Published in final edited form as:

J Stat Comput Simul. 2017 ; 87(14): 2708–2723. doi:10.1080/00949655.2017.1341887.

Fast Bayesian Variable Screenings for Binary Response Regressions with Small Sample Size

S.-M. Chang^a, J.-Y. Tzeng^b, and R.-B. Chen^{a,*}

^aDepartment of Statistics, National Cheng Kung University, Tainan, Taiwan

^bDepartment of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA

Abstract

Screening procedures play an important role in data analysis, especially in high-throughput biological studies where the datasets consist of more covariates than independent subjects. In this article, a Bayesian screening procedure is introduced for the binary response models with logit and probit links. In contrast to many screening rules based on marginal information involving one or a few covariates, the proposed Bayesian procedure simultaneously models all covariates and uses closed-form screening statistics. Specifically, we use the posterior means of the regression coefficients as screening statistics; by imposing a generalized g-prior on the regression coefficients, we derive the analytical form of their posterior means and compute the screening statistics without Markov chain Monte Carlo implementation. We evaluate the utility of the proposed Bayesian screening method using simulations and real data analysis. When the sample size is small, the simulation results suggest improved performance with comparable computational cost.

Keywords

g-prior; Logistic regression; Probit regression; Sure independence screening

1. Introduction

When dealing with high-dimensional data, usually the number of covariates is much larger than the number of independent subjects. Shrinkage methods such as LASSO [1], adaptive LASSO [2], SCAD [3], and Dantzig selector [4] have been commonly applied when the number of covariates, p , is greater than the sample size, n . However, when the variable selection problem is of ultrahigh dimensional, i.e. $\log(p) = O(n^\xi)$ with $0 < \xi < 1$, the shrinkage methods do not perform well [5]. In this circumstance, one practical strategy is to screen out unimportant covariates so that the dimension of the parameter space is reduced to a feasible size. The goal of screening is, therefore, to control false negative rate in a tolerable level rather than to pursue high accuracy among the selected “important” variables. In this article, we refer to those covariates that truly affect the responses as *active* covariates

*Corresponding author. rbchen@mail.ncku.edu.tw.

whereas the rest are considered as *inactive* covariates. We also define those covariates that are considered to affect the response by some statistical methods as important covariates, defining the remainder as unimportant covariates.

One of the pioneer approaches for screening is the sure independent screening (SIS; [5]). The SIS was originally designed for linear regressions and then was extended to generalized linear regression models by [6]. In general, the screening statistic of the j th covariate is the maximum likelihood estimate (MLE) of the regression slope in a working model that only contains the j th covariate. The screening rule is to select the first k covariates that have the largest slope estimates (in absolute value) as important covariates where k is pre-determined. Fan and Lv [5] showed that, under some regularity conditions, the SIS possesses the sure screening property, i.e., the probability of keeping all active covariates as important covariates converges to 1.

In addition to SIS, [7] proposed sure independent ranking and screening (SIRS). Their screening statistic for the j th covariate is the covariance between the j th covariate and the distribution function of the response conditional on the j th covariate. Since the screening statistic requires no model assumption, the SIRS is model-free. Later, [8] considered distance correlation (DC; [9]) to measure the relationship between the response and covariates. The major reason for applying the DC measure is that the DC between two random vectors is 0 if and only if these two random vectors are independent. So the DC should perform better than the SIS if the response and covariate have a nonlinear relationship. Recently, [10, 11] proposed Kolmogorov filters (KF) for screening. They utilized the idea of Kolmogorov-Smirnov test and took the largest distance among empirical distribution functions of the j th covariate conditional on the response as the screening statistic for the j th covariate.

