www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

Check for updates

OPEN

Bayesian Hyper-LASSO Classification for Feature Selection with Application to Endometrial Cancer RNA-seq Data

Lai Jiang[1,2] ✉, Celia M. T. Greenwood[1,2,3], Weixin Yao[4] & Longhai Li[5] ✉

Feature selection is demanded in many modern scientific research problems that use high-dimensional data. A typical example is to identify gene signatures that are related to a certain disease from high-dimensional gene expression data. The expression of genes may have grouping structures, for example, a group of co-regulated genes that have similar biological functions tend to have similar expressions. Thus it is preferable to take the grouping structure into consideration to select features. In this paper, we propose a Bayesian Robit regression method with Hyper-LASSO priors (shortened by BayesHL) for feature selection in high dimensional genomic data with grouping structure. The main features of BayesHL include that it discards more aggressively unrelated features than LASSO, and it makes feature selection within groups automatically without a pre-specified grouping structure. We apply BayesHL in gene expression analysis to identify subsets of genes that contribute to the 5-year survival outcome of endometrial cancer (EC) patients. Results show that BayesHL outperforms alternative methods (including LASSO, group LASSO, supervised group LASSO, penalized logistic regression, random forest, neural network, XGBoost and knockoff) in terms of predictive power, sparsity and the ability to uncover grouping structure, and provides insight into the mechanisms of multiple genetic pathways leading to differentiated EC survival outcome.

The accelerated development of many high-throughput biotechnologies has made it affordable to collect complete sets of measurements of gene expressions. Scientists are often interested in selecting certain genes that are related to a categorical response variable, such as the onset or progression of cancer. These genes are known as *signatures* in the life sciences literature; for the purposes of our paper, we will call them *features*.

When finding features, we require an algorithm that can identify both sparse features and grouping structures. Sparsity is required because in the context of gene expression analysis, there are often only a few important features. Hence, the feature selection is expected to be sparse. Grouping structure is required because biological features often have an innate grouping structure. For example, there might be a high correlation between certain features; a group of genes might relate to the same molecular pathway, or be in close proximity in the genome sequence, or share a similar methylation profile[1,2]. To better understand disease etiology, therefore, one must understand the grouping structure of the genes associated with that disease.

At first, researchers concentrated on the sparsity problem in high dimensional feature spaces. They developed automatic sparse selection methods such as LASSO (Least Absolute Shrinkage and Selection Operator[3]) and knockoff variable selections[4,5]. In the Bayesian literature, LASSO is equivalent to a linear regression with (convex) Laplace penalty function on the coefficients. But for our purposes—namely, uncovering sparse features and grouping structure—the convex penalty functions have several problems. First, they are not sparse enough. Second, these functions are incapable of uncovering grouping structure. Indeed, the traditional convex sparse feature-selection algorithms are either unable to take grouping structure information into account, or else depend on prior knowledge of the specific grouping structure.

[1]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada. [2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada. [3]Gerald Bronfman Department of Oncology, McGill University, Montreal, Canada. [4]Department of Statistics, University of California, Riverside, US. [5]Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Canada. ✉e-mail: lai.jiang@mail.mcgill.ca; longhai@math.usask.ca

Again, some researchers focused on the first problem: the need for greater sparsity. One potential way to overcome this limitation is to amplify the signal. In recent years, researchers developed methods that are even more aggressive, and hence even more sparse, than LASSO. They proposed fitting classification or regression models with continuous non-convex penalty functions to discover features related to a response. Such non-convex penalty functions include, but are not limited to, hyper-LASSO[6], global-local penalties[7], t with small degrees of freedom[8,9], SCAD[10], horseshoe[11–15], MCP[16], NEG[17], adaptive LASSO[18], Dirichlet-Laplace and Dirichlet-Gaussian[19], and generalized double-pareto[20] functions. Reviews of non-convex penalty functions have been provided by[7,21,22] and[23]. These non-convex penalties can shrink the coefficients of unrelated features (noise) to zero more aggressively than LASSO, while enlarging the coefficients of related features (signal).

These non-convex functions work well for sparsity; however, their effectiveness regarding grouping structure has not yet been explored. In fact, non-convex penalty will make selection within a group of highly correlated features: either splitting important features into different modes of penalized likelihood or suppressing less important features in favour of more important features. The within-group selection is indeed a desired property if our goal is selecting a sparse subset of features. Note that the within-group selection does not mean that we will lose other features within a group that are also related to the response because other features can still be identified from the group representatives using the correlation structure. On the other hand, the within-group selection results in a huge number of modes in the posterior (for example, two groups of 100 features can make 100^2 subsets containing one from each group). Therefore, optimization algorithms encounter great difficulty in reaching a global or good mode because in non-convex regions, the solution paths are discontinuous and erratic.

The application of non-convex penalties in grouped feature selection remains limited by the computational difficulties, instead, researchers interested in grouping structure have tended to favor (convex) LASSO penalty. There have been three main approaches developed to consider directly the grouping structure in classification/regression models based on LASSO penalties:

1. The first approach is to develop methods that directly consider the grouping structure or the correlation among features in classification and regression models. This involves fitting classification models on the "new features" constructed from the feature groups (e.g., centroids or means) of features within groups; see[24–27] and[28].
2. The second approach is to fit classification models with penalties that enforce similarity on the coefficients of features within groups, such as group or fused LASSO[29,30]. Group and fused LASSO functions are more predictive than plain LASSO because they consolidate the predictive power of all features within groups.
3. The third approach is known as two-stage selection, and it includes supervised group LASSO (SGL)[31] and ProtoLASSO[28]. SGL works in two stages: 1. apply LASSO to each group, separately, to determine the features of each group; and 2. apply LASSO to the selected features (from step one). ProtoLASSO, on the other hand, works as follows: 1. select prototype features within each group using marginal correlations; and 2. apply LASSO to the prototype features.

There are three more problems with all three of these approaches. First, the Lasso solution is subject to specific choices of regularization parameter λ and may not select the "correct" sparsity pattern in high dimensional data. For example, the solution may not be sparse enough when the signal to noise ratio is low. Second, these methods make selections separately in each group, so they cannot consider the joint effects of features from different groups. This is particularly limiting when analyzing biological interactions, because the individual associations of certain features (genes) with the outcome (disease) are low, but such features could still be useful when joined with other features (with a correlation structure). Third, in order to consider grouping structure, these methods require a pre-specified grouping index. This pre-specified grouping structure is often found via a clustering algorithm. However, the statistical algorithm's results might not be a perfect match with meaningful feature groups (e.g., with biologically accurate clusters of genes). In addition, such grouping is probably too simple to explain complicated biological activities. For example, an outcome may be related to correlations among multiple groups of features—that is, interactions may occur both within and between groups.

Due to the numerous problems with convex functions, we hypothesize that non-convex functions would better suit our needs regarding sparsity and grouping structure. Unlike convex functions, non-convex functions have at least the potential to find grouping structures without prior knowledge. However, limited work has been done to take into account of grouping structure with non-convex penalties[32–34]. Moreover, non-convex methods are usually computationally expensive, which has limited their application to high-throughput biomedical data despite the potential usefulness.

Another problem is that the algorithms to solve these non-convex functions could be unstable[32]. When it comes to solving these non-convex penalty functions, the algorithms that provide solutions traditionally involve the process of *non-convex learning*. Specifically, non-convex learning processes are optimization algorithms for learning the classification and/or regression likelihood penalized by non-convex functions. But thus far, limited research has been done to develop stable optimization algorithm for non-convex function to uncover both sparse features and grouping structure. A review of such algorithms can be found in[32] and[35].

In this paper, we develop an approach to non-convex penalty functions, which we call BayesHL (for Bayesian Hyper-LASSO). BayesHL is a fully Bayesian approach that uses the Markov chain Monte Carlo (MCMC) method to explore the multi-modal posterior. This is a promising alternative to non-convex learning methods, because unlike many non-convex learning methods[32,36], a well-designed MCMC algorithm can explore many modes to find multiple desirable feature subsets. The development of MCMC methods for exploring regression posteriors based on heavy-tailed priors has emerged only recently; the relevant articles include[9,37–41], among others.

More specifically, we develop a sophisticated MCMC method to explore the posterior of a Robit model assigned with a class of heavy-tailed priors (i.e., Cauchy distribution with small scale). We employ the Hamiltonian Monte Carlo method[42] to draw MCMC samples of the regression coefficients in a restricted Gibbs sampling framework. The framework's computational complexity is more dependent on the number of signals than the number of features; this greatly accelerates the MCMC sampling efficiency for very high-dimensional problems. After MCMC sampling, we divide the samples into sub-pools according to the posterior modes to find a list of sparse feature subsets. This process is further aided by cross-validatory evaluation.

To the best of our knowledge, our work here is one of the few attempts to uncover grouping structure using MCMC in the context of a high-dimensional feature selection problem. The development of fully Bayesian (MCMC) methods for exploring regression posteriors based on heavy-tailed priors has emerged only recently; relevant articles include[6,9,37,43].

Compared to other feature selection methods in the literature, BayesHL has two main benefits. First, BayesHL is more parsimonious than traditional convex (including LASSO) and non-convex learning methods. That is because BayesHL automatically makes selections within groups *without* a pre-specified grouping structure. Consequently, a single feature subset found by BayesHL is more parsimonious than the feature subsets selected by LASSO. It is also possible for BayesHL to consider the joint effects of features from different groups. Thus, the results always include representatives from different groups. Such succinct feature subsets are (compared to the results of traditional convex and non-convex methods) much easier to investigate and interpret according to existing biological knowledge. Second, BayesHL is immune to the problem of correlation bias. Other convex and non-convex methods may not be able to identify significant features when there is a high number of correlated features, due to the problem of correlation bias[2]. But BayesHL will always be able to identify the significant features because it enforces within-group selection, so even as the number of correlated features increases, the magnitudes of coefficients will not decrease.

