# Knowledge-Guided Bayesian Support Vector Machine for HighDimensional Data with Application to Analysis of Genomics Data 

Wenli Sun ${ }^{*}, \ddagger$, Changgee Chang ${ }^{*}, \ddagger$, Yize Zhao ${ }^{\dagger}$, and Qi Long ${ }^{*}, \S$<br>*Department of Biostatistics, Epidemiology and Informatics The University of Pennsylvania, Philadelphia, PA, 19104<br>${ }^{\dagger}$ Department of Healthcare Policy and Research Weill Cornell Medicine, Cornell University, New York, NY, 10065


#### Abstract

Support vector machine (SVM) is a popular classification method for the analysis of wide range of data including big data. Many SVM methods with feature selection have been developed under frequentist regularization or Bayesian shrinkage frameworks. On the other hand, the importance of incorporating a priori known biological knowledge, such as gene pathway information which stems from the gene regulatory network, into the statistical analysis of genomic data has been recognized in recent years. In this article, we propose a new Bayesian SVM approach that enables the feature selection to be guided by the knowledge on the graphical structure among predictors. The proposed method uses the spike-and-slab prior for feature selection, combined with the Ising prior that encourages group-wise selection of the predictors adjacent to each other on the known graph. Gibbs sampling algorithm is used for Bayesian inference. The performance of our method is evaluated and compared with existing SVM methods in terms of prediction and feature selection in extensive simulation settings. In addition, our method is illustrated in the analysis of genomic data from a cancer study, demonstrating its advantage in generating biologically meaningful results and identifying potentially important features.


## Keywords

Bayesian support vector machine; knowledge-guided; pathway graph information; Spike-and-slab prior; Ising prior

## I. Introduction

The support vector machine (SVM) [1] is a popular classification method in data mining and machine learning. It has achieved great successes in various data mining tasks such as image classification, pattern recognition and forecasting [2], [3]. Many SVM approaches with feature selection have been introduced in the literature, among which the ones that use a specific penalty on the coefficients (normal vector) are popular. The $L_{1}$ norm penalized SVM ( $L_{1}$ SVM) [4]-[6] applies the LASSO penalty [7] into SVM. The SVM with a non-

[^0]convex penalty [8], [9] (SCADSVM) adopts the smoothly clipped absolute deviation penalty [10] to alleviate the bias in estimating nonzero coefficients. Double regularization SVM (DrSVM) [11] combines the $L_{1}$ and $L_{2}$ norm to encourage the selection of correlated features. $L_{\infty}$ penalized SVM [12] encourages all the features in the same group to be selected simultaneously. These approaches and their variants have proven their superiority during the past two decades. In this era of big data, however, where the multi-omics data need to be analyzed beyond the GWAS or genomic studies, it is imperative that new innovation is required.

In some real world applications, some prior knowledge on data may be available, which can be integrated into the analysis and improve the power of detecting important signals. For example, a comprehensive review [13] summarizes the methods that incorporate such prior knowledge into SVM, while classifying the prior knowledge into two categories: classinvariance and knowledge on the data. The class-invariance stands for the invariance of the class to a transformation of the input pattern, and the knowledge on the data refers to such knowledge as the information in unlabeled samples, the imbalance of the training set, and the quality of the data. This article aims to consider the prior biological knowledge that is represented by the pathway graph information. Enormous genomic studies have revealed that the genes influence phenotypes through a complex regulatory network represented by a directed acyclic graph, where each gene is expressed by a node and the promotion/inhibition relationships between the genes are indicated by the edges. The network is composed of multiple gene pathways and the knowledge on the pathway graphs is publicly available [14] and still growing. Recent works [15]-[18] have attempted to incorporate the pathway graph information, motivated from its biological interpretation, by encouraging group-wise selection of adjacent predictors. They demonstrate that the incorporation of such prior knowledge offers a great promise toward the improved predictive accuracy and the increased power of detecting key molecular signatures and acting pathways. In addition, the resulting prediction models become more interpretable as they help select key biological pathways and likely lead to idenfication of potential molecular targets for treatments [19].

However, only very few works [20] in the SVM framework can incorporate the prior knowledge on the correlation structure among features. At the same time, most penalization based SVM methods [4]-[6], [8]-[12], [20] provide point estimates, failing to systematically quantify the uncertainty of the estimates. Therefore, we propose a knowledge-guided Bayesian SVM (KBSVM), which is a Bayesian approach capable of incorporating the graphical structure of features. As a Bayesian method, our approach can provide not only the uncertainty information but also the ensemble inference, which leads to more accurate and reliable performance in both classification and feature selection. Some Bayesian approaches [21]-[23] have been proposed to perform feature selection by introducing shrinkage priors on the normal vector, but to the best of our knowledge, none of them utilizes the graph structure among the features. Also, note that the exising frequentist approaches [11], [12], [20] either force the coefficients to have similar values or apply smoothing between all the member coefficients in a pathway group, which may cause bias. Unlike those works, our approach uses the pathway graph information, which is more refined than the pathway membership information, and encourages only the joint selection among the adjacent
features rather than smoothing their coefficient estimates. This helps achieve enhanced performance without the expense of bias.

In the proposed model, we employ the spike-and-slab prior [24] for feature selection. The selection status of each feature is represented by a latent binary variable. The gaussian prior with small variance (spike) is assigned for the inactive coefficient, and the gaussian prior with large variance (slab) is assigned for the active coefficient. This prior shrinks the inactive coefficients toward zero and reduces the bias for the active coefficients. In addition to the spike-and-slab prior, we assign the Ising prior [25] to the latent indicator variables to reflect the graphical structure of the predictors. This prior encourages any pair of predictors which are adjacent on the graph to have the same selection status. Note that [16] uses the Markov random field (MRF) prior for the latent indicator variables, which is similar to the Ising prior. The difference is that, while the MRF prior only has the selected features encourage the selection of the adjacent features, the Ising prior also has the unselected features encourage the deselection of the neignboring features. Therefore, our model prefers both group-wise inclusion and exclusion of adjacent features, which further improves the prediction performance.

We present the Gibbs sampling algorithm [26] that performs the Bayesian prediction and feature selection. We employ the the state-of-the-art data augmentation techniques [27] to make our algorithm efficient and easy to implement. Another contribution to the Bayesian SVM literature is that we propose the corrected pseudo-likelihood. Having the proper form of likelihood allows other model parameter to have a better interpretation, which will be elaborated in Section II-A. The performance of the proposed method is evaluated in comparison with other existing SVM methods in terms of prediction and feature selection under extensive simulation scenarios. In addition, we illustrate an application of our method to the analysis of genomic data from a cancer study, further demonstrating its advantage in identifying important features and yielding biologically meaningful results.

The rest of the article is organized as follows. In Sections II and III, we describe the proposed models and the computing algorithms. In Section IV, we conduct simulation to evaluate our approach in comparison with several existing approaches. In Section V, we apply our approach to a TCGA glioblastoma dataset. We conclude with a brief discussion in Section VI.