One limitation of the SIS arises from the fact that only one covariate was applied for each screening statistic. The marginal slope, which is used as a surrogate measure of the true relationship between the response and a covariate, could be misleading especially when the covariates are correlated in linear regression context [12, 13]. To overcome this deficiency, [5] proposed iterative sure independence screening (ISIS) method, in which the screening statistic depends on more than one covariates in one working model. This idea cannot be heuristically applied to any of the SIRS, DC and KF methods. In addition, for binary response regression variable screening, the SIS and DC can be directly applied but not for the SIRS. The KF in [10, 11] is designed for two-class classification problem which is equivalent to binary response regression problem. So we will apply the proposed method, SIS, ISIS, DC, and KF in the latter simulation studies for the comparison purpose.

In contrast to the classical approaches, Bayesian logistic regression models might have the advantage in dealing with small sample data. As pointed out by [14], for logistic regressions, the unique MLE would not exist (i.e. non-unique or infinite) when the data are completely or quasi-completely separated by a hyperplane spanned by the covariates involved in the working model. Bayesian methods mentioned in [15, 16] guarantee finite and unique Bayesian estimates for the regression coefficients. Moreover, some Bayesian regression models are unified for both $n \geq p$ and $n < p$ cases without changing model structures. For

instance, [17] applied independent priors for regression coefficients to select important covariates which can be applied to both $n \geq p$ and $n < p$ cases. Their simulation results are promising although their p - n ratio is not large. Furthermore, Bayesian approaches balance the information provided by data and domain knowledge. However, there are two major challenges for Bayesian joint-screening procedures: First, when $n < p$, the priors used in [15, 16] becomes inapplicable, and, second, the implementation tends to be computationally intensive because the computation would require either thousands of Markov chain Monte Carlo (MCMC) iterations or high-dimensional integrals. While one alternative is to approximate these integrals by using the Laplace approximation [18], the Laplace approximation might not perform well universally as explored by [19–21]. To address these issues, we introduce a generalized g-prior for regression coefficients and derive their closed-form posterior means for probit and logit links, and then use these posterior means as screening statistics.

In this article, we consider a Bayesian approach to tackle the screening problem for binary response regression models and $n < p$. To alleviate the limitation of marginal screenings, all covariates are considered in one generalized linear model and Zellner's g-prior [22] is modified to fit the $n < p$ situation. This modified prior results in the closed-form posterior means of the regression coefficients provided $n < p$ and the posterior mean serves as the screening statistic. The computation cost of the proposed method is dominated by performing one singular value decomposition on the $n \times (p + 1)$ design matrix of the regression model, which is computationally efficient. Hereafter, we abbreviate the proposed fast Bayesian screening by Fabs. Also, we demonstrate by simulations that ISIS outperforms SIS and Fabs when the sample size is as large as 200 whereas Fabs outperforms ISIS when the sample size is as small as 72, the sample size of the case study in this paper.

The rest of this article is arranged as follows. In Section 2, binary response regression models with probit and logit links are introduced. Based on the Gaussian prior for regression coefficients, we derive the posterior means of the regression coefficients and show that the computation of the posterior means involves n -dimensional integrals and is unfeasible for a screening procedure. To avoid the high-dimensional integrals, in Section 3, we propose a generalized g-prior for cases of $n < p$ and then derive the closed-form posterior means of these regression coefficients under the probit and logistic regression models. In Section 4, we perform simulation studies to demonstrate the utility of the Fabs followed by a real case in Section 5. Finally, concluding remarks are provided in Section 6.

2. Bayesian Binary Response Regression Models

We begin by fixing notation. Suppose that, for the i th subject ($i = 1, \dots, n$), we can observe a binary response y_i^* , $y_i^* \in \{0, 1\}$, and a p -dimensional covariate vector $\mathbf{x}_i$, which may or may not include intercept. Let $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$ be the corresponding regression coefficient vector with respect to $\mathbf{x}_i$. With a monotone increasing link function $\Psi(\cdot) : \mathbb{R} \rightarrow [0, 1]$, we set $\Pr(y_i^* = 1) = \Psi(\mathbf{x}_i^T \boldsymbol{\beta})$ and $\Pr(y_i^* = 0) = 1 - \Psi(\mathbf{x}_i^T \boldsymbol{\beta})$. Then if the link function $\Psi(\cdot)$ is chosen such that $\Psi(x) + \Psi(-x) = 1$, for example, probit link and logit link, the likelihood function can be written as follows:

