The Spike-and-Slab Lasso Generalized Linear Models for Prediction and Associated Genes Detection

Zaixiang Tang¹,²,³, Yueping Shen¹,², Xinyan Zhang³, Nengjun Yi³, *

¹ Department of Biostatistics, School of Public Health, Medical College of Soochow University, Suzhou 215123, China;
² Jiangsu Key Laboratory of Preventive and Translational Medicine for Geriatric Diseases, Medical College of Soochow University, Suzhou, 215123, China;
³ Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA
Running title: Spike-and-slab lasso GLMs

Key words: Caner, Double-exponential distribution, Generalized linear model, Outcome prediction, Spike-and-slab lasso

*Corresponding author:

Nengjun Yi
Department of Biostatistics
School of Public Health,
University of Alabama at Birmingham
Birmingham, AL 35294-0022
Phone: 205-934-4924
Fax: 205-975-2540
Email: nyi@uab.edu
Abstract

Large-scale omics data have been increasingly used as an important resource for prognostic prediction of diseases and detection of associated genes. However, there are considerable challenges in analyzing high-dimensional molecular data, including the large number of potential molecular predictors, limited number of samples, and small effect of each predictor. We propose new Bayesian hierarchical generalized linear models, called spike-and-slab lasso GLMs, for prognostic prediction and detection of associated genes using large-scale molecular data. The proposed model employs a spike-and-slab mixture double-exponential prior for coefficients that can induce weak shrinkage on large coefficients and strong shrinkage on irrelevant coefficients. We have developed a fast and stable algorithm to fit large-scale the hierarchal GLMs by incorporating EM (Expectation-Maximization) steps into the fast cyclic coordinate descent algorithm. The proposed approach integrates the nice features of the two popular methods, i.e., penalized lasso and Bayesian spike-and-slab variable selection. The performance of the proposed method is assessed via extensive simulation studies. The results show that the proposed approach can provide not only more accurate estimates of the parameters, but also better prediction. We demonstrate the proposed procedure on two cancer data sets: a well-known breast cancer data set consisting of 295 tumors and expression data of 4919 genes and the ovarian cancer data set from TCGA with 362 tumors and expression data of 5336 genes. Our analyses show that the proposed procedure can generate powerful models for predicting outcomes and detecting associated genes. The methods have been
implemented in a freely available R package BhGLM

(http://www.ssg.uab.edu/bhglm/).
Introduction

The growing recognition of precision medicine reflects the emergence of a field that is accelerating rapidly and will help shape new clinical practice in the future (Collins and Varmus 2015; Jameson and Longo 2015). The important base of precision medicine is to generate knowledge of disease that will enable better assessment of disease risk, understanding of disease mechanisms, and prediction of optimal therapy and prognostic outcome for diseases by using a wide range of biomedical, clinical, and environmental information. Precision medicine needs accurate detection of biomarkers and prognostic prediction (Chin et al. 2011; Barillot et al. 2012). The traditional clinical prognostic and predictive factors often provided poor prognosis and prediction (Barillot et al. 2012). Modern ‘omics’ technologies can robustly generate large-scale molecular data, such as large amount of genomic, transcriptomic, proteomic, and metabolomics data, which provides extraordinary opportunities to detect new biomarkers and to build more accurate prognostic and predictive models. However, these large-scale data sets also introduce computational and statistical challenges.

Various approaches have been applied in analyzing large-scale molecular profiling data to address the challenges (Bovelstad et al. 2007; Bovelstad et al. 2009; Lee et al. 2010; Barillot et al. 2012). The lasso and its extensions are the most commonly used methods (Tibshirani 1996; Hastie et al. 2009; Hastie et al. 2015). These methods put L1-penalty on the coefficients and can shrink many coefficients exactly to zero, thus performing variable selection. With the L1-penalty,
the penalized likelihood can be solved by extremely fast optimization algorithms, for
e.g., the lars and the cyclic coordinate descent algorithms (Efron et al. 2004; Friedman et al. 2010), making them very popular in high-dimensional data analysis.

Recently, these penalization approaches have been widely applied for prediction and

However, the lasso uses a single penalty for all coefficients and thus can either
include a number of irrelevant predictors or over-shrink large coefficients. An ideal
method should induce weak shrinkage on large effects and strong shrinkage on
irrelevant effects (Fan and Li 2001; Zou 2006; Zhang 2010). Statisticians have
introduced hierarchical models with mixture spike-and-slab priors that can adaptively
determine the amount of shrinkage (George and McCulloch 1993; George and
McCulloch 1997). The spike-and-slab prior is the fundamental basis for most
Bayesian variable selection approaches and has proved remarkably successful
(George and McCulloch 1993; Chipman 1996; George and McCulloch 1997;
Chipman et al. 2001; Ročková and George 2014; Ročková and George 2015).
Recently, Bayesian spike-and-slab priors have been applied to predictive modeling
and variable selection in large-scale genomic studies (Yi et al. 2003; Ishwaran and
Rao 2005; De los Campos et al. 2010; Zhou et al. 2013; Lu et al. 2015; Shankar et
al. 2015; Shelton et al. 2015; Partovi Nia and Ghannad-Rezaie 2016). However,
most previous spike-and-slab variable selection approaches use the mixture normal
priors on coefficients and employ Markov Chain Monte Carlo (MCMC) algorithms.
(e.g., stochastic search variable selection) to fit the model. Although statistically sophisticated, these MCMC methods are computationally intensive for analyzing large-scale molecular data. The mixture normal priors cannot shrink coefficients exactly to zero and thus cannot automatically perform variable selection. Ročková and George (2014) developed an EM algorithm to fit large-scale linear models with the mixture normal priors.

Ročková and George (2015) recently proposed a new framework, called the spike-and-slab lasso, for high-dimensional normal linear models by using a new prior on the coefficients, i.e., the spike-and-slab mixture double-exponential distribution. Ročková and George (2015) proved that the spike-and-slab lasso has remarkable theoretical and practical properties and overcomes some drawbacks of the previous approaches. However, the spike-and-slab lasso and most of the previous methods were developed based on normal linear models and cannot be directly applied to other models. Therefore, extensions of high-dimensional methods using mixture priors to frameworks beyond normal linear regression provide important new research directions for both methodological and applied works (Ročková and George 2014; Ročková and George 2015).

In this article, we extend the spike-and-slab lasso framework to generalize linear models, called the spike-and-slab lasso GLMs (sslasso GLMs), for jointly analyzing large-scale molecular data for building accurate predictive models and identifying important predictors. By using the mixture double-exponential priors, the spike-and-slab lasso GLMs can adaptively shrink coefficients (i.e., weakly shrink
important predictors but strongly shrink irrelevant predictors), and thus can result in accurate estimation and prediction. To fit the spike-and-slab lasso GLMs, we propose an efficient algorithm by incorporating EM steps (Expectation-Maximization) into the extremely fast cyclic coordinate descent algorithm. The performance of the proposed method is assessed via extensive simulations and compared with the commonly used lasso GLMs. We apply the proposed procedure to two cancer data sets with binary outcomes and thousands of molecular features. Our results show that the proposed method can generate powerful prognostic models for predicting disease outcome and also can detect associated genes.