## II. Model

## A. Likelihood

Suppose there are $n$ samples in the training set of data where $y_{i} \in\{-1,1\}$ are the binary outcome variables and $\mathbf{x}_{i}$ are the $(p+1)$ dimensional feature vector including the intercept. The classical SVM seeks to find a classification function $f$ to separate the two classes by minimizing

$$
\begin{equation*}
\Theta(\beta)=\kappa \sum_{i=1}^{N} \max \left(1-y_{i} f\left(\mathbf{x}_{i}\right), 0\right)+R(f), \tag{1}
\end{equation*}
$$

Proc IEEE Int Conf Big Data. Author manuscript; available in PMC 2019 December 01.
where $\sum_{i=1}^{N} \max \left(1-y_{i} f\left(\mathbf{x}_{i}\right), 0\right)$ is the hinge loss function and $R$ is a regularization function controlling the complexity of $f$. The tuning parameter $\boldsymbol{\kappa}$ can be seen as part of the regularization parameters. For the linear classifier $f=\mathbf{x}_{i}^{\prime} \beta$, minimizing the objective function
(1) is equivalent to find the mode of the following pseudo-posterior density [28].

$$
\begin{aligned}
p(\beta \mid X, \mathbf{y}, \kappa) & \propto p(\beta) L(\mathbf{y} \mid X, \beta, \kappa) \\
& \propto p(\beta) \prod_{i=1}^{n} k e^{-2 \kappa \max \left(1-y_{i} \mathbf{x}_{i}^{\prime} \beta, 0\right)} .
\end{aligned}
$$

Note that $\kappa e^{-2 \kappa m a x}\left(1-y_{i} \mathbf{x}_{i}^{\prime} \beta, 0\right)$ is the pseudo-likelihood contribution from the $i$-th observation (as it does not sum to a constant) and obviously prefers the coefficients that reduces the hinge loss. Note that this pseudo-likelihood is not exactly same as the one that has been widely used in the Bayesian SVM literature. We correct the one used in [27], [28] by multiplying it by $\kappa$. This newly proposed pseudo-likelihood gives a plausible interpretation for the parameter $\kappa$, the parameter $\kappa$ learns the overall (average) scale of the errors. In fact, the posterior distribution of $\kappa$ converges to a degenerate distribution concentrated at 0 under the previous pseudo-likelihood, as the sample size increases. Note also that another important role of the parameter $\kappa$ is to allow the normal vector $\beta$ to explore its parameter space more freely in MCMC.

We use the Gamma prior for $\kappa \sim \mathscr{G}\left(a_{\kappa}, b_{\kappa}\right)$, where $a_{\kappa}$ and $b_{\kappa}$ are hyperparameters representing the shape and the rate parameters of the Gamma distribution, the values of which can be chosen in an uninformative or data-driven manner.

## B. Spike-and-Slab and Ising Prior

As aforementioned, we use the spike-and slab prior [24] for $\boldsymbol{\beta}$ to perform the feature selection. We introduce the latent binary variables $\gamma_{j}$ indicating the inclusion of the $j$-th feature into the model, and assume $\beta_{j} \mid \gamma_{j} \propto N\left(0, v_{j}^{2}\right)$

$$
p(\beta \mid \gamma)=C \prod_{j=1}^{p+1} v_{j} v^{-\frac{1}{2}} e^{-\frac{\beta_{j}^{2}}{2 v_{j}}},
$$

where $v_{j}=\gamma_{j} \sigma_{1}^{2}+\left(1-\gamma_{j}\right) \sigma_{0}^{2}$ with $\sigma_{0}^{2}<\sigma_{1}^{2}$ and $C$ is the normalizing constant. If $\gamma_{j}=0$, then the prior of $\beta_{j}$ has the spike variance $v_{j}=\sigma_{0}^{2}$ and $\beta_{j}$ is shrunk toward 0 . If $\gamma_{j}=1$, then the prior of $\beta_{j}$ has the slab variance $v_{j}=\sigma_{1}^{2}$ and $\beta_{j}$ is less biased.

Let $\mathscr{G}=\langle V, E\rangle$ be a pathway graph where $V=\{1, \ldots, p+1\}$ is the set of genes and $E \subset\{(j$, $k): j, k \in V, j \neq k\}$ be the set of edges representing (partial) correlations among the genes.

Let $G$ be the adjacency matrix of $\mathscr{G}$. To incorporate the graph structure between predictors, we use the Ising prior for $\gamma$ given as follows.

$$
\begin{equation*}
p(\gamma)=C_{\mu, \eta} e^{\left.-\mu \sum_{j} \gamma_{j}+\eta \sum_{j \neq k} G_{j k} \mathrm{D} \gamma_{j}=\gamma_{k}\right)} \tag{2}
\end{equation*}
$$

where $C_{\mu, \eta}$ is the normalizing constant and $\mathbb{D}(\cdot)$ indicator function. The tuning parameters $\mu$ controls the sparsity of $\gamma$ and $\eta$ controls the smoothness of $\gamma$ over $E$. Note that (2) encourages $\gamma_{k}=1$ if $\gamma_{j}=1$ and $G_{j k}=1$ and promotes $\gamma_{k}=0$ if $\gamma_{j}=0$ and $G_{j k}=1$. Therefore, the group-wise selection of the $j$-th and the $k$-th genes are encouraged if there is an edge between them.

The Ising prior is slightly different from the Markov random field prior proposed in the literature earlier [16], [29].

$$
\begin{equation*}
p(\gamma)=C_{\mu, \eta} e^{-\mu \sum_{j} \gamma_{j}+\eta \sum_{j \neq k} G_{j k} \gamma_{j} \gamma_{k}} \tag{3}
\end{equation*}
$$

Note that (3) only encourages $\gamma_{k}=1$ if $\gamma_{j}=1$ and $G_{j k}=1$. However, there is little difference from the computational point of view because $\mathbb{\square}\left(\gamma_{j}=\gamma_{k}\right)=2 \gamma_{j} \gamma_{k}-\gamma_{j}-\gamma_{k}+1$.

## III. Posterior inference and Computation

Let $\mathbf{z}_{i}=y_{i} X_{i}$ and $\mathrm{Z}=\left[\mathbf{z}_{1}, \ldots, \mathbf{z}_{n}\right]^{\prime}$. To facilitate the Bayesian compututation, we use the variable augmentation technique; see, for example, [27], [28].

$$
\begin{equation*}
e^{-2 \kappa \max \left(1-\mathbf{z}_{i}^{\prime} \mathbf{\beta}, 0\right)}=\int_{0}^{\infty} \frac{\sqrt{\kappa}}{\sqrt{2 \pi \rho_{i}}} e^{-\frac{\kappa\left(\rho_{i}+1-\mathbf{z}_{i}^{\prime} \boldsymbol{\beta}\right)^{2}}{2 \rho_{i}}} d \rho_{i} \tag{4}
\end{equation*}
$$

Note that (4) makes the conditional distribution of $\beta$ become the multivariate Gaussian distribution, which leads to a straightforward Gibbs sampler.