$$\prod_{i=1}^n \Psi(y_i \mathbf{x}_i^T \boldsymbol{\beta}), \quad (1)$$

where $y_i = 2y_i^* - 1$. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ be the $n \times p$ design matrix of the covariates. Define $\mathbf{y}^{*T} = (y_1^*, \dots, y_n^*)$, $\mathbf{y}^T = (y_1, \dots, y_n)$, and $\mathbf{Y} = \text{diag}\{y_i\}$ which is a diagonal matrix with diagonal elements, $y_1, \dots, y_n$. Let $\mathbf{I}_n$ be the n -dimensional identity matrix. Additionally, define $\ell(\cdot)$ and $L(\cdot)$ as the density function and the distribution function of the standard logistic distribution, respectively. Last, for a p -dimensional random vector $\mathbf{w}$ following a multivariate normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$, say $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we denoted its density function by $\Phi(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\phi(\cdot)$ and $\Phi(\cdot)$ act as the density function and distribution function of the standard normal distribution, respectively.

For Bayesian analysis approaches, a commonly considered prior for the regression coefficient vector, $\boldsymbol{\beta}$, is $\mathcal{N}(\boldsymbol{\beta}_0, \mathbf{D})$ where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the prior mean vector and $\mathbf{D} \in \mathbb{R}^{p \times p}$ is a pre-specified covariance matrix. The priors proposed by [22–25] also belong to this class. First, we consider the probit regression with likelihood function (1). Thus $\Psi(\cdot) = \Phi(\cdot)$. Then the moment generating function of the posterior distribution of $\boldsymbol{\beta}$ is

$$M_{\boldsymbol{\beta}}^P(\mathbf{t}) \propto \int_{\mathbb{R}^p} e^{\boldsymbol{\beta}^T \mathbf{t}} \left[\prod_{i=1}^n \Phi(y_i \mathbf{x}_i^T \boldsymbol{\beta}) \right] \phi(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \mathbf{D}) d\boldsymbol{\beta}.$$

Similarly, for logistic regressions, the definition of the posterior moment generating function, $M_{\boldsymbol{\beta}}^L(\mathbf{t})$, is identical to the probit regression, except for replacing $\Phi(y_i \mathbf{x}_i^T \boldsymbol{\beta})$ by $L(y_i \mathbf{x}_i^T \boldsymbol{\beta})$, for $i = 1, \dots, n$. As shown in [26, 27],

$$L(x) = \int_{\mathbb{R}_+} \Phi(x/\sigma) h(\sigma) d\sigma$$

where $\sigma/2$ follows the Kolmogorov–Smirnov distribution [28] and $h(\cdot)$ is the corresponding density function. After some manipulation, we obtained the following results. The proofs are deferred to Appendix A.

Proposition 2.1

The cumulant generating function of the posterior distribution with probit link is

$$S_{\boldsymbol{\beta}}^P(\mathbf{t}) = \log(M_{\boldsymbol{\beta}}^P(\mathbf{t})) = \mathbf{t}^T \boldsymbol{\beta}_0 + \frac{1}{2} \mathbf{t}^T \mathbf{D} \mathbf{t} + C^P + \log \left(\int_{\mathbb{R}_+^n} \phi(\mathbf{w}; \mathbf{Y} \mathbf{X}(\boldsymbol{\beta}_0 + \mathbf{D} \mathbf{t}), \boldsymbol{\Delta}) d\mathbf{w} \right) \quad (2)$$

where $\mathbb{R}_+^n = [0, \infty)^n$ and $\boldsymbol{\Delta} = \mathbf{I}_n + \mathbf{Y} \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{Y}$; and the cumulant generating function of the posterior distribution with logit link is

$$S_{\beta}^L(t) = \log(M_{\beta}^L(t)) = t^T \beta_0 + \frac{1}{2} t^T D t + C^L + \log \left(\int_{\mathbb{R}_+^n} \left\{ \int_{\mathbb{R}_+^n} \phi(w; Y X(\beta + D t), \tilde{\Delta}) h(\sigma) d\sigma \right\} dw \right)$$

where $\tilde{\Delta} = \sum_{i=1}^2 Y X D x^T Y$. Both C^P and C^L are constant in terms of t .