Methods

Generalized linear models

We consider generalized linear models with a large number of correlated predictors. The observed values of a continuous or discrete response are denoted by \( y = (y_1, \cdots, y_n) \). The predictor variables include numerous molecular predictors (e.g., gene expression). A generalized linear model consists of three components: the linear predictor \( \eta \), the link function \( h \), and the data distribution \( p \) (McCullagh and Nelder 1989; Gelman et al. 2014). The linear predictor for the \( i \)-th individual can be expressed as

\[
\eta_i = \beta_0 + \sum_{j=1}^{J} x_{ij} \beta_j = X_i \beta
\]

where \( \beta_0 \) is the intercept, \( x_{ij} \) represents the observed value of the \( j \)-th variable, \( \beta_j \) is the coefficient, \( X_i \) contains all variables, and \( \beta \) is a vector of the intercept and all the
coefficients. The mean of the response variable is related to the linear predictor via a
link function $h$:

$$E(y_i \mid X_i) = h^{-1}(X_i\beta) \quad \text{(2)}$$

The data distribution is expressed as

$$p(y_i \mid X_i, \beta, \phi) = \prod_{i=1}^{n} p(y_i \mid X_i, \beta, \phi) \quad \text{(3)}$$

where $\phi$ is a dispersion parameter, and the distribution $p(y_i \mid X_i, \beta, \phi)$ can take various forms, including Normal, Binomial, and Poisson distributions. Some GLMs (for example, the binomial distribution) do not require a dispersion parameter; that is, $\phi$ is fixed at 1. Therefore, GLMs include normal linear and logistic regressions, and various others as special cases.

The classical analysis of generalized linear models is to obtain maximum likelihood estimates (MLE) for the parameters $(\beta, \phi)$ by maximizing the logarithm of the likelihood function: $l(\beta, \phi) = \log p(y \mid X\beta, \phi)$. However, the classical GLMs cannot jointly analyze multiple correlated predictors, due to the problems of non-identifiability and over-fitting. The lasso is a widely used penalization approach to handling high-dimensional models, which adds the $L_1$ penalty to the log-likelihood function and estimates the parameters by maximizing the penalized log-likelihood (ZOU and HASTIE 2005; HASTIE et al. 2009; FRIEDMAN et al. 2010; HASTIE et al. 2015):

$$pl(\beta, \phi) = l(\beta, \phi) - \lambda \sum_{j=1}^{J} | \beta_j | \quad \text{(4)}$$

The overall penalty parameter $\lambda$ controls the overall strength of penalty and the size of the coefficients; for a small $\lambda$, many coefficients can be large, and for a large
\( \lambda \), many coefficients will be shrunk towards zero. The lasso GLMs can be fit by the extremely fast cyclic coordinate descent algorithm, which successively optimizes the penalized log-likelihood over each parameter with others fixed and cycles repeatedly until convergence (FRIEDMAN et al. 2010; HASTIE et al. 2015).

With a single penalty parameter \( \lambda \), however, the lasso can either over-shrink large effects or include a number of irrelevant predictors. Ideally, one should use small penalty values for important predictors and large penalties for irrelevant predictors. If we have no prior knowledge about the importance of the predictors, however, we cannot appropriately preset penalties. We here propose a new approach, i.e., the spike-and-slab lasso GLMs, which can induce different shrinkage scales for different coefficients and allows us to estimate the shrinkage scales from the data.

### Spike-and-slab lasso GLMs

The spike-and-slab lasso GLMs are more easily interpreted and handled from Bayesian hierarchical modeling framework. It is well known that the lasso can be expressed as a hierarchical model with double-exponential prior on coefficients (TIBSHIRANI 1996; PARK and CASELLA 2008; YI and XU 2008; KYUNG et al. 2010):

\[
\beta_j \mid s \sim DE(\beta_j \mid 0, s) = \frac{1}{2s} \exp \left( -\frac{|\beta_j|}{s} \right)
\]

where the scale \( s \) controls the amount of shrinkage; smaller scale induces stronger shrinkage and forces the estimates of \( \beta_j \) towards zero.

We develop the spike-and-slab lasso GLMs by extending the double-exponential prior to the spike-and-slab mixture double-exponential prior:
\[ \beta_j \mid \gamma_j, s_0, s_1 \sim (1 - \gamma_j)DE(\beta_j \mid 0, s_0) + \gamma_j DE(\beta_j \mid 0, s_1) \]

or equivalently

\[ \beta_j \mid \gamma_j, s_0, s_1 \sim DE(\beta_j \mid 0, S_j) = \frac{1}{2S_j} \exp \left( -\frac{|\beta_j|}{S_j} \right) \]  \hspace{1cm} (6)

where \( \gamma_j \) is the indicator variable, \( \gamma_j = 1 \) or 0, and the scale \( S_j \) equals one of two preset positive value \( s_0 \) and \( s_1 \) (\( s_1 > s_0 > 0 \)), i.e., \( S_j = (1 - \gamma_j)s_0 + \gamma_j s_1 \). The scale value \( s_0 \) is chosen to be small and serves a “spike scale” for modeling irrelevant (zero) coefficients and inducing strong shrinkage on estimation, and \( s_1 \) is set to be relatively large and thus serves as a “slab scale” for modeling large coefficients and inducing no or weak shrinkage on estimation. If we set \( s_0 = s_1 \), or \( \gamma_j = 1 \) or 0, the spike-and-slab double-exponential prior becomes the double-exponential prior. Therefore, the spike-and-slab lasso includes the lasso as a special case.