Our MCMC-based non-convex learning method can effectively identify sparse feature subsets with superior out-of-sample predictive performance. This statement is based on the results of our experiment on a real high-throughput dataset of endometrial cancer. Results show that the BayesHL-selected feature subsets have better predictive power than those selected by competitors. Indeed, after further investigation we found that the BayesHL-selected gene subsets correspond to interactions between gene networks with meaningful biological interpretations.

In brief, in this paper we present a Bayesian feature subset selection method (BayesHL), test it, and use it to uncover an interesting result regarding endometrial cancer. We test our method on simulated datasets with independent or correlated groups of features, to demonstrate its feature subset selection and prediction performance with respect to grouping structures. We apply our method to a high-throughput dataset related to endometrial cancer, and present interesting findings of gene networks regarding the survival outcome of endometrial cancer.

Methodology

Model: heavy-tailed robit model.

The following is an introduction to the notation we use throughout this paper. Suppose we have collected measurements of p features (such as genes) and a binary response (such as a disease indicator) on n training cases. For a case with index i, we use y_i, taking integers 0 or 1, to denote the response value and use a row vector x_i to denote the p features, and the first element of x_i is set to 1 for including intercept term in linear model. Throughout this paper, we will use bold-faced letters to denote vectors or matrices. We will write collectively $y = (y_1, \ldots, y_n)'$, and $X = (x'_1, \ \ldots, \ x'_n)'$ in which rows stand for observations and columns stand for features. Note that, we use index 0 for the intercept term in this paper, ie., the values of the first column of X are all equal to 1, denoted by $x_{,0}$. Using machine learning terminology, we call (y, X) training data, which are used to fit models; in contrast, the data used only in testing the predictive performance is called test data.

For the purposes of feature selection and binary classification, we are interested in modeling the conditional distribution of y_i given x_i. The traditional **probit** models use a normally distributed auxiliary variable z_i to model y_i given x_i as follows:

$$y_i = I(z_i > 0), \quad z_i = x_i\,\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1), \tag{1}$$

where $I(\cdot)$ is the indicator function, and β is a column vector of coefficients with the first element being intercept, denoted by β_0. With z_i integrated out, the above model is equivalent to the following conditional distribution: $P(y_i|x_i, \ \beta) = \Phi(x_i\,\beta)^{y_i}(1 - \Phi(x_i\beta))^{1-y_i}$, for $y_i = 0{,}1$, where Φ is the cumulative distribution function (CDF) of the standard normal distribution.

In high-throughput data, there are typically a large number of extreme outliers. Since probit models cannot accommodate some extreme outliers (due to the light tails of normal distributions), we will use a more robust model: the Robit model[44]. Robit replaces the normal distribution for ε_i with a t distribution, and is thus more robust to outliers than probit and logistic regression (see[44,45] and the references therein).

The Robit model is as follows:

$$y_i = I(z_i \ > \ 0) \quad z_i = x_i\,\beta + \varepsilon_i, \quad \varepsilon_i \sim T(\alpha_0, \omega_0), \tag{2}$$

where $T(\alpha, \omega)$ stands for scaled student's t distribution with degrees of freedom α, scale parameter $\sqrt{\omega}$, and mean parameter 0, with a probability density function (PDF) as $t_{\alpha,\omega}(x) = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\alpha\pi}\ \Gamma\left(\frac{\alpha}{2}\right)}\left(1 + \frac{x^2}{\omega\alpha}\right)^{-\frac{\alpha+1}{2}}\frac{1}{\sqrt{\omega}}$, where $\Gamma(\cdot)$ is the Gamma function. As in probit models, with z_i integrated out, the above model is equivalent to the following conditional distribution of y_i given x_i:

$$P(y_i|x_i, \beta) = T_{\alpha_0,\omega_0}(x_i\beta)^{y_i}(1 - T_{\alpha_0,\omega_0}(x_i\beta))^{1-y_i}, \ for \ \ y_i = 0{,}1, \tag{3}$$

where $T_{\alpha,\omega}(x)$ represents the CDF of $T(\alpha,\omega)$.

$T(\alpha,\omega)$ is given by $T_{\alpha,\omega}(x) = \frac{1}{2} + \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\sqrt{\alpha\pi}\,\Gamma\left(\frac{\alpha}{2}\right)} \times {}_2F_1\left(\frac{1}{2}, \frac{\alpha+1}{2}; \frac{3}{2}; -\frac{x^2}{\alpha\omega}\right) \times \frac{x}{\sqrt{\omega}}$, where ${}_2F_1$ is the hypergeometric function, which is given as the sum of an infinite series[46].

The α_0 is fixed at $\alpha_0 = 1$, which is appropriate for modeling the possible range of outliers. In addition, from the CDF of the *t* distribution, we notice that only $\beta/\sqrt{\omega_0}$ is identifiable in the likelihood of (ω_0, β) given observation $y_1, \ldots, y_n$. Therefore, we fix ω_0 at some reasonable value. We choose to use $\omega_0 = 0.5$ such that the T_{α_0,ω_0} is similar to the logistic distribution near origin zero, but has heavier tails (than the logistic distribution).

Coefficient prior (β): heavy-tailed cauchy prior. Now that we have chosen a model for selecting features (Robit), the next step is to choose a framework for estimating β. The following subsections discuss why we chose to use a heavy-tailed (Cauchy) prior for estimating β, how we raised the sampling efficiency with restricted Gibbs sampling and Hamiltonian Monte Carlo, and why our MCMC algorithm is a good framework for obtaining an estimate of β.

In many problems of linking high-dimensional features to a response variable, it is believed that the non-zero regression coefficients are very sparse—that is, very few features are related to the response *y*. In the past decade, non-convex penalties have drawn the attention of many researchers because they can shrink the coefficients of unrelated features (noise) more aggressively to zero than the convex L_1 penalty. In other words, non-convex penalties provide a *sharper* separation of signal and noise than L_1.

In Bayesian methodologies, a non-convex penalty often corresponds to a prior distribution with a heavier tail than the Laplace distribution (which corresponds to L_1). So in the Bayesian interpretation, a typical sample of β from a heavy-tailed prior has a few extraordinarily large values representing *related features*, and many small values representing *unrelated features*. Therefore, heavy-tailed priors are a better match for our expectations about β than the Laplace prior.

Of the suitable heavy-tailed priors, we have many choices, including Cauchy[8], horseshoe[7,12–14,22,23], and normal-exponential-gamma (NEG)[17], all of which have been proven superior to L_1 in detecting very sparse signals.

However, there are three reasons why we prefer Cauchy to horseshoe and NEG priors:

1. Although although the horseshoe and NEG priors have the same tail heaviness as Cauchy (converging to zero in the rate of $1/\beta^2$), they also have a non-differentiable log PDF at 0; therefore, if penalties are applied, small signals can be shrunken to exactly 0.
2. In our empirical comparison of the predictive performance of classification models using *t* priors with various tail heaviness (including NEG and horseshoe priors), we found that Cauchy had the optimal performance[6].
3. Horseshoe and NEG priors demand additional computation in sampling the hyperparameters (the local variances for each β_j, i.e., λ_j below). Indeed, in our aforementioned paper, the additional computation accounted for half of the whole sampling time, even after we used a restricted Gibbs sampling scheme to greatly shorten the sampling time for regression coefficients.

Therefore, we chose to use the plain Cauchy prior in this paper: *t* with degree of freedom $\alpha_1 = 1$, denoted by $\beta_j \sim T(\alpha_1, \omega_1), \ \text{for} \ \ j = 1, \ldots, p$.

For the purposes of MCMC sampling, we express the *t* prior for β as a scale-mixture normal by introducing a *latent* variance λ_j for each β_j, as follows:

$$\beta_j|\lambda_j \sim N(0,\lambda_j), \tag{4}$$

$$\lambda_j \sim \text{Inverse} - \text{Gamma}\left(\frac{\alpha_1}{2}, \frac{\alpha_1\omega_1}{2}\right). \tag{5}$$

Hereafter, we will refer to this vector as $\lambda = (\lambda_1, \ldots, \lambda_p)$.

In order to shrink small coefficients toward zero, we must choose a very small-scale parameter $\sqrt{\omega_1}$ for Cauchy. In Bayesian methodologies, a typical way to avoid assigning a fixed value to a parameter is to treat it as a hyperparameter such that it will be chosen automatically during MCMC sampling according to marginalized likelihood. However, we have found that this approach does not choose at a sufficiently small scale to yield a very sparse β, because a classification model with *p* features can easily overfit a dataset with sample size $n \ll p$. In order to enforce sparsity in β and to improve the efficiency of MCMC sampling, we choose to fix $\sqrt{\omega_1}$ at a small value $e^{-5} \approx 0.01$. Table 1 shows a number of upper-tailed quantiles of $|\beta_j|$ where $\beta_j \sim \text{Cauchy}(0, e^{-5})$.

From Table 1, we see that this choice of value of ω_1 postulates that 2 of the 1000 features have coefficients with magnitude ≥ 2.228. We believe that this is an appropriate level of sparsity for many high-dimensional feature-selection problems.

Another important reason to fix $\sqrt{\omega_1}$ is the "flatness" (heaviness) of a Cauchy tail. Due to this flatness, very small shrinkage is applied to large coefficients. Since the shrinkage is small, the estimates of large coefficients are robust to $\sqrt{\omega_1}$[6,13]. This is a distinctive property of priors with tails as heavy as Cauchy. (In other priors with similar

Upper probability	0.200	0.100	0.020	0.010	0.002	0.001	0.0001
Quantile of $\|\beta_j\|$	0.022	0.044	0.223	0.446	2.228	4.456	42.895

Table 1. Upper-tailed quantiles of absolute Cauchy with scale e^{-5}.

tail heaviness, like Gaussian, Laplace priors, a careful choice of scale must be made because the shrinkage of large coefficients is large and sensitive to the scale.) Therefore, although $\sqrt{\omega_1}$ is fixed at a very small value around 0.01, the prior does not over-shrink large signals, and can accommodate a wide range of signals.