To derive the posterior means of regression coefficients β_i 's, we take the posterior mean of the probit model for illustration. Based on (2), the posterior mean vector of the regression coefficient vector β is

$$\beta_0 + D X^T Y \Delta^{-1} [E(W) - Y X \beta_0]$$

where W is an n -dimensional random vector that follows a truncated multivariate normal distribution with location vector $Y X \beta_0$, scaling matrix Δ , and left truncation point $\mathbf{0}$. The calculation of $E(W)$ has been studied in [29–32] and others. This calculation requires the evaluation of an n -dimensional integral or n evaluations of $(n-1)$ -dimensional integrals. According to our experience, the evaluation of this general form is too computationally intensive to be practical, even when the sample size is as small as 100. To take these posterior means as practical screening statistics, we need to tackle this high-dimensional integration first. One possible solution to avoid such high-dimensional integrals is to choose a covariance matrix D properly, such that the resulting Δ is diagonal. As a result, the n -dimensional integral in (2) can be rewritten as the product of n 1-dimensional integrals and is numerically feasible for all sizes of n . This motivates us to propose a generalized g -prior, which will be introduced in Sect. 3.1.

3. Bayesian Screening Statistics

3.1. Generalized g -prior

When $p < n$, consider the linear regression problem $Y = X\beta + e$ with $e \sim N(\mathbf{0}, \sigma^2 I_n)$. Denote the MLE of β as $\hat{\beta}$. Then the covariance matrix of $\hat{\beta}$ is $\sigma^2 (X^T X)^{-1}$. [22] mimicked this result and set g -prior for the regression coefficient as $\beta \sim N(\mathbf{0}, g \times (X^T X)^{-1})$, $g > 0$. [24] considered the g -prior for Bayesian variable selection with binary responses via probit link. Based on the g -prior structure, the model matrix X must have full column rank. When we focus on the large p and small n problem, X cannot have full column rank and thus $X^T X$ is singular. To overcome the singularity problem, the ridge type modification, i.e. $\beta \sim N(\mathbf{0}, g \times [\lambda I_n + X^T X]^{-1})$, with $\lambda > 0$ and $g > 0$, has been studied in [23]. Tuning the ridge parameter λ is, therefore, a critical issue. In addition to the ridge type method, [25] suggested using the Moore-Penrose generalized inverse of $X^T X$. However, applying none of the two modifications reduces the computational burden in the view of calculating the posterior means as shown in Proposition 2.1.

To simplify the computation of the cumulant generating functions in Proposition 2.1, especially for $n < p$, another modification form of g -prior is introduced as follows. Let $X =$

USV^T be the singular value decomposition (SVD), where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ are orthogonal matrices, and $S \in \mathbb{R}^{n \times p}$ is a matrix such that its (i, j) th element is nonzero if $i = j$ and is zero otherwise. In this sequel, we define $S = [\Omega, \mathbf{0}]$ where $\Omega \in \mathbb{R}^{n \times n}$ is a diagonal matrix and $\mathbf{0}$ is a zero matrix with conformable dimension. Since

$$X^T X = V S^T U^T U S V^T = V \begin{bmatrix} \Omega^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} V^T$$

is nonnegative definite, instead of the ridge type method, a remedy to make it positive definite can be