The indicator variables \( \gamma_j \) play an essential role on linking the scale parameters with the coefficients. The indicator variables are assumed to follow the independent binomial distribution:

\[ \gamma_j \mid \theta \sim Bin(\gamma_j \mid 1, \theta) = \theta^{\gamma_j}(1 - \theta)^{1-\gamma_j} \]  \hspace{1cm} (7)

where \( \theta \) is the probability parameter. For the probability parameter \( \theta \), we assume the uniform prior: \( \theta \sim U(0,1) \). The probability parameter \( \theta \) can be viewed as the overall shrinkage parameter that equals the prior probability \( p(\gamma_j = 1 \mid \theta) \). The prior expectation of the scale \( S_j \) equals \( E(S_j) = (1 - \theta)s_0 + \theta s_1 \), which lies in the range \([s_0, s_1]\). As will be seen, the scale \( S_j \) for each coefficient can be estimated, leading to different shrinkage for different predictors.
Algorithm for fitting the spike-and-slab lasso GLMs

For high-dimensional data, it is desirable to have an efficient algorithm that can quickly identify important predictors and build a predictive model. We develop a fast algorithm to fit the spike-and-slab lasso GLMs. Our algorithm, called the EM coordinate descent algorithm, incorporates EM (expectation-maximization) steps into the cyclic coordinate descent procedure for fitting the penalized lasso GLMs regression. We derive the EM coordinate descent algorithm based on the log joint posterior density of the parameters ($\beta, \phi, \gamma, \theta$):

$$
\log p(\beta, \phi, \gamma, \theta | y) = \log p(y | \beta, \phi) + \sum_{j=1}^{J'} \log p(\beta_j | S_j) + \sum_{j=1}^{J'} \log p(\gamma_j | \theta) + \log p(\theta)
$$

$$
\propto l(\beta, \phi) - \sum_{j=1}^{J} \frac{1}{S_j} |\beta_j| + \sum_{j=1}^{J'} (\gamma_j \log \theta + (1 - \gamma_j) \log(1 - \theta))
$$

(8)

where $l(\beta, \phi) = \log p(y | X \beta, \phi)$ and $S_j = (1 - \gamma_j) s_0 + \gamma_j s_1$.

The EM coordinate decent algorithm treats the indicator variables $\gamma_j$ as ‘missing values’ and estimates the parameters ($\beta, \phi, \theta$) by averaging the missing values over their posterior distributions. For the E-step, we calculate the expectation of the log joint posterior density with respect to the conditional posterior distributions of the missing data $\gamma_j$. The conditional posterior expectation of the indicator variable $\gamma_j$ can be derived as

$$
p_j = p(\gamma_j = 1 | \beta_j, \theta, y) = \frac{p(\beta_j | \gamma_j = 1, s_j) p(\gamma_j = 1 | \theta)}{p(\beta_j | \gamma_j = 0, s_0) p(\gamma_j = 0 | \theta) + p(\beta_j | \gamma_j = 1, s_i) p(\gamma_j = 1 | \theta)}
$$

(9)

where $p(\gamma_j = 1 | \theta) = \theta$, $p(\gamma_j = 0 | \theta) = 1 - \theta$, $p(\beta_j | \gamma_j = 1, s_i) = DE(\beta_j | 0, s_i)$, and
\[
p(\beta_j | \gamma_j = 0, s_0) = DE(\beta_j | 0, s_0).\]

Therefore, the conditional posterior expectation of \( S_j^{-1} \) can be obtained by

\[
E(S_j^{-1} | \beta_j) = E\left( \frac{1}{(1-\gamma_j)s_0 + \gamma_j s_1} | \beta_j \right) = \frac{1-p_j}{s_0} + \frac{p_j}{s_1} \tag{10}
\]

It can be seen that the estimates of \( p_j \) and \( S_j \) are larger for larger coefficients \( \beta_j \), leading to different shrinkage for different coefficients.

For the M-step, we update \((\beta, \phi, \theta)\) by maximizing the posterior expectation of the log joint posterior density with \( \gamma_j \) and \( S_j^{-1} \) replaced by their conditional posterior expectations. From the log joint posterior density, we can see that \((\beta, \phi)\) and \( \theta \) can be updated separately, because the parameters \((\beta, \phi)\) are only involved in

\[
l(\beta, \phi) - \sum_{j=1}^{J} S_j^{-1} |\beta_j| \quad \text{and the probability parameter } \theta \text{ is only in}
\]

\[
\sum_{j=1}^{J} \left( \gamma_j \log \theta + (1-\gamma_j) \log(1-\theta) \right). \quad \text{Therefore, the parameters } (\beta, \phi) \text{ are updated by maximizing the expression:}
\]

\[
Q_1(\beta, \phi) = l(\beta, \phi) - \sum_{j=1}^{J} \frac{1}{S_j} |\beta_j| \tag{11}
\]

where \( S_j^{-1} \) is replaced by its conditional posterior expectation derived above. Given the scale parameters \( S_j \), the term \( \sum_{j=1}^{J} \frac{1}{S_j} |\beta_j| \) serves as the \( L_1 \) lasso penalty with \( S_j^{-1} \) as the penalty factors, and thus the coefficients can be updated by maximizing \( Q_1(\beta, \phi) \) using the cyclic coordinate decent algorithm. Therefore, the coefficients can be estimated to be zero. The probability parameter \( \theta \) is updated by maximizing the expression:
\[ Q_j(\theta) = \sum_{j=1}^{J} \left( \gamma_j \log \theta + (1 - \gamma_j) \log(1 - \theta) \right) \]  \hspace{1cm} (12)

We can easily obtain: \( \theta = \frac{1}{J} \sum_{j=1}^{J} p_j \).

In summary, the EM coordinate decent algorithm for fitting the spike-and-slab lasso Cox models proceeds as follows:

1) Choose a starting value for \( \beta^0 \), \( \phi^0 \) and \( \theta^0 \). For example, we can initialize \( \beta^0 = 0 \), \( \phi^0 = 1 \) and \( \theta^0 = 0.5 \).

2) For \( t = 1, 2, 3, \ldots \),
   - E-step: Update \( \gamma_j \) and \( S_j^{-1} \) by their conditional posterior expectations.
   - M-step:
     - a) Update \((\beta, \phi)\) using the cyclic coordinate decent algorithm;
     - b) Update \( \theta \).

We assess convergence by the criterion: \( \left| d^{(t)} - d^{(t-1)} \right| / \left( 0.1 + \left| d^{(t)} \right| \right) < \epsilon \), where \( d^{(t)} = -2 \log l(\beta^{(t)}, \phi^{(t)}) \) is the estimate of deviance at the \( t \)th iteration, and \( \epsilon \) is a small value (say \( 10^{-5} \)).

Selecting optimal scale values

The performance of the spike-and-slab lasso approach can depend on the scale parameters \((s_0, s_1)\). Rather than restricting attention to a single model, our fast algorithm allows us to quickly fit a sequence of models, from which we can choose an optimal one based on some criteria. Our strategy is to fix the slab scale \( s_1 \) (e.g., \( s_1 = 1 \)), and consider a sequence of \( L \) decreasing values \( \{ s_0^l \} \): \( s_1 > s_0^1 > s_0^2 > \cdots > s_0^L > 0 \), for
the spike scale $s_0$. We then fit $L$ models with scales \{$(s_{0,l}, s_l) : l = 1, \ldots, L$\}. Increasing the spike scale $s_0$ tends to include more non-zero coefficients in the model. This procedure is similar to the lasso implemented in the widely-used R package glmnet, which quickly fits the lasso model over a grid of values of $\lambda$ covering its entire range, giving a sequence of models for users to choose from (FRIEDMAN et al. 2010; HASTIE et al. 2015).