## A. Gibbs Sampling Algorithm

We sample ( $\boldsymbol{\kappa}, \boldsymbol{\rho}$ ) jointly, by first sampling $\boldsymbol{\kappa}$ with $\rho$ marginalized out and then sampling $\boldsymbol{\rho}$ conditioning on $\mathcal{\kappa}$ (and other parameters). The conditional distribution of $\kappa$ is given by

$$
\kappa \mid \beta, Z \sim \mathcal{G}\left(a_{k}+n, b_{k}+2 \sum_{i} \max \left(1-\mathbf{z}_{i}^{\prime} \beta, 0\right)\right) .
$$

The conditional distribution of $\rho_{i}$ is given by

$$
\rho_{i} \mid \boldsymbol{\beta}, \mathbf{z}_{i}, \kappa \sim \mathscr{G} \mathcal{F} \mathcal{N}\left(1 / 2, \kappa, \kappa\left(1-\mathbf{z}_{i}^{\prime} \boldsymbol{\beta}\right)^{2}\right),
$$

where $\mathscr{G \mathscr { F }} \mathscr{N}(p, a, b)$ stands for the generalized inverse Gaussian distribution. Alternatively, the conditional distribution of $\rho_{i}^{-1}$ given $\left(\boldsymbol{\beta}, \mathbf{z}_{i}, \boldsymbol{\kappa}\right)$ is an inverse Gaussian distribution, denoted by $\mathscr{F} \mathcal{N}$.

$$
\rho_{i}^{-1} \mid \boldsymbol{\beta}, \mathbf{z}_{i}, \kappa \sim \mathcal{F} \mathcal{N}\left(\left|1-\mathbf{z}_{i}^{\prime} \boldsymbol{\beta}\right|^{-1}{ }_{,} \kappa\right),
$$

where the density function of $\mathscr{F} \mathscr{N}(\mu, \lambda)$ is given by

$$
f(x ; \mu, \lambda)=\frac{\sqrt{\lambda}}{\sqrt{2 \pi x^{3}}} e^{\frac{-\lambda(x-\mu)^{2}}{2 \mu^{2} x}} .
$$

The conditional distribution of $\gamma_{j}$ is given by

$$
p\left(\gamma_{j} \mid \beta_{j}, \gamma_{-j}\right) \propto v_{j}^{-1 / 2} e^{-\frac{\beta_{j}^{2}}{2 v_{j}}-\mu \gamma_{j}+\eta \sum_{k} G_{j k} I\left(\gamma_{j}=\gamma_{k}\right)}
$$

where $\gamma_{-j}=\left(\gamma_{1}, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_{p+1}\right)$.
Finally, let 1 be the vector of 1 's, $D_{\rho}=\operatorname{diag}\left(\rho_{1}, \ldots, \rho_{n}\right)$, and $D_{v}=\operatorname{diag}\left(v_{1}, \ldots, v_{p+1}\right)$. The conditional distribution of $\boldsymbol{\beta}$ follows a multivariate Gaussian:

$$
\beta \mid Z, \kappa, \boldsymbol{\rho} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\beta}, \boldsymbol{\Sigma}_{\beta}\right),
$$

where $\boldsymbol{\mu}_{\beta}=\kappa\left(D_{v}^{-1}+\kappa Z^{\prime} D_{\rho}^{-1} Z\right)^{-1} Z^{\prime} D_{\rho}^{-1}(1+\boldsymbol{\rho})$ and $\boldsymbol{\Sigma}_{\beta}=\left(D_{v}^{-1}+\kappa Z^{\prime} D_{\rho}^{-1} Z\right)^{-1}$.

Algorithm 1 summarizes the Gibbs sampling algorithm for KBSVM. Note that the most time consuming step is to sample $\beta$, which requires an $O\left(p^{3}\right)$ operation. For the probit model simulation in Section IV-D, which includes 500 predictors, it takes less than an hour to run 10,000 MCMC iterations in MATLAB with a 2.6 GHz Intel Core with 24 GB of RAM on 64-bit Windows 10 . When $p$ is large, one can consider sampling each individual $\beta_{j}$ one at a time. It requires an $O(p \min (n ; p))$ operation.

```
for \(t=1\) to \(T\) do
    Sample \(\kappa \propto \mathcal{G}\left(a_{\kappa}+n, b_{\kappa}+2 \sum_{i} \max \left(1-\mathbf{z}_{i} \boldsymbol{\beta}, 0\right)\right)\).
    for \(i=1\) to \(n\) do
            Sample \(\rho_{i}^{-1} \propto \mathcal{I N}\left(\left|1-\mathbf{z}_{i} \boldsymbol{\beta}\right|^{-1}, \kappa\right)\)
    end
    for \(j=1\) to \(p+1\) do
        Sample \(\gamma_{j}\) from \(\pi\left(\gamma_{j} \mid \beta_{j}, \gamma_{-j}\right) \propto v_{j}^{-1 / 2}\)
        \(\exp \left(-\frac{\beta_{j}^{2}}{2 v_{j}}-\mu \gamma_{j}+\eta \sum_{k} G_{j k} I\left(\gamma_{j}=\gamma_{k}\right)\right)\)
    end
    Sample \(\boldsymbol{\beta} \propto \mathcal{N}\left(\kappa\left(D_{v}^{-1}+\kappa Z^{\prime} D_{\rho}^{-1} Z\right)^{-1} Z^{\prime} D_{\rho}^{-1}(\mathbf{1}+\right.\)
        \(\left.\rho),\left(D_{v}^{-1}+\kappa Z^{\prime} D_{\rho}^{-1} Z\right)^{-1}\right)\)
0 end
```

Algorithm 1: Gibbs sampling algorithm for KBSVM

## IV. Simulations

## A. Design of Experiment

We use both the linear discrimination analysis (LDA) model and the probit model to generate correlated data to evaluate the performance of our KBSVM method and make comparisons with other existing methods such as the standard SVM ( $L_{2}$ SVM $), L_{1}$ SVM, DrSVM and SCADSVM. We generate $m=100$ datasets, each with a training sample of size $n=200$, a validation sample of size $n=200$ and an independent test sample of size $n=$ 10000. We specify different combinations of the feature dimension $p$ and the nonzero feature dimension $q$ for different models. To assess the performance of the predictive model, we compute the prediction error (PE), prediction sensitivity (PSEN), prediction specificity (PSPEC), Matthews Correlation Coefficients (MCC), feature selection true positive (FSTP) and feature selection false positive (FSFP) averaged across the $m=100$ datasets. The approach for obtaining PE is described in the following section. PSEN is calculated as the proportion of positives $\left(y_{i}=1\right)$ that are correctly identified and PSPEC is calculated as the proportion of negatives $\left(y_{i}=-1\right)$ that are correctly identified. MCC is defined as $\frac{T P \times T N-F P \times F N}{\sqrt{(T P+F P)(T P+F N)(T N+F P)(T N+F N)}}$, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. FSTP and FSFP are the average number of selected relevant and irrelevant features in the training samples.