$$(X^T X)_\varepsilon \equiv V \begin{bmatrix} \Omega^2 & \mathbf{0} \\ \mathbf{0} & \varepsilon I_{p-n} \end{bmatrix} V^T$$

for any $\varepsilon > 0$. In fact, $(X^T X)_\varepsilon$ is modified from the result of the eigenvalue decomposition of $X^T X$ by replacing zero eigenvalues with a small positive number ε . Hereafter, we call

$$\beta \sim N(\beta_0, g(X^T X)_\varepsilon^{-1}) \quad (3)$$

as the generalized g-prior and let D_ε be a shorthand for $(X^T X)_\varepsilon^{-1}$. Two properties of D_ε in this prior are summarized below.

Proposition 3.1—Given any $\varepsilon > 0$,

$$D_\varepsilon X^T = V \begin{bmatrix} \Omega^{-1} \\ \mathbf{0} \end{bmatrix} U^T \quad \text{and} \quad X D_\varepsilon X^T = I_n.$$

Because the proof is straightforward, we have omitted it here. The former property results in a posterior mean of β that is free of ε , such that no effort should be directed toward tuning the parameter ε . The latter plays a key role in deriving closed-form posterior means of β . Recall that the integrand in (2) is a multivariate normal with variance $= I_n + YXDX^TY$. Applying the generalized g-prior results in $YX(gD_\varepsilon)X^TY = g \times YI_nY = g \times I_n$ by Proposition 3.1. Therefore the n -dimensional multivariate normal density in (2) becomes the product of n 1-dimensional normal densities. The benefit of using the proposed generalized g-prior is, therefore, significant.

3.2. Proposed Screening Statistics

For probit models, using the generalized g-prior (3), the covariance matrix in (2) becomes $(1 + g) \times I_n$ due to Proposition 3.1. Then the n -dimensional integral in (2) can be rewritten as the product of n 1-dimensional integrals, such as

$$\prod_{i=1}^n \int_{(-y_i \mathbf{x}_i^T \boldsymbol{\beta}_0 - y_i \mathbf{x}_i^T g \mathbf{D}_\varepsilon \mathbf{t}) / \sqrt{1+g}}^{\infty} \phi(u_i) du_i = \prod_{i=1}^n \Phi \left(\frac{y_i \mathbf{x}_i^T \boldsymbol{\beta}_0 + y_i \mathbf{x}_i^T g \mathbf{D}_\varepsilon \mathbf{t}}{\sqrt{1+g}} \right).$$

The computational cost is therefore greatly reduced compared to directly evaluating $E(W)$. For the logistic model, the benefit of applying the generalized g-prior can also be seen in the calculation of its posterior mean.

Proposition 3.2—With the generalized g-prior (3), the posterior mean of the probit model is

$$\hat{\boldsymbol{\beta}}_P = \boldsymbol{\beta}_0 + \frac{g}{\sqrt{1+g}} \mathbf{D}_\varepsilon \mathbf{X}^T \mathbf{Q} \mathbf{y} \quad \text{where } \mathbf{Q} = \text{diag} \left\{ \frac{\phi(y_i \mathbf{x}_i^T \boldsymbol{\beta}_0 / \sqrt{1+g})}{\Phi(y_i \mathbf{x}_i^T \boldsymbol{\beta}_0 / \sqrt{1+g})} \right\},$$

and the posterior mean of the logistic model is

$$\hat{\boldsymbol{\beta}}_L = \boldsymbol{\beta}_0 + \sqrt{g} \mathbf{D}_\varepsilon \mathbf{X}^T \mathbf{Q} \mathbf{y} \quad \text{where } \mathbf{O} = \text{diag} \left\{ \frac{\int_{\mathbb{R}} \phi \left(\frac{z + y_i \mathbf{x}_i^T \boldsymbol{\beta}_0}{\sqrt{g}} \right) l(z) dz}{\int_{\mathbb{R}} \Phi \left(\frac{z + y_i \mathbf{x}_i^T \boldsymbol{\beta}_0}{\sqrt{g}} \right) l(z) dz} \right\}$$

and $l(z) = e^z / (1 + e^z)^2$.