**Evaluation of predictive performance**

There are several measures to evaluate the performance of a fitted GLM (STEYERBERG 2009), including: (1) Deviance, which is a generic way of measuring the quality of any model, and is defined as: $d = -2 \sum_{i=1}^{n} \log p(y_i | X_i \hat{\beta}, \phi)$. Deviance measures the overall quality of a fitted GLM, and thus is usually used to choose an optimal model; (2) Mean squared error (MSE), defined as $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$; (3) For logistic regression, we can use two additional measures: AUC (area under the ROC curve) and misclassification. The misclassification is defined as: $\frac{1}{n} \sum_{i=1}^{n} I(|y_i - \hat{y}_i| > 0.5)$, where $I(|y_i - \hat{y}_i| > 0.5) = 1$ if $|y_i - \hat{y}_i| > 0.5$, and $I(|y_i - \hat{y}_i| > 0.5) = 0$ if $|y_i - \hat{y}_i| \leq 0.5$; (4) Pre-validated linear predictor analysis (TIBSHIRANI and EFRON 2002; HASTIE et al. 2015); for all GLMs, we can use the cross-validated linear predictor $\hat{\eta}_i = X_i \hat{\beta}$ as a continuous covariate in a univariate GLM to predict the outcome: $E(y_i | \hat{\eta}_i) = h^{-1}(\mu + \hat{\eta}_i b)$. We then look at the $p$-value for testing the hypothesis $b = 0$ or other statistics to evaluate the predictive performance. We also can transform the continuous cross-validated linear predictor
\( \hat{\eta}_i \) into a categorical factor \( c_i = (c_{i1}, \ldots, c_{ik}) \) based on the quantiles of \( \hat{\eta}_i \), for example, 5%, 25%, 50%, 75% and 95% quantiles, and then fit the model:

\[
E(y_i | c_i) = h^{-1}(\mu + \sum_{k=2}^{K} c_{ik} b_k).
\]

This allows us to compare statistical significance and prediction between different categories.

To evaluate the predictive performance of the proposed model, a general way is to fit the model using a data set and then calculate the above measures with independent data. We use the pre-validation method, a variant of cross-validation (TIBSHIRANI and EFRON 2002; HASTIE et al. 2015), by randomly splitting the data to \( K \) subsets of roughly the same size, and using \((K - 1)\) subsets to fit a model. Denote the estimate of coefficients from the data excluding the \( k \)-th subset by \( \hat{\beta}^{(k)} \). We calculate the linear predictor \( \hat{\eta}_{(k)} = X_{(k)} \hat{\beta}^{(k)} \) for all individuals in the \( k \)-th subset of the data, called the cross-validated or pre-validated predicted index. Cycling through \( K \) parts, we obtain the cross-validated linear predictor \( \hat{\eta} \) for all individuals. We then use \((y_i, \hat{\eta}_i)\) to compute the measures described above. The cross-validated linear predictor for each patient is derived independently of the observed response of the patient, and hence the ‘pre-validated’ dataset \{\( y_i, \hat{\eta}_i \}\) can essentially be treated as a ‘new dataset’. Therefore, this procedure provides valid assessment of the predictive performance of the model (TIBSHIRANI and EFRON 2002; HASTIE et al. 2015). To get stable results, we run 10 times 10-fold cross-validation for real data analysis.

**Implementation**

We have created an R function \texttt{bmlasso} for setting up and fitting the
spike-and-slab lasso GLMs, and several other R functions (e.g., `summary.bh`, `plot.bh`, `predict.bh`, `cv.bh`) for summarizing the fitted models and for evaluating the predictive performance. We have incorporated these functions into the freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

**Simulation Study**

**Simulation Design**

We used simulations to validate the proposed spike-and-slab lasso approach, and to compare with the lasso in the R package `glmnet`. Although the proposed method can be applied to any GLMs, we focused on the spike-and-slab lasso logistic regression, because we analyzed binary outcomes in our real data sets (see the next section). In each situation, we simulated two data sets, and used the first one as the training data to fit the models and the second one as the test data to evaluate the predictive values. For each simulation setting, we replicated the simulation 50 times and summarized the results over these replicates. We reported the results on the predictive values including deviance, MSE, AUC, and misclassification in the test data, and the accuracy of parameter estimates and the proportions of coefficients included in the model.

For each dataset, we generated $n (= 500)$ observations, each with a binary response and a vector of $m (= 1000, 3000)$ continuous predictors $X_i = (x_{i1}, \ldots, x_{im})$. The vector $X_i$ was generated with 50 elements at a time, i.e., the sub-vector $(x_{i(50k+1)}, \ldots, x_{i(50k+50)})$, $k = 0, 1, \ldots$, was randomly sampled from multivariate normal distribution $N_{m_0}(0, \Sigma)$, where $\Sigma = (\sigma_{jj'})$ with $\sigma_{jj'} = 1$ and $\sigma_{jj'} = 0.6$ ($j \neq j'$). Thus, the
predictors within a group were correlated and between groups were independent. To simulate a binary response, we first generated Gaussian response $z_i$ from univariate normal distribution $N(\eta_i, 1.6^2)$, where $\eta_i = \beta_0 + \sum_{j=1}^{m} x_{ij} \beta_j$, and then transformed the continuous response to a binary data by setting individuals with the 30% largest continuous response $Z$ as ‘affected’ ($y_i = 1$) and the other individuals as ‘unaffected’ ($y_i = 0$). We set five coefficients $\beta_5, \beta_{20}, \beta_{40}, \beta_{m-50}$, and $\beta_{m-5}$ to be non-zero, two of which are negative, and all others to be zero. Table 1 shows the preset non-zero coefficient values for six simulation scenarios.

(Insert Table 1 here)

We analyzed each simulated data set using the lasso logistic model implemented in the R package glmnet and the proposed spike-and-slab lasso logistic regression in our R package BhGLM. For the lasso approach, we used 10-fold cross-validation to select an optimal value of $\lambda$, which determines an optimal lasso model, and reported the results based on the optimal lasso model. For the proposed spike-and-slab lasso GLMs approach, we fixed slab scale as $s_1 = 1$, and run a grid value of spike scales: $s_0 = 0.01, 0.02, \ldots, 0.07$.

To fully investigate the impact of the scales ($s_0, s_1$) on the results, we also fitted the simulated data sets under scenarios 3 and 6 (see Table 1) with 100 combinations when $s_0$ changed from 0.01 to 0.10 and $s_1$ from 0.7 to 2.5. We analyzed each scale ($s_0, s_1$) combination over 50 replicates. We also presented the solution path under simulation scenarios 3 and 6, averaged over 50 replicates, to show the special characteristic of spike-and-slab lasso GLMs.
Simulation results

Impact of scales \((s_0, s_1)\) and solution path. We analyzed the data with a grid of values of \((s_0, s_1)\) covering their entire range to fully investigate the impact of the prior scales. Figure 1 shows the profiles of the deviance under scenarios 3 and 6. It can be seen that the slab scale \(s_1\) within the range \([0.7, 2.5]\) had little influence on the deviance, while the spike scale \(s_0\) strongly affected the model performance. These results show that our approach with a fixed slab scale \(s_1\) (e.g., \(s_1 = 1\)) is reasonable.