## B. Parameter Tuning

For each of the existing methods, we use the penalizedSVM R-package [30] to fit the model on the training datasets, tune the parameters in the validation datasets and obtain the results from the testing datasets. $\sigma_{1}^{2}$ is set to 100 to account for large variances for the slab. $\eta$ is set to 1 or 0 , to account for the prior knowledge used or not. $\sigma_{0}^{2}, \mu$ need to be tuned to achieve the best performance. To tune the parameters $\sigma_{0}^{2}$ and $\mu$, we apply our algorithm on each training data and draw 1000 samples from the joint posterior distribution of $\boldsymbol{\beta}$ and $\gamma$. Each sample of $\beta$ and the corresponding $\gamma$ values are plugged into the model to make predictions on the validation sample. If $\gamma_{j}=1$, the corresponding $\boldsymbol{\beta}_{j}$ is selected. If $\gamma_{j}=0$, the corresponding $\boldsymbol{\beta} j$
is set to zero. Then the prediction can be obtained by $\hat{y}=\operatorname{sign}(X \beta)$, where $X$ is the observation matrix of the validation sample. PE can be calculated as the number of non-zero elements of $(y-\hat{y})$ divided by the number of observations of the validation sample ( $n=200$ ). Then the averaged PE across the 1000 posterior samples will be acquired and used for choosing the optimal parameters, and the corresponding 1000 samples are plugged into the model again to make predictions on the independent test sample. We repeat this procedure on the $m=100$ datasets to obtain the average PE and the corresponding standard errors.

## C. Simulation I: LDA model in the absence of the graph

The LDA model is used to evaluate the prediction and variable selection performance of our KBSVM method without incorporating the prior graph information. We adapt the same setting of $(\rho=-0.2, p=400, q=5)$ as in [31]. The similar results for the existing methods such as $L_{2}$ SVM, $L_{1}$ SVM and SCADSVM are obtained. Furthermore, the cases for $\rho=0$ and 0.2 is also included to investigate the impact of different correlation structure of $X$ on the performance of our method and other methods.

Model: $P(\mathbf{y}= \pm 1)=0.5, X \mid \mathbf{y} \sim \mathcal{N}(\operatorname{sign}(\mathrm{y}) \boldsymbol{\mu}, \Sigma), \boldsymbol{\mu}=(0.1,0.2,0.3,0.4,0.5,0, \ldots, 0)$ and

$$
\Sigma=\left(\begin{array}{lll}
\left(\begin{array}{lll}
1 & & \rho \\
& \ddots & \\
\rho & & 1
\end{array}\right)_{q \times q}, \\
& 0 & \mathbf{I}
\end{array}\right)_{p \times p}
$$

where $\rho= \pm 0.2$ or $0, q=5$ and $p=400$.
Table I compares different methods for the LDA model with the negative correlation, independent or positive correlation between genes. The numbers in the parentheses are the corresponding standard errors over the 50 datasets. It is not surprising to see that the performance deteriorates when $\rho$ increases from -0.2 to 0.2 for all the methods, because in general, the variance of $\beta$ is proportional to the inverse of the covariance matrix $\Sigma$. When $\rho=$ -0.2 ,

$$
\left.\mathbf{\Sigma}^{-1}=\left(\begin{array}{lll}
1.67 & & 0.83 \\
& \ddots & \\
0.83 & 1.67
\end{array}\right)_{5 \times 5} \quad \mathbf{0} \mathbf{O}_{5} \quad \mathbf{I}\right)_{400 \times 400}
$$

and when $\rho=0.2$,

$$
\Sigma^{-1}=\left(\begin{array}{lll}
\left(\begin{array}{lll}
1.11 & & -0.14 \\
& \ddots & \\
-0.14 & & 1.11
\end{array}\right)_{5 \times 5} & \mathbf{0} \\
& \mathbf{0} & \mathbf{I}
\end{array}\right)_{400 \times 400}
$$

Therefore, $\beta$ learned from the training set with positive correlation will have smaller variance and may not be particularly stable when making predictions for the testing set.

When $\rho=-0.2$, DrSVM has similar performance as $L_{2}$ SVM and also a very high FSTP because it tends to select more variables. SCADSM and KBSVM achieve significantly lower PE and greater MCC, which may be due to the negative correlation structure, while our method KBSVM has the least PE, largest MCC and highest FSTP. When $\rho=0$, genes in $X$ are independent, DrSVM still has the highest FSTP, as well as the highest FSFP. PE for SCADSVM and KBSVM are close, while KBSVM has significantly lower FSFP than the other methods. When $\rho=0.2, \mathrm{PE}$ and MCC for $L_{1}$ SVM, SCADSVM and KBSVM are similar, while $L_{1}$ SVM has the highest FSTP, SCADSVM has the highest FSFP and KBSVM has the moderate FSTP and the lowest FSFP. In sum, Our KBSVM method outperforms the presented methods in terms of PE, PSEN, MCC and FSFP. Even without the guidance of prior knowledge, the performance of our method doesn't degrade.

## D. Simulation II: Probit model in the presence of the graph

This section is to illustrate how to model the prior structure information and how to incorporate it in our method.

1) Graph simulation: Note that the true correlation structure of the genes is unknown in practice. As mentioned, we use the undirected graph $\mathscr{G}$ to represent the relationship between genes. In our simulation, we distinguish the underlying true graph $\mathscr{G}$ which is used for generating the data, and the working graph $\mathscr{G} *$ which is providing the guidance to KBSVM algorithms.

In our simulation examples, the true graph $\mathscr{G}$ is predefined. Let
$X=\left(\mathrm{x}_{1}, \mathrm{x}_{2}, \ldots, \mathrm{x}_{p}\right) \sim \mathcal{N}\left(0, \Omega^{-1}\right)$, where the precision matrix $\Omega=\left(\omega_{i j}\right)$ is such that $(i, j)-\notin E$ implies $\omega_{i j}=0$. We then say that $X$ follows a Gaussian graphical model (GMM) with respect to the graph $\mathscr{G}$. In order to convert the graph $\mathscr{G}$ to the precision matrix $\Omega$, the Gaussian graphical model is adopted and several steps are performed. First, a matrix is created by assigning uniformly distributed random numbers over an interval of $[-1,1]$ to the off diagonal elements corresponding to the edges in the graph $\mathscr{E}$; second, the absolute value of the lowest eigen-value of the resulting matrix in the first step is obtained and added to a small positive number, denoted as $|\lambda|+\Delta$; third, the elements on the diagonal of the matrix are reset to $|\lambda|+\Delta$, and therefore, all the eigenvalues of the resulting matrix are positive. Then the precision matrix can be obtained through scaling the resulting matrix by making the diagonal elements equal to 1's. Correspondingly, the covariance matrix $\Sigma$ can be obtained by normalizing the inverse of the precision matrix. An example of the three matrices are illustrated in Figure 1(b, c, d).