Implementation: Estimating β using Restricted Gibbs Sampling with Hamiltonian Monte Carlo.

There is great difficulty in maximizing the penalized likelihood function using heavy-tailed and small-scaled priors. For example, using a small scale for $\sqrt{\omega_1}$ such as e^{-5}, the R function bayesglm in R package ARM (which implements penalized logistic regression with Cauchy priors) will converge to a mode where almost all coefficients are shrunken to very small values, even when the number of features (p) is small. On the other hand, using the default 2.5 value, bayesglm does not provide a sparse solution (to be presented in this article). The difficulty in optimization is further intensified by the severe multi-modality in the posterior because heavy-tailed and small-scaled priors can split coefficients of a group of correlated features into different modes rather than shrinking them simultaneously as Gaussian priors do. Therefore, although good theoretical properties of non-convex penalties have been proved in statistics literature (e.g.[16]), many researchers and practitioners have been reluctant to embrace these methods because optimization algorithms often produce unstable solutions[32]. This motivated us to develop MCMC algorithms for exploring the posterior with many modes due to the use of heavy-tailed priors.

Our MCMC algorithm will sample from the joint posterior, $f(\beta, \lambda| y, X)$, which is based on the hierarchical models given by Eqs. (3–5) with α_0, ω_0, α_1, ω_1 fixed (so omitted in the following model descriptions). The log posterior can be written as follows:

$$\begin{aligned}\log(f(\beta, \lambda|y, X)) &= \sum_{i=1}^{n}\log(P(y_i|x_i, \beta)) + \sum_{j=0}^{p}\log(f(\beta_j|\lambda_j)) \\ &\quad + \sum_{j=1}^{p}\log(f(\lambda_j)) + C,\end{aligned} \tag{6}$$

where the first three terms come from the models defined by (3), (4), (5) respectively, and C is the log of the normalization constant unrelated to β and λ. The first three terms in (6) are given as follows:

$$\begin{aligned}\log(P(y_i|x_i, \beta)) &= y_i\log(T_{\alpha_0,\omega_0}(x_i\beta)) + (1-y_i)\log(T_{\alpha_0,\omega_0}(-x_i\beta)) \\ &\equiv lp(y_i|x_i\beta),\end{aligned} \tag{7}$$

$$\log(f(\beta_j|\lambda_j)) = -\frac{1}{2}\log(\lambda_j) - \frac{\beta_j^2}{2\lambda_j} + C_1, \quad \text{for } j = 0, \ldots, p \tag{8}$$

$$\log(f(\lambda_j)) = -\left(\frac{\alpha_1}{2} + 1\right)\log(\lambda_j) - \frac{\alpha_1\omega_1}{2\lambda_j} + C_2, \text{ for } j = 1, \ldots, p. \tag{9}$$

where C_1, C_2 are two constants unrelated to (β, λ); the function $lp(y_i| x_i\beta)$ is introduced to indicate that the probability of y_i given x_i is a function of $x_i\beta$. An ordinary Gibbs sampling procedure to draw samples from (6) is to alternatively draw samples from the conditional posterior of λ given β with a log density equal to the sum of the last two terms of (6), and draw samples from the conditional posterior of β given λ with a log density equal to the sum of the first two terms of (6).

The challenge in sampling from the (6) comes from two aspects of high-dimensional features. One is the high dimension p of β (or X); the other is the high correlation among features X, which results in the high correlation in the conditional posterior of β given λ, and correspondingly the multi-modality in the marginal posterior of β (with λ integrated out). To combat these two difficulties, we propose an MCMC sampling algorithm that uses Gibbs sampling with Hamiltonian Monte Carlo (HMC) for sampling β in a restricted way. Our MCMC algorithm is sketched below and followed with explanations:

Starting from a previous state for (β, λ), a new state denoted by (β, λ) is obtained with these steps:

1. For each j, draw a new $\hat{\lambda}_j$ from the conditional distribution $f(\lambda_j|\beta_j)$ with log PDF equal to the sum of (8) and (9). It is well-known that λ_j given β_j has an Inverse-Gamma distribution given as follows:

$$\lambda_j|\beta_j \sim \text{Inverse} - Gamma\left(\frac{\alpha_1+1}{2}, \frac{\alpha_1\omega_1+\beta_j^2}{2}\right). \tag{10}$$

2. With the new values of $\hat{\lambda}_j$ drawn in step 1, determine a subset, β_U, of β to update in step 3 below. We update β_j if $\hat{\lambda}_j$ is large enough. That is, given a pre-scribed threshold value η, the subset is defined as $U = \{j|\hat{\lambda}_j > \eta\}$. The β_U is defined as $\{\beta_j|j \in U\}$. The subset of $\beta_F = \{\beta_j|j \in F = \{0, \ldots, p\}\backslash U\}$ will be kept unchanged in step 3.
3. Update the set of β_j with $j \in U$, denoted by β_U, by applying HMC to the conditional distribution of β_U given as follows:

$$\log(f(\beta_U|\beta_F, \lambda, X, y))$$

$$=\sum_{i=1}^{n} lp(y_i|x_{i,U}\beta_U + x_{i,F}\beta_F) + \sum_{j \in U}\log(f(\beta_j|\hat{\lambda}_j)) + C_3, \tag{11}$$

where the function lp for computing log likelihood is defined in (7), and $x_{i,U}$ is the subset of x_i with feature index in U. After updating β_U, the new value of β is denoted by β in which β_F does not change. Note that, because HMC is a Metropolis algorithm, the new β may be equal to β if a rejection occurs.
4. Set $(\beta, \lambda) = (\beta, \lambda)$, and go back to step 1 for the next iteration.

A typical sampling method for classification models is to augment a latent continuous value z_i for each categorical variable y_i[47], and sample from the joint distribution of $z_{1:n}$ along with β and λ (e.g.[38]) with Gibbs sampling; we then can borrow algorithms developed for regression models with heavy-tailed priors[37,43]. Given λ_j, the prior for β_j is a normal distribution. It is well-known that the posterior of β for normal regression given normal priors is a multivariate normal distribution with a covariance matrix involving $X' X$. Note that this multivariate normal has a dimension p. When p is very large (e.g. thousands), drawing independent samples from a multivariate normal is extremely inefficient, because the required computation time for decomposing the covariance matrix will increase in the order of p^3. Therefore, for drawing samples from $f(\beta| \lambda, X, y)$, we choose to use Hamiltonian Monte Carlo (HMC), a special case of Metropolis-Hasting (M-H) algorithms, which explore the posterior in a local fashion without the need to decompose a high-dimensional matrix. HMC requires computing the log-posterior and its gradient. The gradient of $\log(f(\beta| \lambda, X, y))$ given by the following expression:

$$\frac{\partial \mathscr{U}}{\partial \beta_j} = \sum_{i=1}^{n}\left[\frac{\Gamma\left(\frac{\alpha_0+1}{2}\right)}{\sqrt{\alpha_0\pi}\ \ \Gamma\left(\frac{\alpha_0}{2}\right)} \times \frac{x_{ij}}{\sqrt{\omega_0}} \times \frac{\left(1 + \frac{(x_{i,U}\,\beta_U + x_{i,F}\,\beta_F)^2}{\alpha_0\omega_0}\right)^{-\frac{\alpha_0+1}{2}}}{1 - y_i - T_{\alpha_0,\omega_0}\left(x_{\,i,U}\beta_U + x_{i,F}\,\beta_F\right)}\right] + \frac{\beta_j}{\hat{\lambda}_j}, \tag{12}$$

where $\mathscr{U}$ is the function defined in (11). We can see that once the linear combination $X\beta$ has been computed, the log posterior and its gradient can be obtained with very little computation. Computing $X\beta$ is significantly cheaper than decomposing a matrix of dimension p. However, the random-walk behaviour of ordinary M-H algorithms limits the sampling efficiency of M-H algorithms. In HMC, the gradient of log posterior is used to construct a trajectory along the least constraint direction, therefore, the end point of the trajectory is distant from the starting point, but has high probability to be accepted; for more discussions of HMC, one is referred to a review paper by[42].

From the above discussion, we see that obtaining the value of $X\beta$ is the primary computation in implementing HMC. To further accelerate the computation for very large p, we introduce a trick called restricted Gibbs sampling; this is inspired by the fact that updating the coefficients with small λ_j (small prior variance in the conditional posterior of β_j given λ) in HMC does not change the likelihood as much as updating the coefficients with large λ_j but updating β_j with small or large λ_j consumes the same time. Therefore, we use λ in step 2 to select only a subset of β, denoted by β_U, those have large prior variance λ_j, to update in step 3 (HMC updating). We can save a great deal of time for computing $X\beta$ in step 3 by caching values of $X_F\beta_F$ from the previous iteration because it does not change in the whole step 3; this greatly accelerates the construction of HMC trajectory. We typically choose η in step 2 so that only 10% of β are updated in step 3.

We clarify that although β_F (sometimes the whole β) are kept the same in an iteration, the choice of U in step 2 for the next iteration will be updated because λ will be updated in step 1. Thus, β_j will not get stuck to a very small absolute value, unlike that in optimization algorithms this typically occurs.

The above restricted Gibbs sampling is a valid Markov chain transition for the joint posterior (6). To understand this, let us recall that, in Gibbs sampling we can arbitrarily choose any variables to update with a Markov chain transition that leaves the conditional distribution of chosen variables invariant, provided that the choice of variables to be updated does not depend on the *values* of the chosen variables in the previous iteration. For example, it is not a valid Markov chain transition if we choose β_j with large $|\beta_j|$ in the previous iterations; by contrast, it is a valid Markov chain transition if we choose β_j to update by referring to variances of β. In step 3, the choice of β_U does not depend on the values of β in the previous step. Instead, the choice only depends on the value of λ_j in the previous step, which partially determines the variances of β in $f(\beta|\hat{\lambda}, X, y)$. Therefore, the updates of β_U in step 3 is reversible with respect to $f(\beta|\hat{\lambda}, X, y)$.