(Insert Figure 1 here)

Figure 2(A) and Figure 2(B) present the solution path for scenario 3 by the proposed model and the lasso model, respectively. The Figure 2(C and D) shows the profiles of deviance by 10 times 10-fold cross-validation for the proposed model and the lasso model, respectively. For the proposed spike-and-slab lasso GLMs, the minimum of deviance was achieved by including 10.7 non-zero coefficients (averaged over 50 replicates) when \(s_0\) scale was 0.04. For the lasso model, the optimal model included 29.8 non-zero coefficients when \(\lambda\) value was 0.035. Similar to the lasso, the spike-and-slab lasso GLMs is a path-following strategy for fast dynamic posterior exploration. However, its solution path is essentially different from that of the lasso model. For the lasso solution as shown in Figure 2(B), the number of non-zero coefficients could be a few, even zero if a strong penalty was adopted. However, a spike-and-slab mixture prior can help larger coefficients escape the gravitational pull of the spike. Larger coefficients will be always included in the model with none or...
weak shrinkage. **Figure S1** shows the solution path of the proposed model and the lasso model under scenario 6. **Figure S2** shows the adaptive shrinkage amount, along with the different effect size. The same conclusion could be reached, that the spike-and-slab prior shows self-adaptive and flexible characteristics.

(Insert Figure 2 here)

**Predictive performance.** **Table 2** shows the deviance, MSE, AUC, and misclassification in the test data under different simulated scenarios. The smallest deviance is boldfaced in **Table 2**. As described earlier, the deviance measures the overall quality of a model and thus is usually used to choose an optimal model. From these results, we can see that the spike-and-slab lasso GLMs with an appropriate value of $s_0$ performed better than the lasso. **Table 2** also shows that the optimal spike-and-slab GLMs usually had higher AUC value than the lasso. The AUC measures the discriminative ability of a logistic model (STEYERBERG 2009). Thus, the spike-and-slab lasso GLMs generated better discrimination.

(Insert Table 2 here)

**Accuracy of parameter estimates.** **Figure 3** and Figures S3 and S4 show the estimates of coefficients from the best spike-and-slab lasso GLMs and the lasso model over 50 replicates of testing data. It can be seen that the spike-and-slab lasso GLMs provided more accurate estimation in most situations, especially for larger coefficients. This result is expected, because the spike-and-slab prior can induce weak shrinkage on larger coefficients. In contrast, the lasso employs a single penalty on all
the coefficients and thus can over-shrink large coefficients. As a result, five non-zero coefficients were shrunk and underestimated compared to true values in our simulation.

(Insert Figure 3 here)

**Proportions of coefficients included in the model.** We calculated the proportions of the coefficients included in the model over the simulation replicates. Like the lasso model, the proposed spike-and-slab lasso GLMs can estimate coefficients to be zero, and thus can easily return these proportions. **Figure 4** and **Figure S5** show the inclusion proportions of the non-zero coefficients and the zero coefficients for the best spike-and-slab lasso GLMs and the lasso model. It can be seen that the inclusion proportions of the non-zero coefficients were similar for the two approaches in most situations. However, the lasso included zero coefficients in the model more frequently than the spike-and-slab lasso GLMs. This indicates that the spike-and-slab lasso approach can reduce noisy signals.

(Insert Figure 4 here)

We summarized the average numbers of non-zero coefficients and the mean absolute errors (MAE) of coefficient estimates, defined as $\text{MAE} = \frac{\sum |\hat{\beta}_j - \beta_j|}{m}$, in **Table 3**. In most simulated scenarios, the average numbers of non-zero coefficients in the spike-and-slab lasso GLMs were much lower than those in the lasso model. We also found that the average numbers of non-zero coefficients detected by the proposed models were close to the number of the simulated non-zero coefficients in most
scenarios. However, the lasso usually included many zero coefficients in the model. This suggests that the noises can be controlled by the spike-and-slab prior.

(Insert Table 3 here)

**Application to real data**

**Dutch breast cancer data**

We applied our spike-and-slab lasso GLMs to analyze a well-known Dutch breast cancer data set, and compared the results with that of the lasso in R packages glmnet. This data was described in van DE Vijer et al. (2002) and is publicly available in the R package 'breastCancerNKI' (https://cran.r-project.org). This data set contains the microarray mRNA expression measurements of 4919 genes and the event of metastasis after adjuvant systemic therapy from 295 women with breast cancer (VAN 'T VEER et al. 2002; VAN DE VIJVER MJ et al. 2002). The 4919 genes were selected from 24885 genes, for which reliable expression is available (VAN 'T VEER et al. 2002). Among 295 tumors, 88 had distant metastases. Our analysis was to build a logistic model for predicting the metastasis event using the 4919 gene-expression predictors. Prior to fitting the models, we standardized all the predictors. For hierarchical (and also penalized) models, it is important to use a roughly common scale for all predictors.

We fixed the slab scale $s_1$ to 1, and varied the spike scale $s_0$ over the grid of values: $0.005 + k \times 0.005; k = 0, 1, \cdots, 39$, leading to 40 models. We performed 10 times 10-fold cross-validation to select an optimal model based on the deviance.
Figure 5(A) shows the profile of the pre-validated deviance for Dutch breast cancer dataset. The minimum value of deviance appears to be 336.880(4.668), when the spike scale $s_0$ is 0.09. Therefore, the spike-and-slab lasso GLMs with the prior scale $(0.09, 1)$ was chosen for the model fitting and prediction. We further used 10-fold cross-validation over 10 replicates to evaluate the predictive values of the chosen model with the prior scale $(0.09, 1)$. As comparison, we fitted the model by the lasso approach, and also performed 10-fold cross-validation over 10 replicates. Table 4 summarized the measures of performance. The result of the proposed spike-and-slab lasso GLMs was better than the result of the lasso. The cross-validated AUC by the proposed method was estimated to be 0.684(0.011), which is significantly larger than 0.5, showing the discriminative ability of the prognostic model. Totally, 51 genes were detected, and the effect sizes for most of these genes were small (Figure S6).