The working graph $\mathscr{G} *$ represents the prior knowledge we now have to incorporate into our algorithm, thus it could be the true graph $\mathscr{G}$ indicating that the truth is known, a partial graph indicating that the truth is partially known or a noisy graph indicating that the prior knowledge is wrong. To simulation the partial graph, we adopt the Gaussian graphical model and set a threshold value on the precision matrix to remove some weak correlations. We first define a threshold value $t$, then compare the absolute values of each element of the precision matrix to $t$ : if less than $t$, the element is set to zero; if equal or greater than $t$, the element remains the same value. Then the adjacency matrix of the partial graph is acquired by setting
all the off-diagonal nonzero values of the resulting matrix to 1 's, indicating the connection between nodes, while setting the diagonal elements to 0's. A pair of the adjacency matrices of the true graph and partial graph are shown in Figure 2.

To simulate the noisy graph, we can directly work on the lower triangle part of the corresponding adjacency matrix. First, we create a dimension $0_{(p+1) \times(p+1)}$ matrix, define a maximum number of connections $n$ and generate a uniformly distributed random integer $k$ over the interval of $[0, n]$. Second, we count the total number of the elements of the lower triangle part without including the diagonal elements, denoted as $m$, then generate $m$ standard uniformly distributed random numbers and sort them. Third, the first $k$ elements in the ordered $m$ samples are assigned 1 's and the left elements are assigned 0 's. Then we apply some transformations to create a symmetric adjacency matrix from the lower part.
2) Probit model: The probit model is used to demonstrate the benefits of incorporating prior knowledge into our KBSVM method. The model can be written as : $X \sim \mathcal{N}(0, \Sigma)$, $\Sigma=f(\mathscr{G}), P(\mathbf{y}=1 X)=\Phi\left(X \boldsymbol{\beta}+\beta_{0}\right) . \mathscr{G}$ is the true underlying true structure among predictors. The covariance structure of $\Sigma$ should have a similar pattern to the adjacency matrix of $\mathscr{G}$, in other words, a function of $\mathscr{G} . \Phi$ is the CDF of the standard normal distribution. $\beta_{0}$ is the intercept set to 0.5 and $\boldsymbol{\beta}=(0.8,0.8, . ., 0.8,0.8,0,0)$ is the $p$ dimension coefficient with the first $q$ non-zero elements.

We specify four settings for our model and compare them to $L_{2}$ SVM, $L_{1}$ SVM, DrSVM and SCADSVM. The four settings are: no working graph incorporated ( $\eta=0$ ), the working graph $\mathscr{G} *$ is assigned by a noisy graph (noisy $\mathscr{G}$ ), a partial graph (partial $\mathscr{G}$ ) and the true underlying graph $(\mathscr{G})$. Table 2 summarizes the simulation results for both $n>p$ and $n<p$ cases. Clearly, for all the cases, when the working graph $\mathscr{G} *$ is assigned by the true graph $\mathscr{G}$, our model KBSVM performs the best among the other settings as well as other existing methods. When $p=20$ and $q=10, L_{2}$ SVM gives the largest PE and the lowest MCC, the prediction performance for $L_{1}$ SVM, DrSVM, SCADSVM, KBSVM $(\eta=0)$ and KBSVM $(\mathscr{G} *=$ noisy $\mathscr{G})$ are similar, while $L_{1}$ SVM has a very high FSFP, and tends to select a larger model. When $p=100$ and $q=20, \operatorname{PE}$ for $\operatorname{KBSVM}(\mathscr{G} *=\mathscr{G})$ is significantly decreasing comparing to the other settings and other existing methods. When $\eta=0$, the performance is the worst, among the four settings, but still outperforms $L_{2}$ SVM, DrSVM and SCADSVM. We also note that $L_{1}$ SVM still has the highest FSFP, and DrSVM has the second highest FSFP, which case is a little different from the case with $p=20$. When $p=500$, the prediction errors of $L_{2} \mathrm{SVM}$ and DrSVM are similar, $L_{1}$ SVM and $\operatorname{KBSVM}(\mathscr{G} *=$ partial $\mathscr{G})$ are similar, while $L_{1}$ SVM has the much higher high FSTP and FSFP. SCADSVM and KBSVM $\mathscr{G} *=\mathscr{G}$ achieve the best results in terms of PE. In general, our method gives the smallest PE, the greatest MCC, a very low FSFP and BS. Even when $\mathscr{G} *$ is assigned by noisy $\mathscr{G}$, the performance of our method doesn't deteriorate too much.

In addition, we generate a new set of data from the independent correlation structure and thus we only need specify two settings for our model: $\eta=0$ and $\mathscr{G} *=$ noisy $\mathscr{G}$. The results are summarized in Table III. When $p=20$ and $100, \operatorname{KBSVM}(\eta=0)$ outperforms the other methods in terms of PE and MCC. $L_{1}$ SVM, DrSVM, SCADSVM tend to select more
variables with a very high FSFP. Both of two settings for KBSVM give a significantly lower FSFP but keep the relatively high FSTP, showing the consistent ability of feature selection. When $p=500, L_{1}$ SVM gives the best performance in terms of PE, MCC and FSTP, while our model with $\eta=0$ achieves satisfactory performance and also agrees with the findings in the LDA model.

In this simulation section, we consider two models under two conditions which are absence of the graph and presence of the graph. We observe that if the graphical network information is associate with the outcome and we utilize the true network information in the model, our KBSVM model outperforms other methods in terms of both prediction and selection accuracy. If the prior graph is not available, the performance doesn't degrade. Such stability is desirable and the results demonstrate encouraging gene selection ability and prediction power for our method.

## V. Data Analysis

In this section, we apply our methods as well as other existing methods to classify a glioblastoma data set obtained from the Cancer Genome Atlas Network. Glioblastoma is a highly malignant brain tumor, also related to other cancer. This data set includes survival times ( Y ) and the gene expression levels for $p=12,999$ genes ( X ) and 303 glioblastoma patients. For the purpose of classification, we define a new indicator variable Z to account for the one year survival outcome by setting

$$
Z=\left\{\begin{array}{l}
1, Y<365, \Delta=0, \\
0, Y>365,
\end{array}\right.
$$

where $\Delta$ represents censoring. Those subjects with $Y<365, \Delta=1$ are removed so the total number of subjects is 286 with $P(Z=1)=45 \%, P(Z=0)=55 \%$. First, we use the generanking methods to select important genes. For each gene, the p value is acquired from the logistic regression and the top 1000 genes corresponding to the smallest 1000 p values are selected. Second, we obtain the network $\mathscr{G}$ for all the 12, 999 genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, use an algorithm to search the connections within the top 1000 genes, and then map them to the working graph $\mathscr{G} *$. We specify two settings ( $\eta=0$ and 1 ) for our model to compare with other methods. The optimal tuning parameters for each methods are chosen by the minimum 20-fold crossvalidation error. The average cross-validation error and the number of selected genes are summarized in Table IV.