The advantage of HMC is that it can explore highly correlated posterior quickly with a long leapfrog trajectory without suffering from the random-walk problem. This ability of HMC also plays an important role in travelling quickly between multiple modes of the posterior. This is explained as follows. When $\hat{\lambda}_j$ and $\hat{\lambda}_k$ for two correlated features j and k are large after a draw in step 1, the joint conditional posterior of (β_j, β_k) given $\left(\hat{\lambda}_j, \hat{\lambda}_k\right)$ are highly

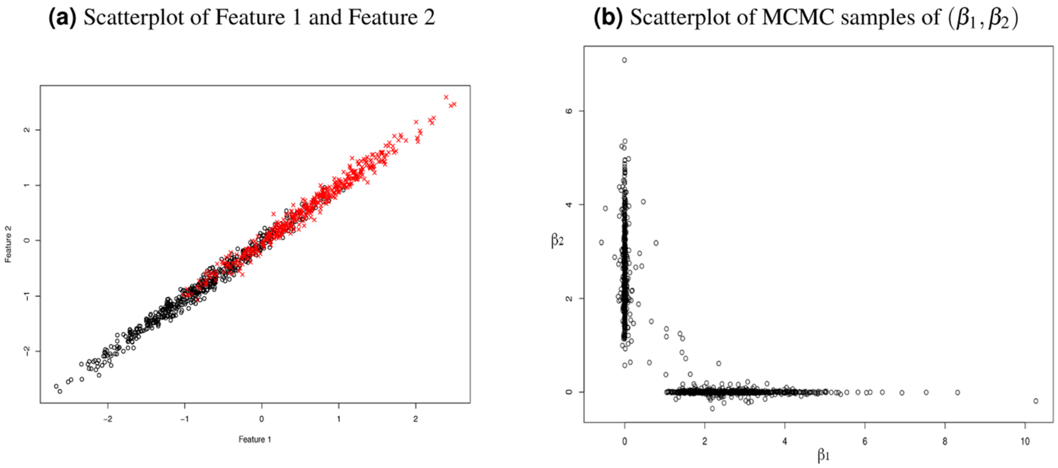


Figure 1. Demonstration of within-group selection with two correlated features for binary response. Color denotes the response value for each case. Note that the two features together do not provide significantly more information than only one for classifying the response.

negatively-correlated. For such distributions, HMC can move more quickly than random-walk algorithms along the least constrained direction, and this move will lead to the change of modes in joint distribution of (β_j, β_k) with λ integrated out.

There are a huge number of modes in the posterior even when p is moderate when there are a large number of correlated features. In the empirical studies reported in this paper, we use a two-stage procedure. In **Stage 1**, we run the restricted Gibbs sampling with HMC using the dataset containing all p features. Then we calculate MCMC means of all coefficients β and choose only the top $p^* = 100$ features with largest absolute values of MCMC means. The stage 1 is very time consuming. In **Stage 2** we re-run the MCMC sampling with only the selected features once again. Our feature selection will be based on the MCMC samples obtained from Stage 2. A list of setting parameters with recommended values for ease in reference are given in the Supplementary Information.

Feature subset selection: MCMC sample subdivision. We run the MCMC sampling to obtain samples from the posterior of β. With the intercept β_0 removed, this sample is denoted by a matrix $B = (\beta_{j,i})_{p\times R}$, in which $\beta_{j,i}$ represents the value of β_j in the i th sample and R is the number of MCMC samples. The posteriors of β for Robit models with heavy-tailed priors are severely multi-modal. For a demonstration, one can look at Fig. 1, which shows a scatterplot of MCMC samples of two β_j's for two correlated features. Therefore, we should divide the whole Markov Chain samples B into sub-pools according to the mode that each sample represent. However, the number of such feature subsets may be huge even the number of features p is small. Therefore, we only consider dividing Markov Chain samples according to the multiple modes for the Markov Chain samples obtained in **Stage 2** in which a large number of weakly related features have been omitted. In this article, we use a scheme that looks at the relative magnitude of β_j to the largest value in all features. The scheme is rather ad-hoc. However, it is very fast and works well in problems of moderate dimension, such as $p = 100$. More advanced methods for collecting feature subsets from MCMC is our priority for future research. The scheme used in this article is described as follows:

1. We set $I_{j,i} = 1$ if $|\beta_{j,i}| > 0.1 \times \max\{|\beta_{1,i}|, \ldots, |\beta_{p,i}|\}$, and $I_{j,i} = 0$ otherwise. By this way, we obtain a boolean matrix $(I_{j,i})_{p\times R}$ with its entry $I_{j,i}$ denotes whether the j th feature is selected or not in i th sample.
2. Discard the features with overall low frequency in step 1. We calculate $f_j = \frac{1}{R}\sum_{i=1}^{R} I_{j,i}$. We will discard a feature j if f_j is smaller than a pre-defined threshold, which is set to be 5% as an ad-hoc choice in this article. Let $D = \{j|f_j < 5\%\}$. For each $j \in D$, we set $I_{j,i} = 0$ for all $i = 1, \ldots, R$. This step is to remove the features that come into selection in step 1 due to MCMC randomness.
3. Find a list of feature subset by looking at the column vectors of I. Each unique column in I represents a different feature subset.

The above algorithm is not the best for dividing the MCMC samples according to the posterior modes of β. The reason is that the MCMC simulation introduces small jitters into the β_j's of the features not selected in the mode. The above algorithm aims to get rid of the jitters by using thresholding in step 1 and step 2. However, they may not eliminate some jitters. This will result in some feature subsets with very small frequency. The optimal algorithm may be to find the posterior modes starting from each MCMC sample using a certain optimization algorithm. However, finding the posterior modes for a large number of MCMC samples is time-consuming. In this paper, we present the results based on the above fast algorithm for simplicity. From our informal comparisons, the results are fairly close to the feature subsets found by hunting the mode from each MCMC sample.

Predictive performance metrics. The frequencies of feature subsets in MCMC samples may not exactly reflect their predictive power. In this paper, we evaluate the predictive power of each feature subset using leave-one-out cross-validation (LOOCV) using the same training cases for simulating MCMC. Specifically, for each collected feature subset, we apply logistic regression model with *t* penalty[8] and evaluate its predictive performance using 4 criteria: **error rate**, average of minus log predictive probabilities (**AMLP**), the area under ROC (receiver operating characteristic) curve (**AUROC**) and the area under precision-Recall curve (**AUPRC**). AMLP is calculated at each actually observed y_i: $\frac{1}{n}\sum_{i=1}^{n} - \log(\hat{P}_i(y_i))$, and it punishes heavily the small predictive probabilities at the true class labels. We use an R-package pROC[48] to compute AUROC and AUPRC. We propose three prediction methods based on BayesHL MCMC sampling results. The prediction methods based on the *top* feature subset with the highest posterior frequency, and the *optimal* feature subset (with the smallest cross-validated AMLP) are referred, respectively, as **BayesHLtop** and **BayesHLopt**. We will refer to the result by averaging the predictive likelihood over MCMC samples as the default **BayesHL** method.

We prefer AMLP and AUPRC as better metrics (than AUROC and error rate) for evaluating model performance in our study. In fact, AUROC values are not informative when (i) data is imbalanced with few cases of the minority class[49] (e.g. high risk patients) or (ii) the cost of misclassifying the minor class (high risk patients) is more of concern[50]. Under scenario (i), AUPRC is usually preferred over AUROC because the latter can be misleading in such cases and provide deceptively optimistic results[51]. AMLP (i.e. estimate of entropy[52,53]) is a better choice than AUROC under both scenarios (i) and (ii), since the former incorporates "certainty" of classification directly in the calculation. In general, AMLP is a more scalable metric than AUROC and AUPRC. It also penalizes severely misclassifications and thus favors more robust methods. In contrast, these misclassifications have less influence on AUROC and AUPRC, but may carry significant costs in applications.

Overview: BayesHL. In summary, we proposed a heavy-tailed Robit model with heavy-tailed Cauchy priors for coefficients β to explore the potential useful posterior modes. The model (β) is then estimated using our adapted restricted Gibbs sampling method with Hamiltonian Monte Carlo technique. Finally, multiple feature subsets, corresponding to multiple posterior modes, were collected from the MCMC samples and used to perform classification problems. This method is referred to as the Bayes Hyper Lasso method **BayesHL**. The prediction performance of **BayesHL** will be tested and compared to other methods in simulation analysis and a gene expression data analysis.

Simulation Studies

An example with independent groups of features. In this section, we compare BayesHL with other existing feature selection methods on simulated datasets with independent groups of features. Each dataset has $p = 2000$ features and $n = 1200$ cases, 200 of which are used as training cases and the other 1000 cases are used as test cases. With z_{ij}, ε_{ij}, e_i generated from $N(0, 1)$, we generate the feature values x_{ij} for $i = 1, ..., n, j = 1, ..., p$ in four groups and the class label y_i as follows:

$$x_{il} = z_{i1} + 0.5\varepsilon_{il},\ i = 1, \ldots, n,\ l = 1, \ldots, 50,\ (\textbf{Group 1}) \tag{13}$$

$$x_{im} = z_{i2} + 0.5\varepsilon_{im},\ i = 1, \ldots, n,\ m = 51, \ldots, 100,\ (\textbf{Group 2}) \tag{14}$$

$$x_{ik} = z_{i3} + 0.5\varepsilon_{ik},\ i = 1, \ldots, n,\ k = 101, \ldots, 150,\ (\textbf{Group 3}) \tag{15}$$

$$x_{ij} \sim N(0,1), i = 1, \ldots, n,\ j = 151, \ldots, 2000,\ \left(\textbf{Group 4}\right) \tag{16}$$

$$y_i = 1 \text{ if } (z_{i1} + z_{i2} + z_{i3})/\sqrt{3} + 0.1e_i > 0;\ = 0\,\text{otherwise}. \tag{17}$$

The z_{i1}, z_{i2} and z_{i3} are common factors for features in respective group. Since the features within each of Group 1–3 are related to a common factor, they are highly correlated. However, the features across groups are independent. The response y_i is generated with the values of the common factors z_{i1}, z_{i2}, z_{i3}. Therefore, y_i is related to all the features in Group 1–3. The y_i is unrelated to all the features in Group 4. The true model of y_i given x_{ij} has non-zero coefficients for all features in Group 1–3.