(Insert Figure 5 and Table 4 here)

We further estimated the pre-validated linear predictor, $\hat{\eta}_i = X_i \hat{\beta}$, for each patient, and then grouped the patients on the basis of the pre-validated linear predictor into categorical factor according to 5th, 25th, 50th, 75th, and 95th percentiles, denoted by $c_i = (c_{i1}, \cdots, c_{ik})$. We fitted the univariate model $E(y_i | \hat{\eta}_i) = h^{-1}(\mu + \hat{\eta}_i b)$ and multivariate model $E(y_i | c_i) = h^{-1}(\mu + \sum_{k=1}^{K} c_{ik} b_k)$ by using the pre-validated linear predictor and the categorical factors, respectively. The results are summarized in Table 5. As expected, the two models were significant, indicating that the proposed prediction model was very informative.

(Insert Table 5 here)
**TCGA Ovarian Cancer (OV) dataset**

The second data set that we analyzed was microarray mRNA expression data for ovarian cancer (OV) downloaded from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) (update at March 2016). The outcome of interest is the tumor event within two years (started the date of initial pathologic diagnosis), including progression, recurrence and new primary malignancies. Clinical data are available for 513 OV patients, where 110 tumors have no interested outcome records. Microarray mRNA expression data (Agilent Technologies platform) includes 551 tumors that have 17785 features profiles after removing the duplications. Even though we can analyze all these features, considering that the genes with small variance might contribute little to the outcome, we selected the top 30% genes filtered by variance for predictive modeling. We merged the expression data and the tumor with new tumor event. After removing individuals with missing response, totally, 362 tumors with 5336 genes were included in our analysis.

Similar to the above analysis of the Dutch breast cancer dataset, we fixed the slab scale $s_1$ to 1 and considered the spike scale $s_0$ over the grid of values: $0.005 \times k$; $k = 0, 1, \cdots, 39$. We performed 10-fold cross-validation with 10 replicates to select an optimal model based on the pre-validated deviance. Figure 5(B) shows the profile of the pre-validated deviance. The minimum value of deviance appears to be 394.796(7.684) when the prior $s_0 = 0.095$. Therefore, the scale (0.095, 1) was chosen for the model fitting and prediction. We also fitted the optimal lasso model using
10-fold cross-validation over 10 replicates. **Table 4** summarized the measures of performance of the proposed spike-and-slab lasso GLMs and the lasso model. We can see that the spike-and-slab lasso was slightly better than the lasso model. The cross-validated AUC was estimated to be 0.647(0.017), which is significantly larger than 0.5, showing the discriminative ability of the prognostic model. Totally, 85 genes were detected, and the effect sizes for most of these genes were small (**Figure S7**).

We further estimated the cross-validated linear predictor, and performed similar analysis as above Dutch breast cancer dataset. The results are summarized in **Table 5**. The univariate and multivariate analyses suggested that the proposed prediction model was very informative for predicting the new tumor event in ovarian tumors.

**Discussion**

Oomics technologies allow researchers to produce tons of molecular data about cancer and other diseases. These multilevel molecular and also environmental data provide an extraordinary opportunity to predict the likelihood of clinical benefit for treatment options and promote more accurate and individualized health predictions (**CHIN et al.** 2011; **COLLINS** and **VARMUS** 2015). However, there are huge challenges to make prediction and detect associated biomarkers from large-scale molecular datasets. In this article, we have developed a new hierarchical model approach, i.e., the spike-and-slab lasso GLMs, for detecting important variables and prognostic prediction (e.g., the event such as tumor recurrence or death). Although focusing on
molecular profiling data and binary outcome in our simulations and real data analysis, the proposed approach also can be used for analyzing general large-scale data and other generalized linear models.

The key to our spike-and-slab lasso GLMs is proposing the new prior distribution, i.e., the mixture spike-and-slab double-exponential prior, on the coefficients. The mixture spike-and-slab prior can induce different amounts of shrinkage for different predictors depending on their effect sizes, and thus have the effect of removing the irrelevant predictors, while supporting the larger coefficients, thus improving the accuracy of coefficient estimation and prognostic prediction. The theoretical property of the spike-and-slab prior is characterized by the solution path. Without the slab component, the output would be equivalent or similar to the lasso solution path. Instead, the solution path of the model with the spike-and-slab prior is different from the lasso. The large coefficients are usually included in the model, while irrelevant coefficients are shrunk to zero.

The large-scale spike-and-slab lasso GLMs can be effectively fitted by the proposed EM coordinate descent algorithm, which incorporates EM steps into the cyclic coordinate descent algorithm. The E-steps involve calculating the posterior expectations of the indicator variable $\gamma_j$ and the scale $S_j$ for each coefficient, and the M-steps employ the existing fast algorithm, i.e., the cyclic coordinate descent algorithm (FRIEDMAN et al. 2010; SIMON et al. 2011; HASTIE et al. 2015), to update the coefficients. As shown in our extensive simulation and real data analyses, the EM coordinate descent algorithm converges rapidly, and is capable of identifying |
important predictors and building promising predictive models from numerous candidates.

The spike-and-slab lasso remains the advantages of two popular methods for high-dimensional data analysis (ROČKOVÁ and GEORGE 2015), i.e., Bayesian variable selection (GEORGE and MCCulloch 1993; CHIPMAN 1996; GEORGE and MCCulloch 1997; CHIPMAN et al. 2001; ROČKOVÁ and GEORGE 2014) and the penalized lasso (TIBSHIRANI 1996; TIBSHIRANI 1997; HASTIE et al. 2015), and bridges these two methods into one unifying framework. Similar to the lasso, the proposed method can shrink many coefficients exactly to zero, thus automatically achieving variable selection and yielding easily interpretable results. More importantly, due to using the spike-and-slab mixture prior, the shrinkage scale for each predictor can be estimated from the data, yielding weak shrinkage on important predictors but strong shrinkage on irrelevant predictors and thus diminishing the well-known estimation bias of the lasso. This self-adaptive strategy is very much in contrast to the lasso model which shrinks all estimates equally with a constant penalty. These remarkable properties of the mixture spike-and-slab priors have been theoretically proved previously for normal linear models (ROČKOVÁ and GEORGE 2015), and also have been observed in our methodological derivation and empirical studies.

The performance of the spike-and-slab lasso GLMs depends on the scale parameter of the double-exponential prior. Optimally presetting two scale values \((s_0, s_1)\) can be difficult. A comprehensive approach is to perform a two-dimensional search on all plausible combinations of \((s_0, s_1)\), and then to select an optimal model.
based on cross-validation. However, this approach can be time-consuming and inconvenient to use. We performed simulations by the two-dimensional search and observed that the slab scale $s_1$ within the range $[0.75, 2.5]$ has little influence on the fitted model, while the spike scale $s_0$ can strongly affect the model performance. Therefore, we suggest a path-following strategy for fast dynamic posterior exploration, which is similar to the approach of Ročková and George (2014, 2015) (Ročková and George 2014; Ročková and George 2015). We fixed the slab scale $s_1$ (e.g., $s_1 = 1$) and run a grid values of spike scale $s_0$ from a reasonable range, e.g., $(0, 0.1)$, and then selected an optimal according to cross-validation. Therefore, rather than restricting analysis based on a single value for the scale $s_0$, the speed of the proposed algorithm makes it feasible to consider several or tens of reasonable values for the scale $s_0$.