The proof is deferred to Appendices B. Clearly, $\hat{\boldsymbol{\beta}}_P$ and $\hat{\boldsymbol{\beta}}_L$ are free of ε because both posterior means associate with ε only through the term $\mathbf{D}_\varepsilon \mathbf{X}^T$.

Next, we briefly discuss the role of the generalized g-prior in the screening statistic when there is no prior information about the importance of covariates, i.e., $\boldsymbol{\beta}_0 = \mathbf{0}$. The screening statistics are therefore proportional to $\mathbf{D}_\varepsilon \mathbf{X}^T \mathbf{y}$. According to Sect. 3.1, we first partition $\mathbf{V}$ into $[\mathbf{V}_1 \ \mathbf{V}_2]$ where $\mathbf{V}_1 \in \mathbb{R}^{p \times n}$ and $\mathbf{V}_2 \in \mathbb{R}^{p \times (p-n)}$ and, therefore,

$$(\mathbf{X}^T \mathbf{X})_\varepsilon = \mathbf{V} \begin{bmatrix} \boldsymbol{\Omega}^2 & \mathbf{0} \\ \mathbf{0} & \varepsilon \mathbf{I} \end{bmatrix} \mathbf{V}^T = [\mathbf{V}_1 \ \mathbf{V}_2] \begin{bmatrix} \boldsymbol{\Omega}^2 & \mathbf{0} \\ \mathbf{0} & \varepsilon \mathbf{I}_{p-n} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1^T \\ \mathbf{V}_2^T \end{bmatrix} = \mathbf{V}_1 \boldsymbol{\Omega}^2 \mathbf{V}_1^T + \varepsilon \mathbf{V}_2 \mathbf{V}_2^T = \mathbf{X}^T \mathbf{X} + \varepsilon \mathbf{V}_2 \mathbf{V}_2^T.$$

Thus, the screening statistic $\mathbf{D}_\varepsilon \mathbf{X}^T \mathbf{y}$ is the solution of the least square problem

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \varepsilon \boldsymbol{\beta}^T \mathbf{V}_2 \mathbf{V}_2^T \boldsymbol{\beta}.$$

In this sequel, the proposed screening statistic can also be viewed as a solution of generalized ridge regression.

There are a few remarks to be made on the derived posterior means in Proposition 3.2. First, we note that such a simplified calculation would only be valid for $n < p$, as when $n > p$,

$\mathbf{XDX}^T$ is singular for any symmetric and positive definite matrix $\mathbf{D}$. Therefore, we cannot find a $\mathbf{D}$ that results in $\mathbf{I}_n$ being proportional to $\mathbf{I}_n$ when $n > p$. Second, for the probit model, the posterior distribution follows a multivariate skew normal with multiple constraints and may even be skewed and multimodal [33]. To illustrate this point, consider the case that $n = 1$ and $\boldsymbol{\beta}_0 = \mathbf{0}$ and then, by setting $\mathbf{z} = [\mathbf{gD}_\varepsilon]^{-1/2}\boldsymbol{\beta}$, the posterior distribution becomes $2\boldsymbol{\Phi}(\mathbf{z}; 0, \mathbf{I}_d)\Phi(\boldsymbol{\alpha}^T\mathbf{z})$ where $\boldsymbol{\alpha} = [\mathbf{gD}_\varepsilon]^{1/2}\mathbf{x}_{1Y1}$. This distribution takes the form of a multivariate skewed normal, and its skewness and kurtosis might vary with respect to $\boldsymbol{\alpha}$. Similar conclusions can be made for the logistic model.