We apply BayesHL and other methods including LASSO, Group LASSO (GL), supervised Group LASSO (SGL), Random Forest (RF), Penalized Logistic Regression (PLR) with hyper-LASSO penalty, neural network (NN[54]), eXtreme Gradient Boosting (XGBoost[55]), and knockoff variable selection method (Knockoff[4]) to fit the training cases and then test their performance with the 1000 test cases. The implementation details of all the competitors can be found in the supplement. BayesHL is conducted with the default parameter settings as listed in supplement. We run BayesHL first with all 2000 features in **stage 1**, and then rerun with $p^* = 100$ top features selected with posterior means, both with the aforementioned settings. The feature selection and prediction use the MCMC samples in the **stage 2** with the top 100 features. Because of the large p in stage 1, we ran BayesHL hours to ensure convergence. We allowed BayesHL to run about 30 minutes to obtain the results reported throughout this article.

Table 2 shows the top (by frequency) five feature subsets selected by BayesHL. According to the AMLP, the top feature subset (1,57,140) is identified as the optimal feature subset too. We see that the top 4 feature subsets selected by BayesHL contain exactly one feature from each of Group 1–3 (each with 50 features) and none from Group 4 (noise).

	fsubsets	freqs	AMLP	ER	AUROC	AUPRC
1	1,57,140	0.22	0.13	0.09	0.99	0.99
2	1,51,140	0.11	0.13	0.08	0.99	0.99
3	16,57,140	0.10	0.14	0.08	0.99	0.99
4	1,51,101	0.09	0.14	0.08	0.99	0.99
5	12,57	0.04	0.41	0.39	0.89	0.90

Table 2. Top 5 feature subsets selected by BayesHL, and their within-sample leave-one-out cross-validatory predictive power. "fsubsets" gives I.D. of features in each subset, "coefs" is the vector of regression coefficients found with the posterior means, "AMLP" - "AUPRC" are cross-validatory predictive power measures of each feature subset.

(a) Numbers of selected features in respective group											
	BayesHLtop	**BayesHLopt**	**BayesHL**	**LASSO**	**GL**	**SGL**	**RF**	**PLR**	**NN**	**XGBoost**	**Knockoff**
Group 1	1	1	4	6	49	7	49	50	50	32	49
Group 2	1	1	4	5	50	10	49	50	50	0	50
Group 3	1	1	3	6	50	6	48	50	50	0	50
Group 4	0	0	0	13	341	12	14	1305	1252	0	16
Total	3	3	11	30	490	35	160	1455	1402	32	165
(b) Out-of-sample predictive performance											
ER	0.10	0.10	0.06	0.09	0.07	0.10	0.08	0.08	0.10	0.34	0.10
AMLP	0.22	0.22	0.15	0.21	0.22	0.24	0.38	0.18	0.20	0.63	0.31
AUROC	0.97	0.97	0.99	0.97	0.99	0.97	0.98	0.98	0.98	0.75	0.97
AUPRC	0.97	0.98	0.99	0.96	0.98	0.98	0.98	0.97	0.98	0.75	0.97

Table 3. Comparison of feature selection and out-of-sample prediction performance of different methods on a dataset with independent group of features. The number of features used by the others other than BayesHL are counted after thresholding the absolute coefficients by 0.1 times the maximum. BayesHLopt: optimal feature subset from BayesHL. BayesHLtop: top feature subset from BayesHL. BayesHL: average prediction probability across feature subsets identified by BayesHL. RF: random forest. NN: neural network. XGboost: eXtreme Gradient Boosting. Knockoff: knock off variable selection followed with logistic regression. AMLP: average minus log-probabilities. AUROC: area under ROC. AUPRC: area under precision-recall curve.

We also compare the out-of-sample predictive power of the top and optimal feature subset found by BayesHL with the "complete" feature subsets selected by other methods. We compare 4 predictive measures (ER, AMLP, AUROC, AUPRC) against the number of features used in making predictions, as shown in Table 3. The numbers of features used in BayesHLtop and BayesHLopt are the number of features in the top and optimal subsets. To count the number of features selected by the methods other than BayesHL, if automatic sparse selection is not available, we threshold their absolute coefficients by 0.1 of the maximum to decide whether or not they are used in predictions. We choose 0.1 as a threshold because we use the same thresholding to obtain the top and and optimal feature subsets of BayesHL.

Table 3, demonstrates that the BayesHL has the best predictive performance, which is better than the best performer in non-BayesHL methods—Group LASSO with more than 490 features. Specifically, BayesHL achieved the lowest ER (0.06), lowest AMLP (0.15), highest AUROC (0.99) and AUPRC (0.99) among all methods. The values of ER, AUPRC and AUROC stay at fairly similar levels for all methods except XGBoost. Only AMLP values (i.e. cross-entropy estimates[52,53]) are substantially different across methods and BayesHL outperforms others on this metric. BayesHLtop and BayesHLopt have slightly worse predictive performance than non-BayesHL methods, however, they use only 3 features, one from each signal group. In terms of efficiency in selecting useful features, BayesHLtop and BayesHLopt do the best jobs if we look at the ratio of predictive measure to number of used features. In comparison, other methods are all less sparse and select much larger subsets from noise group (Group 4). Particularly, Group LASSO enforces the similarity of coefficients in each group, therefore, all the features in signal groups along with a large number (341) of noise features are selected.

An example with correlated weakly differentiated features.

In this section we will compare the performance of BayesHL in a simulated scenario such that two groups of features are weakly differentiated but have a strong joint effect on the response. Specifically, a dataset with $n = 1200$ cases and $p = 2000$ features is generated as follows:

$$P(y_i = c) = \frac{1}{2}, \quad \text{for} \quad c = 1,2, \tag{18}$$

(a) Numbers of selected features in respective group											
	BayesHLtop	**BayesHLopt**	**BayesHL**	**LASSO**	**GL**	**SGL**	**RF**	**PLR**	**NN**	**XGboost**	**Knockoff**
Group 1	1	1	6	3	155	4	3	172	200	136	36
Group 2	1	1	8	3	123	5	5	177	200	0	9
Group 3	1	1	6	7	176	12	102	192	200	0	170
Group 4	0	0	1	10	215	22	3	1020	1259	0	17
Total	4	3	21	23	669	43	113	1561	1859	136	232
(b) Out-of-sample predictive performance											
ER	0.18	0.17	0.12	0.14	0.10	0.16	0.15	0.10	0.10	0.14	0.43
AMLP	0.47	0.48	0.33	0.34	0.25	0.46	0.37	0.26	0.31	0.35	16.88
AUROC	0.90	0.91	0.95	0.93	0.97	0.92	0.93	0.95	0.92	0.92	0.55
AUPRC	0.91	0.92	0.94	0.93	0.96	0.95	0.93	0.96	0.92	0.92	0.55

Table 4. Comparison of feature selection and out-of-sample prediction performance of different methods on a simulated dataset with correlated weakly differentiated features. BayesHLopt: optimal feature subset from BayesHL. BayesHLtop: top feature subset from BayesHL. BayesHL: average prediction probability across feature subsets identified by BayesHL. RF: random forest. NN: neural network. XGboost: eXtreme Gradient Boosting. Knockoff: knock off variable selection followed with logistic regression. AMLP: average minus log-probabilities. AUROC: area under ROC. AUPRC: area under precision-recall curve.

$$x_{ij} = \mu_{y_i,1} + z_{i1} + 0.5\varepsilon_{ij}, \quad \text{for } j = 1, \ldots, 200, \ (\textbf{Group 1}) \tag{19}$$

$$x_{ij} = \mu_{y_i,2} + 0.8z_{i1} + 0.6z_{i2} + 0.5\varepsilon_{ij}, \quad \text{for } j = 201, \ldots, 400, \ (\textbf{Group 2}) \tag{20}$$

$$x_{ij} = \mu_{y_i,3} + z_{i3} + 0.5\varepsilon_{ij}, \quad \text{for } j = 401, \ldots, 600, \ (\textbf{Group 3}) \tag{21}$$

$$x_{ij} \sim N(0,1), \quad \text{for } j = 601, \ldots, 2000, \ (\textbf{Group 4}) \tag{22}$$

where z_{ij} and ε_{ij} are from $N(0, 1)$, and the means of features in Group 1–3 in two classes are given by the following matrix $\mu_{c,1:3}$, where $\mu_{1,1:3} = (-0.3,0.3,1)$ and $\mu_{2,1:3} = (0.3, -0.3, -1)$.

A dataset generated as above has 200 features in each of Group 1–3 related to the response and the remaining 1400 are completely noisy. Each feature in Group 1 and Group 2 is weakly differentiated due to the small difference in class means (0.3 vs −0.3). The features within each group are positively correlated with correlation 0.8. Additionally, a feature from Group 1 and a feature from Group 2 has a correlation coefficient 0.64 because they share a common factor z_{i1}. Therefore, a combination of two features from Group 1 and 2 respectively has clear joint effect for y_i.