As can be seen, $L_{1}$ SVM selects most of the 1000 genes and has a similar performance to $L_{2}$ SVM. DrSVM and SCADSVM give the very close CV errors while DrSVM select fewer number of genes. Our method KBSVM $(\eta=1)$ achieves the lowest CV error and BS and identifies a moderate number of genes. $\operatorname{KBSVM}(\eta=0)$ imposes more sparsity on the model and select only 69 genes, yet provides the satisfactory cross-validation error. In addition, all the genes selected by KBSVM $(\eta=0)$ are contained in the set of genes selected by KBSVM ( $\eta=1$ ), which confirms the stability.

We also conduct the pathway enrichment analysis for the selected genes for our method via ToppGene Suite [32]. When $\eta=0$, our method doesn't encourage the inclusion of the connected genes, therefore, fewer genes and pathways are detected. However, several important genes are still selected, such as PICK1, IL22, BHLHE40 and NTN1, which are the members of the glioma pathways. When $\eta=1$, the pathways detected by our method are highly enriched, such as protein processing in endoplasmic reticulum $\left(1.16 \times 10^{-6}\right)$, asparagine N -linked glycosylation $\left(6.69 \times 10^{-3}\right.$ ), ATF6 (ATF6-alpha) activates chaperone genes $\left(7.86 \times 10^{-3}\right)$, and unfolded protein response $\left(1.08 \times 10^{-2}\right)$. The numbers in the parentheses are the Bonferroni-adjusted $p$ values. These pathways were found to be linked with the cancer cell proliferation and survival [33]-[36]. Moreover, the most highly enriched diseases are glioblastoma, mammary neoplasms and malignant tumor of colon. Therefore, the detected pathways and diseases further confirm our method can offer great promises of improved power in detection of key molecular signatures and provide valuable insights on biological bases of diseases.

In sum, for our method KBSVM, when the prior network incorporated, the cross-validation error is reduced and the related pathways are significantly enriched, yielding biologically meaningful results. Therefore, we believe that our method KBSVM enjoys the benefits of incorporating prior knowledge to improve predictive performance.

## VI. Discussion

In this article we have developed a knowledge-guided Bayesian SVM approach, which enables feature selection and incorporation of the prior structural information. The numerical results confirm the performance of our method in terms of the improved prediction and variable selection accuracy. Our method yields significant performance when the working graph is correctly specified, and is fairly robust when the working graph is misspecified. One limitation of our model is that it can be computationally expensive to tune the hyper-parameters, especially in high-dimensional settings. Future work may extend our approach to the non-linear Bayesian SVM model.

## Acknowledgment

This work is partly supported by NIH grants P30CA016520 and R01GM124111. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## REFERENCES

[1]. Vapnik VN and Vapnik V, Statistical learning theory Wiley New York, 1998, vol. 1.
[2]. Salcedo-Sanz S, Rojo-Á lvarez JL, Martínez-Ramón M, and Camps-Valls G, "Support vector machines in engineering: an overview," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 3, pp. 234-267, 2014.
[3]. Nayak J, Naik B, and Behera H, "A comprehensive survey on support vector machine in data mining tasks: applications \& challenges," International Journal of Database Theory and Application, vol. 8, no. 1, pp. 169-186, 2015.
[4]. Bradley PS and Mangasarian OL, "Feature selection via concave minimization and support vector machines." in ICML, vol. 98, 1998, pp. 82-90.
[5]. Song M, Breneman CM, Bi J, Sukumar N, Bennett KP, Cramer S, and Tugcu N, "Prediction of protein retention times in anion-exchange chromatography systems using support vector regression," Journal of chemical information and computer sciences, vol. 42, no. 6, pp. 13471357, 2002. [PubMed: 12444731]
[6]. Zhu J, Rosset S, Tibshirani R, and Hastie TJ, "1-norm support vector machines," in Advances in neural information processing systems, 2004, pp. 49-56.
[7]. Tibshirani R, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B (Methodological), pp. 267-288, 1996.
[8]. Zhang HH, Ahn J, Lin X, and Park C, "Gene selection using support vector machines with nonconvex penalty," bioinformatics, vol. 22, no. 1, pp. 88-95, 2005. [PubMed: 16249260]
[9]. Becker N, Toedt G, Lichter P, and Benner A, "Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data," BMC bioinformatics, vol. 12, no. 1, p. 138, 2011. [PubMed: 21554689]
[10]. Fan J, "Runze li variable selection via penalized likelihood," J. Amer. Stat. Assoc, 2001.
[11]. Wang L, Zhu J, and Zou H, "The doubly regularized support vector machine," Statistica Sinica, pp. 589-615, 2006.
[12]. Zou H and Yuan M, "The f-norm support vector machine," Statistica Sinica, pp. 379-398, 2008.
[13]. Lauer F and Bloch G, "Incorporating prior knowledge in support vector machines for classification: A review," Neuro-computing, vol. 71, no. 7, pp. 1578-1594, 2008.
[14]. S. C. F., "Pathway databases," Annals of the New York Academy of Sciences, vol. 1020, no. 1, pp. 77-91.
[15]. Pan W, Xie B, and Shen X, "Incorporating predictor network in penalized regression with application to microarray data," Biometrics, vol. 66, no. 2, pp. 474-484, 2010. [PubMed: 19645699]
[16]. Stingo FC, Chen YA, Tadesse MG, and Vannucci M, "Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes," The annals of applied statistics, vol. 5, no. 3, 2011.
[17]. Zhao Y, Chung M, Johnson BA, Moreno CS, and Long Q, "Hierarchical feature selection incorporating known and novel biological information: Identifying genomic features related to prostate cancer recurrence," Journal of the American Statistical Association, vol. 111, no. 516, pp. 1427-1439, 2016. [PubMed: 28435175]
[18]. Chang C, Kundu S, and Long Q, "Scalable bayesian variable selection for structured highdimensional data," Biometrics, in press, 2018.
[19]. Chuang H-Y, Lee E, Liu Y-T, Lee D, and Ideker T, "Network-based classification of breast cancer metastasis," Molecular systems biology, vol. 3, no. 1, p. 140, 2007. [PubMed: 17940530]
[20]. Zhu Y, Shen X, and Pan W, "Network-based support vector machine for classification of microarray samples," BMC bioinformatics, vol. 10, no. 1, p. S21, 2009.
[21]. Luts J and Ormerod JT, "Mean field variational bayesian inference for support vector machine classification," Computational Statistics \& Data Analysis, vol. 73, pp. 163-176, 2014.
[22]. Bhosale D and Ade R, "Feature selection based classification using naive bayes, j48 and support vector machine," International Journal of Computer Applications (0975-8887) Volume, vol. 99, 2014.
[23]. Yang X, Pan W, and Guo Y, "Sparse bayesian classification and feature selection for biological expression data with high correlations," PloS one, vol. 12, no. 12, p. e0189541, 2017. [PubMed: 29281700]
[24]. George EI and McCulloch RE, "Variable selection via gibbs sampling," Journal of the American Statistical Association, vol. 88, no. 423, pp. 881-889, 1993.
[25]. Ising E, "Beitrag zur theorie des ferromagnetismus," Zeitschrift für Physik, vol. 31, no. 1, pp. 253-258, 1925
[26]. Gilks WR, Richardson S, and Spiegelhalter DJ, "Introducing markov chain monte carlo," Markov chain Monte Carlo in practice, vol. 1, p. 19, 1996.
[27]. Polson NG, Scott SL et al., "Data augmentation for support vector machines," Bayesian Analysis, vol. 6, no. 1, pp. 1-23, 2011. [PubMed: 22247752]
[28]. Henao R, Yuan X, and Carin L, "Bayesian nonlinear support vector machines and discriminative factor modeling," in Advances in Neural Information Processing Systems, 2014, pp. 1754-1762.
[29]. Li F and Zhang NR, "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics," Journal of the American statistical association, vol. 105, no. 491, pp. 1202-1214, 2010.
[30]. Becker N, Werft W, Toedt G, Lichter P, and Benner A, "penalizedsvm: a r-package for feature selection svm classification," Bioinformatics, vol. 25, no. 13, pp. 1711-1712, 2009. [PubMed: 19398451]
[31]. Zhang X, Wu Y, Wang L, and Li R, "Variable selection for support vector machines in high dimensions"
[32]. Chen J, Bardes EE, Aronow BJ, and Jegga AG, "Toppgene suite for gene list enrichment analysis and candidate gene prioritization," Nucleic acids research, vol. 37, no. suppl 2, pp. W305-W311, 2009. [PubMed: 19465376]
[33]. Clarke HJ, Chambers JE, Liniker E, and Marciniak SJ, "Endoplasmic reticulum stress in malignancy," Cancer cell, vol. 25, no. 5, pp. 563-573, 2014. [PubMed: 24823636]
[34]. Kurtoglu M, Gao N, Shang J, Maher JC, Lehrman MA, Wangpaichitr M, Savaraj N, Lane AN, and Lampidis TJ, "Under normoxia, 2-deoxy-d-glucose elicits cell death in select tumor types not by inhibition of glycolysis but by interfering with n-linked glycosylation," Molecular cancer therapeutics, vol. 6, no. 11, pp. 3049-3058, 2007. [PubMed: 18025288]
[35]. Grantham NS, Reich BJ, Borer ET, and Gross K, "Mimix: a bayesian mixed-effects model for microbiome data from designed experiments," arXiv preprint arXiv:1703.07747, 2017.
[36]. Hiramatsu N, Joseph VT, and Lin JH, "Monitoring and manipulating mammalian unfolded protein response," in Methods in enzymology Elsevier, 2011, vol. 491, pp. 183-198. [PubMed: 21329801]
(a) True graph with four networks