The proposed framework is highly extensible. The benefits of hierarchical modeling are flexible and easy to incorporate structural information about the predictors into predictive modeling. With the spike-and-slab priors, the biological information (for example, biological pathways, molecular interactions of genes, and genomic annotation, etc.) and multi-level molecular profiling data (e.g., clinical, gene expression, DNA methylation, somatic mutation, etc.) could be effectively integrated. These prior biological knowledges will help improve prognosis and prediction (Barillot et al. 2012). Another important extension is to incorporate a polygenic random component for modeling small-effect predictors or relationship among individuals into the spike-and-slab lasso GLMs. Zhou et al. (2013) proposed a
Bayesian sparse linear mixed model that includes numerous predictors with mixture
normal priors and a polygenic random effect and developed a MCMC algorithm to fit
the model. We will incorporate the idea of Zhou et al. (2013) into the framework of
our spike-and-slab lass GLMs and develop faster model-fitting algorithms.
Acknowledgments

We thank the associate editor and two reviewers for their constructive suggestions and comments. This work was supported in part by the research grant: NIH2 R01GM069430, and was also supported by grants from China Scholarship Council, the National Natural Science Foundation of China (81573253), and project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions at Soochow University.

Competing interests

The authors declare that they have no competing interests.

Contribution

Study conception and design: NY, ZXT

Simulation study and data summary: NY, ZXT, XYZ, YPS

Real data analysis: ZXT, XYZ, YPS

Drafting of manuscript: NY, ZXT
REFERENCES


Hastie, T., R. Tibshirani and J. Friedman, 2009 The Elements of Statistical Learning. Springer-Verlag, New York, NY, USA.


Ročková, V., and E. I. George, 2015 The spike-and-slab lasso. manuscript.


Zhao, Q., X. Shi, Y. Xie, J. Huang, B. Shia et al., 2014 Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Briefings in Bioinformatics.


Table 1. The simulated effect sizes of five non-zero coefficients under different scenarios. m is the number of simulated predictors. For all scenarios, the number of individuals (n) is 500.

<table>
<thead>
<tr>
<th>Simulated scenarios</th>
<th>$\beta_5$</th>
<th>$\beta_{20}$</th>
<th>$\beta_{40}$</th>
<th>$\beta_{m-50}$</th>
<th>$\beta_{m-5}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>m=1000</td>
<td>0.362</td>
<td>0.395</td>
<td>-0.418</td>
<td>-0.431</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>m=1000</td>
<td>-0.457</td>
<td>-0.491</td>
<td>0.521</td>
<td>0.550</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>m=1000</td>
<td>0.563</td>
<td>-0.610</td>
<td>0.653</td>
<td>-0.672</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>m=3000</td>
<td>0.357</td>
<td>-0.388</td>
<td>0.414</td>
<td>-0.429</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>m=3000</td>
<td>-0.455</td>
<td>0.509</td>
<td>0.528</td>
<td>0.552</td>
</tr>
<tr>
<td>Scenario 6</td>
<td>m=3000</td>
<td>0.560</td>
<td>-0.618</td>
<td>0.654</td>
<td>-0.673</td>
</tr>
</tbody>
</table>
Table 2. Estimates of four measures over 50 replicates under different simulated scenarios. Values in the parentheses are standard errors. “sslasso” represents the spike-and-slab lasso GLMs. The slab scales, $s_1$, are 1 in all scenarios. The smallest deviance values are boldfaced to indicate the optimal model.

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>n=500, m=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>638.983(17.431)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.01$</td>
<td>685.229(18.922)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.02$</td>
<td>643.119(23.694)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.03$</td>
<td>637.226(22.197)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.04$</td>
<td>636.930(18.928)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.05$</td>
<td>639.385(16.732)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.06$</td>
<td>639.784(17.359)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.07$</td>
<td>645.151(19.752)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>n=500, m=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>601.872(15.666)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.01$</td>
<td>640.816(26.885)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.02$</td>
<td>580.940(24.945)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.03$</td>
<td>581.661(28.271)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.04$</td>
<td>583.037(21.964)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.05$</td>
<td>590.185(19.343)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.06$</td>
<td>595.879(19.388)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.07$</td>
<td>603.756(20.202)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 3</th>
<th>n=500, m=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>561.917(14.623)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.01$</td>
<td>585.600(34.703)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.02$</td>
<td>531.956(26.214)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.03$</td>
<td>532.747(26.343)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.04$</td>
<td>530.781(24.638)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.05$</td>
<td>541.192(24.496)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.06$</td>
<td>550.971(25.065)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.07$</td>
<td>559.430(24.311)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 4</th>
<th>n=500, m=3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>665.349(11.253)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.01$</td>
<td>680.988(16.432)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.02$</td>
<td>655.714(23.241)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.03$</td>
<td>646.877(20.963)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.04$</td>
<td>645.278(16.039)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.05$</td>
<td>654.349(16.241)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.06$</td>
<td>665.488(18.227)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.07$</td>
<td>675.374(20.660)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 5</th>
<th>n=500, m=3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>620.034(16.209)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.01$</td>
<td>642.083(30.947)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.02$</td>
<td>597.547(34.288)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.03$</td>
<td>593.701(32.304)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.04$</td>
<td>596.421(30.906)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.05$</td>
<td>610.549(24.024)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.06$</td>
<td>623.014(24.530)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.07$</td>
<td>634.536(26.023)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 6</th>
<th>n=500, m=3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>570.158(17.989)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.01$</td>
<td>568.332(35.346)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.02$</td>
<td>537.665(28.103)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.03$</td>
<td>530.081(29.097)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.04$</td>
<td>530.553(26.149)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.05$</td>
<td>542.091(26.825)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.06$</td>
<td>557.014(27.697)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.07$</td>
<td>572.405(28.018)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 7</th>
<th>n=500, m=3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>570.158(17.989)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.01$</td>
<td>568.332(35.346)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.02$</td>
<td>537.665(28.103)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.03$</td>
<td>530.081(29.097)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.04$</td>
<td>530.553(26.149)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.05$</td>
<td>542.091(26.825)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.06$</td>
<td>557.014(27.697)</td>
</tr>
<tr>
<td>sslasso: $s_1=0.07$</td>
<td>572.405(28.018)</td>
</tr>
</tbody>
</table>
Table 3. Average number of non-zero coefficients and mean absolute error (MAE) of coefficient estimates over 50 simulation replicates. Values in the parentheses are standard errors.