We run BayesHL and other methods to a dataset generated as above and use 1000 test cases to compare the out-of-sample predictive power of the top and also optimal feature subset with the "complete" feature subsets found by other methods using the same procedure for obtaining Table 4. Table 4 shows BayesHL methods have slightly worse predictive performance than Group LASSO and PLR, which successfully combine the power of all signal features from Group 1–3 to make better predictions. However, the ROC curves of Group LASSO (AUROC = 0.97) and BayesHL (AUROC = 0.95) are not significantly different under Delong's two-sided test[56] (p-value = 0.47). Moreover, the feature subsets selected by Group LASSO and PLR are significantly less sparse and include many noise features in Group 4, while BayesHL, LASSO, SGL and RF have more sparse results. Particularly, BayesHL selected 6, 8, and 6 features respectively from each of the three signal groups, which demonstrated the clear strength of our method in sparsity and interpretability. In conclusion, BayesHL delivered similarly good prediction performance as its competitors, while using much more sparse feature subsets.

Endometrial Cancer RNA-Seq Data Analysis

Endometrial cancer (EC) starts in the cells of the inner lining (endometrium) of the uterus. It is one of the most common cancers of the female reproductive system and is particularly common in women over age 60.

We chose to analyze the endometrial cancers from The Cancer Genome Atlas (TCGA) Research Network (http://cancergenome.nih.gov/) since there is a large number of tumour samples with matched gene expression profiles and clinical information. The original TCGA-EC dataset contains around 500 samples of EC with matched gene expression profiles and clinical information. This is one of the largest samples in TCGA's database.

We obtained TCGA data from the Broad GDAC Firehose (using bioconductor Rpackages TCGAbiolinks and TCGA2STAT), which includes $N = 269$ samples with matched RNASeq profile and clinical information, after filtering steps such as restricting to primary solid tumor as sample type and endometrioid endometrial adenocarcinoma as histological type. We further filtered 3 patients with missing radiation information. The $N =$ 266 matched RNASeq profiles were downloaded in RPKM (Reads Per Kilobase Million)-normalized format and log2-transformed. We then performed univariate feature selection and retained $P = 7298$ (out of 20502) genes with high values of coefficient of variation ($\geq$5), or high values of mean expression of log2 RPKM ($\geq$3).

	BayesHL	LASSO	GL	SGL	RF	PLR	NN	XGBoost	knockoff
False Predictions	29	35	33	32	34	49	55	37	39
AMLP	0.30	0.34	0.31	0.32	Inf	0.98	Inf	0.34	0.60
AUROC	0.78	0.72	0.80	0.79	0.73	0.65	0.70	0.77	0.70
AUPRC	0.54	0.31	0.41	0.42	0.36	0.25	0.29	0.35	0.32

Table 5. Comparison of cross-validated predictive performance on endometrial cancer data (N = 266) using all methods. AMLP: average minus log probabilities. AUROC: area under ROC, AUPRC: area under precision-recall curve. GL: Group LASSO, SGL: supervised Group LASSO, RF: Random Forest, PLR: Penalized Logistic Regression with a hyper-LASSO penalty, NN: neural network, XGBoost: eXtreme Gradient Boosting, knockoff: knockoff variable selection followed with logistic regression.

The rest of this section is organized as follows: we first compare the predictive performance of our algorithm vs. the competitors with a classification analysis on 5-year EC survival outcome, then introduce and discuss the results of survival analysis and pathway analysis. Finally, we discuss possible biological explanations for the results of these analyses.

Comparison of the algorithms' results via classification analysis. We applied all methods to the dataset ($n = 266$ samples, $p = 7298$ features), with leave-one-out cross-validation, to perform binary classification for 5-year overall survival (OS) outcome of patients ($\mathbf{Y}$). Specifically, classification methods were trained on each leave-one-out training set (265 samples) and then used to predict on the leave-out test case. Covariates such as age, diagnosis year and radiation therapy are also included in the model. The cross-validated prediction probabilities across folds were then collected to evaluate the overall performance for all methods. For the purpose of comparison, we performed the same steps detailed above with all algorithms: LASSO, Group LASSO (GL), supervised Group LASSO (SGL), Random Forest (RF), Penalized Logistic Regression (PLR) with a hyper-LASSO penalty, neural network (NN), eXtreme Gradient Boosting (XGBoost), knockoff (knockoff) variable selection, and BayesHL. Note that we report the prediction probabilities from BayesHL as explained in the method section. The implementation details of other methods can be found in the supplement.

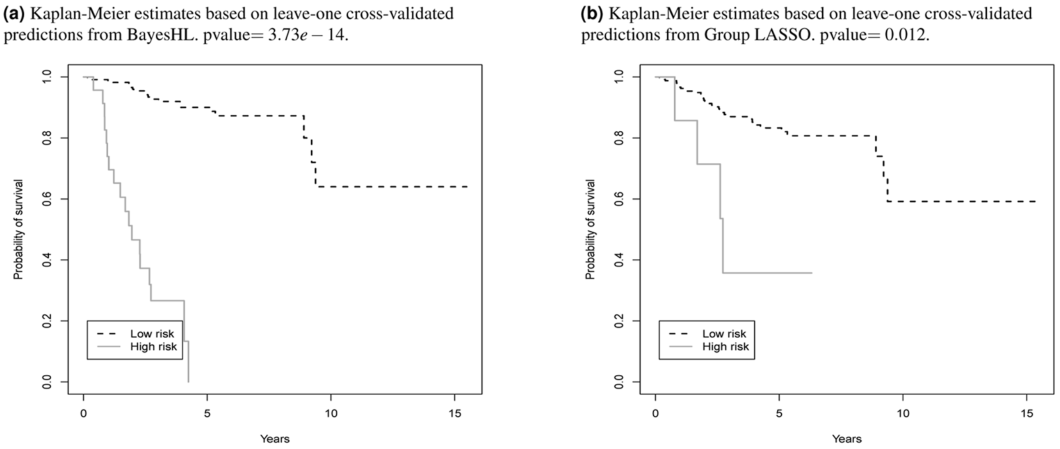
Table 5 presents the results of our analysis. As can be seen, BayesHL had the best performance out of all the classifiers with respect to three of the four measures (false predictions, AMLP, and area under precision-recall curve (AUPRC)). GL demonstrated slightly better predictive performance than BayesHL with respect to area under ROC (AUROC). However, the differences of ROC curves from BayesHL and Group LASSO are not statistically significant under Delong's two-sided test[56] (p-value = 0.79). The AUPRC values are substantially different across methods and BayesHL provide the optimal result. BayesHL also have the optimal AMLP values. Note that the infinite AMLP values from the Random Forest and Neural Network methods imply that there exist extreme misclassifications. Finally, both Group LASSO and SGL performed relatively better than the rest of methods (except BayesHL), suggesting potential existence of grouping structures among genes. In conclusion, we prefer AMLP and AUPRC (over AUROC and ER) to evaluate the performance of all methods for this imbalanced classification problem (12.8% high risk patients), and under both metrics our BayesHL method clearly demonstrated better prediction power by using more succinct feature subset selections.

Comparison via survival analysis. Next, we compared the algorithms' results by performing a Kaplan-Meier correlation. The 5-year OS for each patient was calculated using leave-one-out cross-validation. Patients were classified as either high-risk or low-risk according to the minimum survival probability, with 0.5 as the cutoff threshold. The Kaplan-Meier curve in Fig. 2 shows that patients in the high-risk group had significantly lower 5-year OS than those in the low-risk group (log-rank test $p = 3.73e{-}14$) based on the BayesHL predictions. By comparison, Group LASSO did not achieve statistical significance $p = 0.012$ in the same Kaplan-Meier test. Note that we only compare the results of BayesHL with Group LASSO here because of their superior classification performance in Table 5.

Comparison via IPA pathway analysis. Finally, we conducted a pathway analysis of the identified gene signatures using the Ingenuity Pathways Analysis (IPA) system[57]. The resulting network is derived from the Ingenuity Knowledge Base and indicates which roles the input genes play in a global molecular network. The purpose of the IPA analysis is to see whether the genes output by our method and the competitors are related to potentially cancer-linked subnetworks. Figure 3a shows the results of our IPA analysis for the BayesHL-selected features on the whole dataset. As can be seen, the BayesHL-selected features are related to eight subnetworks, one of which is clearly linked to a cancer outcome ("Cancer, organismal injury and abnormalities, and tumor morphology").

Explanation of the results. In this section, we examine the genes selected by BayesHL and explore why some of these genes might be linked to EC survival outcomes. Table 6 provides a list of the BayesHL-selected feature subsets, along with measures of their leave-one-out cross-validated predictive performance on the whole dataset.

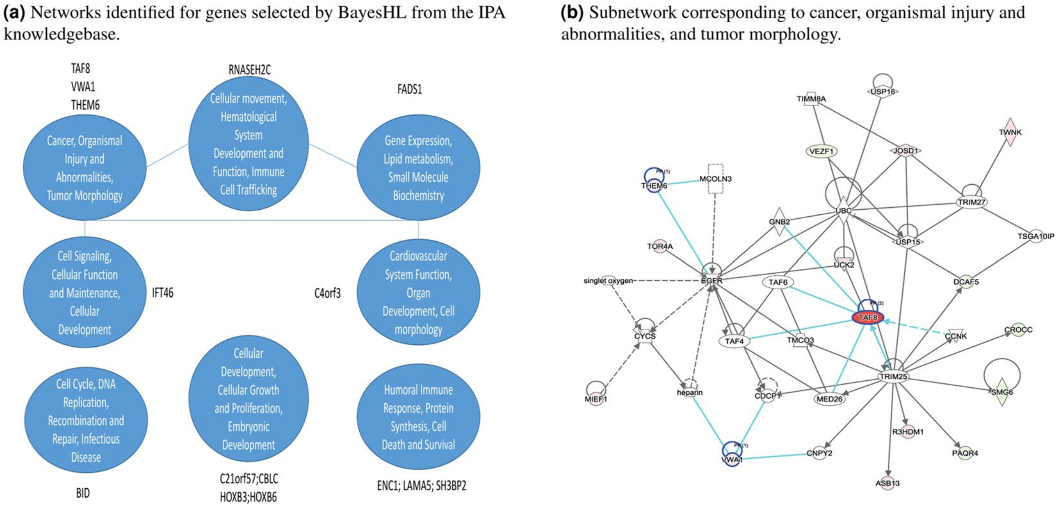
We observe that genes HOXB3 and SH3BP2 both appear twice in Table 6. This is supporting evidence for our claim of BayesHL's effectiveness, because these genes are known to be prognostic markers for endometrial cancer. HOXB3 can induce the transformation and proliferation of tumor cells in breast cancer[58] and ovarian cancer[59]. It has recently been tested as a regulation target to control endometrial cancer[60]. The protein encoded by SH3BP2 functions in the cell that signals various immune response pathways[61]. This information, combined with



(a) Kaplan-Meier estimates based on leave-one cross-validated predictions from BayesHL. pvalue= $3.73e-14$.