Figure 1.
The true graph $\mathscr{G}$ and its corresponding adjacency matrix $G$, precision matrix $\Omega$ and covariance matrix $\Sigma$.


Figure 2.
The simulation steps of the partial graph $G^{*}$.
Author Manuscript
Author Manuscript
Author Manuscript
Table II
COMPARISON OF THE PREDICTION PERFORMANCE AND VARIABLE SELECTION WHEN THE DIMENSION OF PREDICTIONS $p$ CHANGES FROM 20 TO 500 AMONG DIFFERENT METHODS. $q$ IS THE NUMBER OF RELEVANT VARIABLES. $\eta=0$ REPRESENTS THE WORKING GRAPH $G^{*}$ IS NOT INCORPORATED IN OUR KBSVM mOdel.

| Method | PE (\%) | PSen (\%) | PSpec (\%) | MCC (\%) | FSTP (\%) | FSFP (\%) |
| :---: | :---: | :---: | :---: | :---: | :---: | ---: |
|  | $p=20, q=10$ |  |  |  |  |  |
| L2SVM | $14.10(0.11)$ | $89.97(0.31)$ | $80.75(0.47)$ | $71.52(0.22)$ | - | - |
| L1SVM | $11.82(0.09)$ | $90.09(00.22)$ | $85.75(0.30)$ | $76.06(0.19)$ | $98.05(0.51)$ | $52.30(2.51)$ |
| DrSVM | $11.92(0.08)$ | $89.69(0.23)$ | $86.04(0.26)$ | $75.88(0.16)$ | $99.80(0.14)$ | $18.20(1.33)$ |
| SCADSVM | $11.84(0.11)$ | $89.63(0.21)$ | $86.30(0.28)$ | $76.04(0.23)$ | $98.10(0.42)$ | $28.60(3.08)$ |
| KBSVM, $\boldsymbol{\eta}=0$ | $11.92(0.11)$ | $89.83(0.22)$ | $85.86(0.30)$ | $75.87(0.22)$ | $96.23(0.62)$ | $16.44(2.22)$ |
| KBSVM, $\mathscr{G} *=$ noisy $\mathscr{G}$ | $12.02(0.11)$ | $89.59(0.24)$ | $85.94(0.29)$ | $75.69(0.22)$ | $97.20(0.45)$ | $25.93(2.56)$ |
| KBSVM, $\mathscr{G} *=$ partialG $\mathscr{G}$ | $11.59(0.11)$ | $90.00(0.20)$ | $86.41(0.30)$ | $76.56(0.22)$ | $98.36(0.48)$ | $12.94(2.06)$ |
| KBSVM, $\mathscr{G} *=\mathscr{G}$ | $11.55(0.11)$ | $90.00(0.22)$ | $86.48(0.30)$ | $76.64(0.21)$ | $98.56(0.41)$ | $11.00(1.91)$ |


|  | $p=100, q=20$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | ---: |
| L2SVM | $20.96(0.11)$ | $83.23(0.44)$ | $73.87(0.62)$ | $57.77(0.27)$ | - | - |
| L1SVM | $17.27(0.19)$ | $84.93(0.30)$ | $80.03(0.46)$ | $65.18(0.39)$ | $90.41(0.83)$ | $40.57(1.66)$ |
| DrSVM | $19.74(0.15)$ | $83.13(0.31)$ | $76.73(0.46)$ | $60.16(0.30)$ | $84.40(2.13)$ | $28.43(2.12)$ |
| SCADSVM | $18.18(0.27)$ | $83.89(0.34)$ | $79.28(0.51)$ | $63.35(0.54)$ | $73.65(1.46)$ | $9.85(1.63)$ |
| KBSVM, $\boldsymbol{\eta = 0}$ | $17.92(0.29)$ | $83.67(0.33)$ | $80.13(0.42)$ | $63.85(0.58)$ | $78.71(1.16)$ | $8.10(0.56)$ |
| KBSVM, $\mathscr{G} *=$ noisy $\mathscr{G}$ | $17.29(0.25)$ | $84.30(0.30)$ | $80.76(0.42)$ | $65.15(0.50)$ | $79.11(1.12)$ | $9.63(0.66)$ |
| KBSVM, $\mathscr{G} *=$ partial $\mathscr{G}$ | $15.76(0.24)$ | $85.67(0.27)$ | $82.51(0.42)$ | $68.25(0.49)$ | $87.83(0.91)$ | $7.19(0.46)$ |
| KBSVM, $\mathscr{G} *=\mathscr{G}$ | $14.40(0.11)$ | $87.08(0.27)$ | $83.79(0.35)$ | $70.97(0.41)$ | $96.66(0.55)$ | $6.69(0.41)$ |


