

[25] L. Held and C. C. Holmes, "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Analysis*, vol. 1, no. 1, pp. 145–168, 2006.

[26] A. V. Ivshina, J. George, O. Senko et al., "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer," *Cancer Research*, vol. 66, no. 21, pp. 10292–10301, 2006.

[27] L. J. van 't Veer, H. Dai, M. J. van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[28] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[29] M. A. Shipp, K. N. Ross, P. Tamayo et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[30] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[31] M. H. Cheok, W. Yang, C. H. Pui et al., "Treatment-specific changes in gene expression discriminate _in vivo_ drug response in human leukemia cells," *Nature Genetics*, vol. 34, no. 1, pp. 85–90, 2003.

[32] V. Ferrari, *Libsvm : A Library for Support Vector Machines*, 2008.

[33] M. Momenzadeh, M. Sehhati, and H. Rabbani, "A novel feature selection method for microarray data classification based on hidden Markov model," *Journal of Biomedical Informatics*, vol. 95, p. 103213, 2019.

[34] A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm.," *Expert Systems with Applications*, vol. 166, p. 114012, 2021.

[35] M. B. Mobasheri, R. Shirkoohi, and M. H. Modarressi, "Synaptonemal complex protein 3 transcript analysis in breast cancer," *Iranian Journal of Public Health*, vol. 45, no. 12, pp. 1618–1624, 2016.

[36] J. Lubiński, B. Górski, T. Huzarski et al., "BRCA1-positive breast cancers in young women from Poland," *Breast Cancer Research and Treatment*, vol. 99, no. 1, pp. 71–76, 2006.

[37] A. Ueda, K. Oikawa, K. Fujita et al., "Therapeutic potential of PLK1 inhibition in triple-negative breast cancer," *Laboratory Investigation*, vol. 99, no. 9, pp. 1275–1286, 2019.

[38] N. Mukherjee, H. Dasgupta, R. Bhattacharya et al., "Frequent inactivation of MCC/CTNNBIP1 and overexpression of phospho-beta- catenin^{Y654} are associated with breast carcinoma: clinical and prognostic significance," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1862, no. 9, pp. 1472–1484, 2016.

[39] A. H. Birch, M. C. J. Quinn, A. Filali-Mouhim, D. M. Provencher, A. M. Mes-Masson, and P. N. Tonin, "Transcriptome analysis of serous ovarian cancers identifies differentially expressed chromosome 3 genes," *Molecular Carcinogenesis*, vol. 47, no. 1, pp. 56–65, 2008.

[40] A. Kwon, H. L. Lee, K. M. Woo, H. M. Ryoo, and J. H. Baek, "SMURF1 plays a role in EGF-induced breast cancer cell migration and invasion," *Molecules and Cells*, vol. 36, no. 6, pp. 548–555, 2013.

[41] S.-H. Kuo, S. Y. Yang, S. L. You et al., "Polymorphisms of ESR1, UGT1A1, HCN1, MAP3K1 and CYP2B6 are associated with the prognosis of hormone receptor-positive early breast cancer," *Oncotarget*, vol. 8, no. 13, pp. 20925–20938, 2017.

[42] K. H. Tiong, B. S. Tan, H. L. Choo et al., "Fibroblast growth factor receptor 4 (FGFR4) and fibroblast growth factor 19 (FGF19) autocrine enhance breast cancer cells survival," *Oncotarget*, vol. 7, no. 36, pp. 57633–57650, 2016.

[43] C. Zhang, H. Zhao, J. Li et al., "The identification of specific methylation patterns across different cancers," *PLoS One*, vol. 10, no. 3, article e0120361, 2015.

[44] J. Xie, Y. Yan, F. Liu et al., "Knockdown of Rab7a suppresses the proliferation, migration, and xenograft tumor growth of breast cancer cells," *Bioscience Reports*, vol. 39, no. 2, 2019.

[45] W. Yu, X. He, Y. Ni, J. Ngeow, and C. Eng, "Cowden syndrome-associated germline SDHD variants alter PTEN nuclear translocation through SRC-induced PTEN oxidation," *Human Molecular Genetics*, vol. 24, no. 1, pp. 142–153, 2015.

[46] J. Hatakeyama, J. H. Wald, I. Printsev, H. Y. H. Ho, and K. L. Carraway, "Vangl1 and Vangl2: planar cell polarity components with a developing role in cancer," *Endocrine-Related Cancer*, vol. 21, no. 5, pp. R345–R356, 2014.

[47] X. Pu, S. J. Storr, Y. Zhang et al., "Caspase-3 and caspase-8 expression in breast cancer: caspase-3 is associated with survival," *Apoptosis*, vol. 22, no. 3, pp. 357–368, 2017.

[48] F. Alimirah, X. Peng, A. Gupta et al., "Crosstalk between the vitamin D receptor (VDR) and miR-214 in regulating SuFu, a hedgehog pathway inhibitor in breast cancer cells," *Experimental Cell Research*, vol. 349, no. 1, pp. 15–22, 2016.