(b) Kaplan-Meier estimates based on leave-one cross-validated predictions from Group LASSO. pvalue= 0.012.

Figure 2. Kaplan-Meier estimates of the 5-year OS for all (266) EC patients, according to the cross-validated prediction probabilities from BayesHL and Group LASSO.



(a) Networks identified for genes selected by BayesHL from the IPA knowledgebase.

(b) Subnetwork corresponding to cancer, organismal injury and abnormalities, and tumor morphology.

TAF8
VWA1
THEM6
Cancer, Organismal Injury and Abnormalities, Tumor Morphology
RNASEH2C
Cellular movement, Hematological System Development and Function, Immune Cell Trafficking
FADS1
Gene Expression, Lipid metabolism, Small Molecule Biochemistry
Cell Signaling, Cellular Function and Maintenance, Cellular Development
IFT46
C4orf3
Cardiovascular System Function, Organ Development, Cell morphology
Cell Cycle, DNA Replication, Recombination and Repair, Infectious Disease
Cellular Development, Cellular Growth and Proliferation, Embryonic Development
Humoral Immune Response, Protein Synthesis, Cell Death and Survival
BID
C21orf57;CBLC
HOXB3;HOXB6
ENC1; LAMA5; SH3BP2
USP16
TIMM8A
TWNK
VEZF1
JOSD1
THEM6
MCOLN3
UBC
TRIM27
GNB2
USP15
TSGA10IP
TOR4A
UCK2
singlet oxygen
EGFR
TAF6
DCAF5
TAF8
CYCS
TAF4
TMCO3
CCNK
CROCC
TRIM25
MIEF1
heparin
CDCP1
MED26
SMG6
R3HDM1
VWA1
CNPY2
PAQR4
ASB13

Figure 3. Molecular network information of the genes selected by BayesHL from the IPA knowledgebase. (**a**) All networks identified for genes selected by BayesHL. (**b**) Subnetwork corresponding to cancer, organismal injury and abnormalities, and tumor morphology. Arrows with solid lines represent direct interactions and arrows with broken lines represent indirect interactions. Direction of the arrows represents causal effects from upstream to downstream or protein self-bindings. The shapes of blocks correspond to different classes of general molecular functions in IPA knowledgebase. The color inside each block reflects the (averaged) gene coefficients from BayesHL model. Red indicates that the expression of the gene has negative impact on survival outcome and cyan indicates positive impact. White denotes no impact. The blocks with blue circle and green edges denote genes that occur in the selected 19 feature subsets from BayesHL. Both figures were generated through the use of IPA[57] software (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis).

our finding (that SH3BP2 might be a survival-related gene for EC patients) suggest that EC cells may induce an abnormal immune system response that leads to a worse survival outcome. For example, we know that SH3BP2 protein helps to regulate signaling pathways that activate B cells and macrophages, whose infiltration in necrosis in the tumor center (hot-spot tumor associated macrophages) is a hazard factor to relapse-free survival of endometrial cancer patients[62].

From Table 6, we also observe that the gene TAF8 appears numerous times. Moreover, TAF8 plays a central role in the cancer-related subnetwork we found during our IPA analysis (see Fig. 3b below for a detailed view of how BHTF-selected genes fit into the cancer and morphology subnetwork). Specifically, BayesHL assigns gene TAF8 the highest coefficient value, and the IPA analysis located TAF8 in one of the central positions in the cancer and morphology subnetwork (highlighted in red in Fig. 3b). It is known that TAFs contribute to the differentiation and proliferation of cells, and several TAFs (including TAF2, TAF4B, TAF9) have been identified as tumor promoters or suppressors in ovarian cancer[63,64]. However, more research must be done before we can establish a link between TAF8 and EC; and as of yet, no study exists on that topic.

	frequency	Genes	False Prediction	AMLP	AUROC	AUPRC
1	0.1	FADS1, TAF8	28	0.33	0.73	0.45
2	0.08	HOXB6, TAF8	29	0.32	0.75	0.44
3	0.07	CBLC, TAF8	31	0.34	0.72	0.42
4	0.07	SH3BP2, TAF8	32	0.33	0.75	0.45
5	0.07	C4orf3, TAF8	28	0.30	0.78	0.51
6	0.06	BID, TAF8	31	0.33	0.76	0.42
7	0.06	C21orf57, TAF8	30	0.32	0.76	0.43
8	0.06	FADS1	34	0.39	1.00	0.07
9	0.05	SH3BP2	35	0.37	0.66	0.28
10	0.05	C8orf55, ENC1	34	0.37	0.63	0.24
11	0.05	IFT46, TAF8	33	0.31	0.81	0.45
12	0.04	HOXB3, TAF8	28	0.32	0.75	0.48
13	0.04	C8orf55	34	0.37	0.64	0.17
14	0.04	RNASEH2C, TAF8	30	0.34	0.72	0.43
15	0.04	LAMA5	33	0.36	0.65	0.25
16	0.04	C4orf3, HOXB3, TAF8	25	0.29	0.81	0.56
17	0.03	TAF8, VWA1	32	0.32	0.75	0.45
18	0.03	C21orf57	33	0.36	0.67	0.29
19	0.03	C4orf3, FADS1, TAF8	27	0.31	0.77	0.51

Table 6. BayesHL-selected feature subsets and the corresponding cross-validated prediction performance on the endometrial cancer dataset (N = 266). AMLP: average minus log probabilities. AUROC: area under ROC, AUPRC: area under precision-recall curve.

In conclusion, these results suggest a potential central role of TAF8, VWA1 and THEM6 in the endometrial cancer development and survival outcome, by their repeated occurrence in most of feature subsets identified by FBRHT. Another interesting network identified by IPA is the one associated with cellular development and growth, proliferation and embryonic development, which includes C21orf57, CBLC, HOXB3, and HOXB6. Our finding that the representatives from these two sub-networks are repeatedly selected by FBRHT may suggest that the development of endometrial cancer, as well as the corresponding survival outcome, could be influenced by the regulation factors of cell proliferation and pathways of protein binding process. In conclusion, this study identified several candidate genes and sub-networks that may play an important role in key aspects of endometrial cancer development, and eventually lead to different survival outcome.

Discussion

In this paper, we proposed a feature selection method, Bayesian Robit regression with Hyper-LASSO priors (BayesHL), that employs MCMC to explore the posteriors of Robit classification models with heavy-tailed priors. We conducted experiments— with real data— that demonstrate BayesHL's ability to find sparse feature subsets with good predictive power, and to automatically make selections within groups of correlated features (without a pre-specified grouping structure). In future work, we would like to improve the accuracy of our feature subset selection method, and apply our Bayesian Inference framework to other models and non-convex penalties.

Regarding the first goal, we hope to optimize the selection of feature subsets. Currently, MCMC introduces small random jitters into the sample values of a β_j, which inadvertently lead to the selection of certain undesirable features. To address this problem, we use a large arbitrary threshold of 0.1 (on the relative magnitudes of coefficients) to eliminate such undesirable features. But this results in overly sparse feature subsets and risks omitting features with small coefficients. Future work should aim to resolve this optimization problem without introducing over-sparsity. There are three general approaches we could take. First, we could consider a fast optimization algorithm, which can take the sparsity in coefficients into consideration, to find the exact modes from the MCMC samples. Second, we could use mixture modeling or clustering methods to divide MCMC samples according to their modes, and third, we could use a "reference approach" to find the feature subset (from among the MCMC samples) that gives similar predictions as the global mode of all the MCMC samples (not the best within-sample predictive power)[65].

Finally, another possible direction for future work is to apply the Bayesian inference we developed in this paper to many other models (e.g., linear, graphical) and non-convex penalties to address feature selection problems in different application domains.

In conclusion, we would like to highlight two interesting findings from this study. First, our experiments with high-dimensional data— show that BayesHL results are comparable, in terms of their predictive power, to those of competitors (including LASSO, group LASSO, supervised group LASSO, random forest, penalized logistic regression, neural network, XGBoost and Knockoff) using far more sparse feature subsets. Secondly, in verifying the efficacy of BayesHL, we not only uncovered sparse feature subsets; we also identified genes that may be biologically meaningful in determining the survival outcome of endometrial cancer patients. Although we know that much work remains to be done, our results demonstrate that BayesHL has enormous potential for use in gene expression analysis.