|  | $p=500, q=20$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | ---: |
| L2SVM | $33.61(0.29)$ | $75.02(1.92)$ | $56.96(2.26)$ | $33.44(0.45)$ | - | - |
| L1SVM | $24.34(0.46)$ | $79.39(0.97)$ | $71.07(0.87)$ | $50.87(0.91)$ | $67.59(1.47)$ | $9.65(0.85)$ |
| DrSVM | $32.26(0.24)$ | $75.19(1.28)$ | $58.57(1.26)$ | $34.75(0.47)$ | $31.67(5.06)$ | $1.56(0.36)$ |
| SCADSVM | $24.16(0.56)$ | $77.48(0.77)$ | $73.83(1.02)$ | $51.36(1.12)$ | $48.00(2.38)$ | $1.38(0.12)$ |
| KBSVM, $\boldsymbol{\eta}=0$ | $24.87(0.54)$ | $76.89(0.80)$ | $72.97(1.00)$ | $49.96(1.08)$ | $45.64(2.53)$ | $2.06(0.59)$ |
| KBSVM, $\mathscr{G} *=$ noisy $\mathscr{G}$ | $24.67(0.52)$ | $77.30(0.66)$ | $72.90(0.91)$ | $50.28(1.05)$ | $42.86(2.04)$ | $1.23(0.39)$ |
| KBSVM, $\mathscr{G} *=$ partial $\mathscr{G}$ | $24.32(0.51)$ | $77.72(0.63)$ | $73.11(0.81)$ | $50.94(1.04)$ | $46.60(2.71)$ | $1.25(0.22)$ |

Sun et al.
Author Manuscript
COMPARISON OF THE PREDICTION PERFORMANCE AND VARIABLE SELECTION WHEN THE PREDICTORS ARE INDEPENDENT.

| Method | PE (\%) | PSen (\%) | PSpec (\%) | MCC (\%) | FSTP (\%) | FSFP (\%) |
| :---: | :---: | :---: | :---: | :---: | :---: | ---: |
|  | $p=20, q=10$ |  |  |  |  |  |
| L2SVM | $16.74(0.20)$ | $87.76(0.55)$ | $77.22(0.86)$ | $65.90(0.39)$ | - | - |
| L1SVM | $14.13(0.13)$ | $88.40(0.37)$ | $82.45(0.45)$ | $71.15(0.25)$ | $100.00(0.00)$ | $56.98(3.12)$ |
| DrSVM | $14.31(0.12)$ | $88.20(0.37)$ | $82.33(0.48)$ | $70.82(0.25)$ | $100.00(0.00)$ | $45.80(3.47)$ |
| SCADSVM | $13.89(1.48)$ | $87.95(0.37)$ | $83.65(0.42)$ | $71.71(0.29)$ | $100.00(0 .(\mathrm{X}))$ | $25.80(4.51)$ |
| KBSMM, $\boldsymbol{\eta}=0$ | $13.67(0.12)$ | $88.03(0.34)$ | $84.03(0.39)$ | $72.15(0.24)$ | $99.95(0.03)$ | $10.48(2.27)$ |
| KBSVM, $\mathscr{G} *=$ noisy $\mathcal{G}$ | $13.90(0.14)$ | $87.91(0.34)$ | $83.68(0.42)$ | $71.69(0.29)$ | $99.88(0.08)$ | $15.84(2.97)$ |


|  | $p=100, q=20$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | ---: | :---: |
| L2SVM | $22.93(0.19)$ | $83.56(0.68)$ | $69.08(0.90)$ | $53.74(0.37)$ | - | - |  |
| L1SVM | $17.68(0.25)$ | $85.90(0.51)$ | $77.90(0.57)$ | $64.29(0.51)$ | $99.39(0.23)$ | $39.69(2.27)$ |  |
| DrSVM | $21.14(0.21)$ | $82.25(0.46)$ | $74.68(0.52)$ | $57.26(0.43)$ | $99.00(0.35)$ | $43.38(2.95)$ |  |
| SCADSVM | $19.58(0.58)$ | $82.66(0.60)$ | $77.65(0.87)$ | $60.46(1.17)$ | $89.50(1.68)$ | $25.75(5.08)$ |  |
| KBSVM, $\eta=0$ | $16.61(0.39)$ | $85.07(0.49)$ | $81.32(0.53)$ | $66.49(0.78)$ | $94.48(0.75)$ | $7.93(1.05)$ |  |
| KBSVM, $\mathscr{G} *=$ noisy $\mathscr{G}$ | $17.11(0.42)$ | $84.57(0.53)$ | $80.82(0.60)$ | $65.50(0.85)$ | $93.67(1.18)$ | $9.34(1.05)$ |  |
|  |  | $p=500, q=20$ |  | - |  |  |  |
| L2SVM | $36.16(0.27)$ | $81.90(1.09)$ | $41.55(1.82)$ | $26.27(0.53)$ |  | - |  |
| L1SVM | $26.33(0.68)$ | $78.36(0.68)$ | $67.86(0.72)$ | $46.65(0.99)$ | $88.75(1.30)$ | $18.49(0.27)$ |  |
| DrSVM | $35.43(0.17)$ | $74.23(0.90)$ | $52.67(1.21)$ | $27.90(0.35)$ | $43.70(4.51)$ | $9.10(0.94)$ |  |
| SCADSVM | $27.07(0.70)$ | $76.64(0.67)$ | $68.34(1.19)$ | $45.22(1.44)$ | $71.90(1.81)$ | $14.98(0.38)$ |  |
| KBSVM, $\eta=0$ | $26.90(0.61)$ | $76.07(0.63)$ | $69.43(0.85)$ | $45.68(1.27)$ | $64.03(1.77)$ | $13.33(0.37)$ |  |
| KBSVM, $\mathscr{G} *=$ noisy $\mathscr{G}$ | $27.95(0.59)$ | $74.80(0.66)$ | $68.67(0.81)$ | $43.55(1.20)$ | $59.11(2.16)$ | $12.31(0.45)$ |  |

Table IV
Results of the analysis of TCGA data. $n=286, p=1000$.

|  | CV error (\%) | \# selected genes |
| :--- | ---: | ---: |
| L2SVM | 30.45 | 1000 |
| L1SVM | 29.85 | 957 |
| DrSVM | 27.52 | 399 |
| SCADSVM | 27.31 | 864 |
| KBSVM, $\eta=0$ | 28.92 | 69 |
| KBSVM, $\eta=1$ | 26.49 | 821 |


[^0]:    § Corresponding author: qlong@pennmedicine.upenn.edu.
    ${ }^{\text {T}}$ Two authors equally contributed to this work.