<table>
<thead>
<tr>
<th>Simulation scenarios</th>
<th>sslasso ($s_1=1$)</th>
<th>lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average number</td>
<td>MAE</td>
</tr>
<tr>
<td>Scenario 1, $s_0=0.04$</td>
<td>15.800</td>
<td>1.431(0.420)</td>
</tr>
<tr>
<td>Scenario 2, $s_0=0.03$</td>
<td>5.780</td>
<td>0.750(0.452)</td>
</tr>
<tr>
<td>Scenario 3, $s_0=0.04$</td>
<td>10.420</td>
<td>0.669(0.286)</td>
</tr>
<tr>
<td>Scenario 4, $s_0=0.04$</td>
<td>32.900</td>
<td>1.984(0.352)</td>
</tr>
<tr>
<td>Scenario 5, $s_0=0.03$</td>
<td>6.460</td>
<td>0.906(0.498)</td>
</tr>
<tr>
<td>Scenario 6, $s_0=0.03$</td>
<td>5.920</td>
<td>0.629(0.362)</td>
</tr>
</tbody>
</table>
### Table 4. The measures of optimal spike-and-slab lasso (sslasso) and lasso models for Dutch breast cancer dataset and TCGA ovarian cancer dataset by 10 times 10-fold cross validation. Values in the parentheses are standard errors.

<table>
<thead>
<tr>
<th></th>
<th>deviance</th>
<th>MSE</th>
<th>AUC</th>
<th>misclassification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dutch breast cancer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sslasso</td>
<td>336.880(4.668)</td>
<td>0.192(0.003)</td>
<td>0.684(0.011)</td>
<td>0.284(0.014)</td>
</tr>
<tr>
<td>lasso</td>
<td>342.134(3.383)</td>
<td>0.197(0.003)</td>
<td>0.656(0.018)</td>
<td>0.297(0.007)</td>
</tr>
<tr>
<td><strong>TCGA ovarian cancer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sslasso</td>
<td>394.796(7.684)</td>
<td>0.179(0.004)</td>
<td>0.647(0.017)</td>
<td>0.258(0.010)</td>
</tr>
<tr>
<td>lasso</td>
<td>393.152(5.392)</td>
<td>0.179(0.003)</td>
<td>0.636(0.017)</td>
<td>0.254(0.005)</td>
</tr>
</tbody>
</table>
Table 5. Univariate and multivariate analyses using the pre-validated linear predictors and their categorical factors.

<table>
<thead>
<tr>
<th>model</th>
<th>coefficients</th>
<th>Estimates</th>
<th>Std. Error</th>
<th>p values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch breast cancer</td>
<td>univariate</td>
<td>b</td>
<td>0.816</td>
<td>0.162</td>
</tr>
<tr>
<td></td>
<td>multivariate</td>
<td>$b_2$</td>
<td>-0.619</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_3$</td>
<td>0.251</td>
<td>0.700</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_4$</td>
<td>0.412</td>
<td>0.697</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_5$</td>
<td>1.556</td>
<td>0.696</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_6$</td>
<td>1.520</td>
<td>0.827</td>
</tr>
<tr>
<td>TCGA ovarian cancer</td>
<td>univariate</td>
<td>b</td>
<td>0.683</td>
<td>0.162</td>
</tr>
<tr>
<td></td>
<td>multivariate</td>
<td>$b_2$</td>
<td>0.347</td>
<td>0.519</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_3$</td>
<td>0.906</td>
<td>0.518</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_4$</td>
<td>1.426</td>
<td>0.536</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_5$</td>
<td>1.719</td>
<td>0.572</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$b_6$</td>
<td>2.035</td>
<td>0.877</td>
</tr>
</tbody>
</table>
Figures Legend

Figure 1. The profiles of deviance under scenarios 3 and 6 over 50 replicated testing datasets. \((s_0, s_1)\) is the prior scale for the spike-and-slab lasso GLMs.

Figure 2. The solution path and deviance profiles of the spike-and-slab lasso GLMs (A and C) and the lasso model (B and D) for scenario 3. The colorful points on the solution path represent the estimated values of assumed five non-zero coefficients, and the circles represent the true non-zero coefficients. The vertical lines correspond to the optimal models.

Figure 3. The parameter estimation averaged over 50 replicates for the spike-and-slab lasso GLMs and the lasso model under scenarios 3 and 6. The blue cycles are the assumed true values. The black points and lines represent the estimated values and the interval estimates of coefficients over 50 replicates.

Figure 4. The inclusion proportions of the non-zero and zero coefficients in the model over 50 simulation replicates under scenarios 3 and 6. The black points and red cycles represent the proportions of non-zero coefficients for the spike-and-slab lasso GLMs and the lasso model, respectively. The blue points and grey cycles represent the proportions of zero coefficients for the spike-and-slab lasso GLMs and the lasso model, respectively.

Figure 5. The profiles of pre-validated deviance under varied prior scales \(s_0\) and fixed \(s_1=1\) for Dutch breast cancer dataset (A) and TCGA Ovarian Cancer dataset (B).
Supplemental Figures Legend

Figure S1. The solution path and deviance profiles of the spike-and-slab lasso GLMs (A and C) and the lasso model (B and D) for scenario 6. The colorful points on the solution path represent the estimated values of assumed five nonzero coefficients, and the circles represent the true non-zero coefficients. The vertical lines correspond to the optimal models.

Figure S2. The adaptive shrinkage prior scale ($S_j$) with varying effect size from -2.0 to 2.0 for the simulation $(n, m) = (500, 1000)$ and $(s_0, s_1) = (0.04, 1)$.

Figure S3. The parameter estimation averaged over 50 replicates for the spike-and-slab lasso GLMs and the lasso model under scenarios 1 and 2. The blue cycles are the assumed true values. The black points and lines represent the estimated values and the interval estimates of coefficients over 50 replicates.

Figure S4. The parameter estimation averaged over 50 replicates for the spike-and-slab lasso GLMs and the lasso model under scenarios 4 and 5. The blue cycles are the assumed “true” values. The black points and lines represent the estimated values and the interval estimates of coefficients over 50 replicates.

Figure S5. The inclusion proportions of the non-zero and zero coefficients in the model over 50 simulation replicates under scenarios 1, 2, 4 and 5. The black points and red cycles represent the proportions of non-zero coefficients for the spike-and-slab lasso GLMs and the lasso model, respectively. The blue points and grey cycles represent the proportions of zero coefficients for the spike-and-slab lasso GLMs and the lasso model, respectively.

Figure S6. The detected genes and their standardized effect sizes estimated by the spike-and-slab lasso model with the prior scale $(0.09, 0.1)$ for Dutch breast cancer dataset.

Figure S7. The detected genes and their standardized effect sizes estimated by the spike-and-slab lasso model under the prior scale $(0.095, 0.1)$ for TCGA ovarian cancer dataset.
Figure 1
Figure 2
Figure 3
Figure 4
Figure 5