Received: 18 November 2019; Accepted: 29 April 2020;
Published online: 16 June 2020

References

1. Clarke, R. *et al*. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer* **8**, 37–49 (2008).
2. Tolosi, L. & Lengauer, T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* **27**, 1986–1994, http://bioinformatics.oxfordjournals.org/content/27/14/1986.short (2011).
3. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288 (1996).
4. Candes, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: model-x. knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 551–577 (2018).
5. Sesia, M., Sabatti, C. & Candès, E. Gene hunting with hidden markov model knockoffs. *Biometrika* **106**, 1–18 (2018).
6. Li, L. & Yao, W. Fully bayesian logistic regression with hyper-lasso priors for high-dimensional feature selection. *Journal of Statistical Computation and Simulation* **88**, 2827–2851 (2018).
7. Polson, N. G. & Scott, J. G. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics* **9**, 501–538, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.727&rep=rep1&type=pdf (2010).
8. Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2**, 1360–1383 (2008).
9. Yi, N. & Ma, S. Hierarchical shrinkage priors and model fitting for high-dimensional generalized linear models. *Statistical applications in genetics and molecular biology* **11**, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3658361/. PMID: 23192052 PMCID: PMC3658361 (2012).
10. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360, https://doi.org/10.1198/016214501753382273.
11. Gelman, A. Prior distributions for variance parameters in hierarchical models. *Bayesian analysis* **1**, 515–533 (2006).
12. Carvalho, C. M., Polson, N. G. & Scott, J. G. Handling sparsity via the horseshoe. *Journal of Machine Learning Research* **5** (2009).
13. Carvalho, C. M., Polson, N. G. & Scott, J. G. The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–465 (2010).
14. Polson, N. G. & Scott, J. G. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* **7**, 887–902, http://projecteuclid.org/euclid.ba/1354024466 (2012).
15. Van Der Pas, S. *et al*. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2585–2618 (2014).
16. Zhang, C. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942, http://projecteuclid.org/euclid.aos/1266586618. MR: MR2604701 Zbl: 05686523 (2010).
17. Griffin, J. E. & Brown, P. J. Bayesian Hyper-Lassos with Non-Convex penalization. *Australian & New Zealand Journal of Statistics* **53**, 423–442, https://doi.org/10.1111/j.1467-842X.2011.00641.x/abstract (2011).
18. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429, https://doi.org/10.1198/016214506000000735 (2006).
19. Bhattacharya, A., Pati, D., Pillai, N. S. & Dunson, D. B. Bayesian shrinkage. *arXiv preprint arXiv:1212.6088*, http://arxiv.org/abs/1212.6088 (2012).
20. Armagan, A., Dunson, D. & Lee, J. Bayesian generalized double pareto shrinkage. *Biometrika*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.89&rep=rep1&type=pdf (2010).
21. Kyung, M., Gill, J., Ghosh, M. & Casella, G. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis* **5**, 369–412 (2010).
22. Polson, N. G. & Scott, J. G. Good, great, or lucky? screening for firms with sustained superior performance using heavy-tailed priors. *The Annals of Applied Statistics* **6**, 161–185, http://projecteuclid.org/euclid.aoas/1331043392 (2012).
23. Polson, N. G. & Scott, J. G. Local shrinkage rules, levy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 287–311, https://doi.org/10.1111/j.1467-9868.2011.01015.x/full (2012).
24. Jager, J., Sengupta, R. & Ruzzo, W. L. Improved gene selection for classification of microarrays. In *Proceedings of the eighth Pacific Symposium on Biocomputing: 3–7 January 2003; Lihue, Hawaii*, 53–64, https://books.google.com/books?hl=en&lr=&id=5_fRL7r-SSX0C&oi=fnd&pg=PA53&dq=+%22Improved+gene+selection+for+classification+of+microarrays%22+J+Jager&ots=I6swz4gcTp&sig=rXMzVsHbHI4mwPoPAG_wTuQOU0U (2002).
25. Huang, E. *et al*. Gene expression predictors of breast cancer outcomes. *The Lancet* **361**, 1590–1596 (2003).
26. Dettling, M. & Buhlmann, P. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* **90**, 106–131 (2004).
27. Park, M. Y., Hastie, T. & Tibshirani, R. Averaged gene expressions for regression. *Biostatistics* **8**, 212–227 (2007).
28. Reid, S. & Tibshirani, R. Sparse regression and marginal testing using cluster prototypes. *Biostatistics* **17**, 364–376 (2016).
29. Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 53–71 (2008).
30. Rapaport, F., Barillot, E. & Vert, J. Classification of arrayCGH data using fused SVM. *Bioinformatics* **24**, i375–i382 (2008).
31. Ma, S., Song, X. & Huang, J. Supervised group lasso with applications to microarray data analysis. *BMC bioinformatics* **8**, 60 (2007).
32. Breheny, P. & Huang, J. Coordinate Descent Algorithms For Nonconvex Penalized Regression, With Applications To Biological Feature Selection. *The annals of applied statistics* **5**, 232–253 URL, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3212875/. PMID: 22081779 PMCID: PMC3212875 (2011).
33. Breheny, P. & Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing* **25**, 173–187 (2015).
34. She, Y. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis* **56**, 2976–2990 (2012).
35. Wang, Z., Liu, H. & Zhang, T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics* **42**, 2164, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276088/ (2014).
36. Loh, P.-L. & Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, 476–484 (2013).
37. Polson, N. G., Scott, J. G. & Windle, J. The bayesian bridge. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 713–733, https://doi.org/10.1111/rssb.12042/abstract (2014).
38. Zucknick, M. & Richardson, S. MCMC algorithms for bayesian variable selection in the logistic regression model for large-scale genomic applications. *arXiv preprint arXiv:1402.2713*, http://arxiv.org/abs/1402.2713 (2014).
39. Piironen, J. & Vehtari, A. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, **54**, 905–913 (AISTATS, 2017).
40. Nalenz, M. & Villani, M. Tree ensembles with rule structured horseshoe regularization. *arXiv:1702.05008 [stat]*, http://arxiv.org/abs/1702.05008. arXiv: 1702.05008 (2017).

41. Johndrow, J. E. & Orenstein, P. Scalable MCMC for bayes shrinkage priors. *arXiv:1705.00841 [stat]*, http://arxiv.org/abs/1705.00841. arXiv: 1705.00841 (2017).
42. Neal, R. M. *et al*. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* **2**, 2 (2011).
43. Piironen, J. *et al*. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11**, 5018–5051 (2017).
44. Liu, C. Robit regression: A simple robust alternative to logistic and probit regression. *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 227–238 (2004).
45. Lange, K. L., Little, R. J. & Taylor, J. M. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881–896 (1989).
46. Abramowitz, M. & Stegun, I. A. *Handbook of Mathematical Functions* (Dover publications, 1972).
47. Holmes, C. C. & Held, L. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168, http://projecteuclid.org/euclid.ba/1340371078. Mathematical Reviews number (MathSciNet) MR2227368 (2006).
48. Robin, X. *et al*. pROC: an open-source package for r and s+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77, http://www.biomedcentral.com/1471-2105/12/77 (2011).
49. andLuis Torgo, P. B. & Ribeiro, R. A survey of predictive modeling under imbal-anced distributions. *ACM Comput. Surv* **49**, 1–31 (2016).
50. He, H. & Ma, Y. *Imbalanced learning: foundations, algorithms, and applications* (John Wiley & Sons, 2013).
51. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10** (2015).
52. Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research* **13**, 27–66 (2012).
53. Murphy, K. P. *Machine learning: a probabilistic perspective* (MIT press, 2012).
54. Guh, R.-S. & Hsieh, Y.-C. A neural network based model for abnormal pattern recognition of control charts. *Computers & Industrial Engineering* **36**, 97–108 (1999).
55. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (ACM, 2016).
56. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845 (1988).
57. Krämer, A., Green, J., Pollard, J. Jr. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2014).
58. Bodey, B., Bodey, J. B., Siegel, S. E. & Kaiser, H. E. Immunocytochemical detection of the homeobox b3, b4, and c6 gene products in breast carcinomas. *Anticancer research* **20**, 3281–3286 (2000).
59. Hong, J. H. *et al*. Expression pattern of the class i homeobox genes in ovarian carcinoma. *Journal of gynecologic oncology* **21**, 29–37 (2010).
60. Chen, H. *et al*. mir-10b inhibits apoptosis and promotes proliferation and invasion of endometrial cancer cells via targeting hoxb3. *Cancer Biotherapy and Radiopharmaceuticals* **31**, 225–231 (2016).
61. Deckert, M. & Rottapel, R. The adapter 3bp2: how it plugs into leukocyte signaling. In *Lymphocyte Signal Transduction*, 107–114 (Springer, 2006).
62. Ohno, S. *et al*. Correlation of histological localization of tumor-associated macrophages with clinicopathological features in endometrial cancer. *Anticancer research* **24**, 3335–3342 (2004).
63. Voronina, E. *et al*. Ovarian granulosa cell survival and proliferation requires the gonad-selective tfiid subunit taf4b. *Developmental biology* **303**, 715–726 (2007).
64. Ribeiro, J. R., Lovasco, L. A., Vanderhyden, B. C. & Freiman, R. N. Targeting tbp-associated factors in ovarian cancer. *Frontiers in oncology* **4**, 45 (2014).
65. Piironen, J. & Vehtari, A. Comparison of bayesian predictive methods for model selection. *Statistics and Computing* **27**, 711–735, https://doi.org/10.1007/s11222-016-9649-y (2017).

Acknowledgements

The research of Lai Jiang and Celia Greenwood is supported by CIHR grant PJT-363444. The research of Longhai Li is supported by funding from Natural Sciences and Engineering Research Council of Canada (NSERC IDG: RGPIN-2019-07020), and Canada Foundation of Innovations (CFI: 22217). The research of Weixin Yao is supported by NSF grant DMS-1461677. Data analysis in this paper was enabled in part by support provided by the GRAHAM computer cluster of Compute Canada (nzt-671-ab).

Author contributions

L.J. conceived the study, carried out the data analysis and drafted the manuscript. L.L. conceived the study and helped to draft the manuscript. C.G. coordinated the data analysis and edited the manuscript. W.Y. participated in the study design. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-66466-z.

Correspondence and requests for materials should be addressed to L.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.