To compute these screening statistics, we apparently need to specify the hyperparameters $\boldsymbol{\beta}_0$, ε , and g in the generalized g-prior. At the screening stage, there is no information about active covariates, in general. In such cases, the zero prior mean vector, i.e. $\boldsymbol{\beta}_0 = \mathbf{0}$, is reasonable. For choosing ε , according to Proposition 3.1, both probit and logistic screening statistics, $\hat{\beta}_p$ and $\hat{\beta}_L$, are free of ε . Additionally, [34] provided an insightful discussion about choosing g for the g-prior which should be a promising guidance of choosing g for our generalized g-prior. However, when $\boldsymbol{\beta}_0 = \mathbf{0}$, both $\hat{\beta}_p$ and $\hat{\beta}_L$ are proportional to $\mathbf{D}_\varepsilon\mathbf{X}^T\mathbf{y}$ which is independent of g . Thus the values of ε and g are not essential to the rankings of both screening statistics. The computation of the proposed screening statistic is very simple when $\boldsymbol{\beta}_0 = \mathbf{0}$. We demonstrate it by R language [35] in Appendix C.

4. Simulation Studies

4.1. Simulation Setup

For each simulation dataset, we independently sample n observations from the following data generating process:

$$\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T \sim N(\mathbf{0}, \mathbf{V}_K(r)), \quad K \in \{I, M, A, C\},$$

$$m_i = m_g(\mathbf{x}_i) + \varepsilon_i, \quad g=1, 2, 3, \quad \varepsilon_i \sim N(0, 1), \quad \text{and} \quad (4)$$

$$y_i^* = \delta(m_i \geq 0) \quad (i=1, \dots, n),$$

where $\delta(a)$ is the indicator function of a , i.e., $\delta(a) = 1$ if a is true and 0, otherwise; x_j 's and ε_j 's are independent; and $\mathbf{V}_K(r) \in \mathbb{R}^{p \times p}$ is a valid covariance matrix characterized by the parameter r . We consider four structures of the covariance:

$$\mathbf{V}_I: cov(x_{ij}, x_{ik}) = \delta(j=k),$$

$$\mathbf{V}_M: cov(x_{ij}, x_{ik}) = \delta(j=k) + r \times \delta(|j-k|=1),$$

$$\mathbf{V}_A: cov(x_{ij}, x_{ik}) = r^{|j-k|}, \text{ and}$$

$$\mathbf{V}_C: cov(x_{ij}, x_{ik}) = \delta(j=k) + r \times \delta(j \neq k).$$

$\mathbf{V}_I(r)$ is the identity matrix; $\mathbf{V}_M(r)$ is a first-order moving average correlation matrix; $\mathbf{V}_A(r)$ is a first-order autoregressive correlation matrix; and $\mathbf{V}_C(r)$ is a compound symmetric correlation matrix, for $r \in (0, 1)$. Some properties of these covariance matrices were discovered in [36]. In the following simulations, we set $r = 0.5$. Also, we considered three mean models:

$$\text{M1: } m_1(x_i) = -1 + x_{i1} + x_{i2} - x_{i3} - x_{i4},$$

$$\text{M2: } m_2(x_i) = -1 + x_{i1} + x_{i2} + x_{i3} - 3x_{i4}, \text{ and}$$

$$\text{M3: } m_3(x_i) = -1 + x_{i1}x_{i2} + \delta(x_{i3} > 0) - \exp\{x_{i4}\}.$$

The mean model M1 is a regular linear part of generalized linear regressions. The mean model M2 is designed so that x_{i4} is uncorrelated with $m_2(x_i)$ under $\mathbf{V}_C(r)$, for any valid r , so that SIS should have little chance to identify x_{i4} as an important covariate. Last, the mean model M3, also applied in [8], represents three nonlinear relationships between response and covariates so that model-free approaches such as the KF and DC should perform better than model-based methods such as the SIS, ISIS, and Fabs.

Below, we consider SIS, ISIS, KF, DC, and Fabs approaches to screen covariates. For ISIS, following [5], a two-iteration ISIS was applied. Let κ^* be the largest integer that is smaller than or equal to $n/\log(n)$, and κ_1 be an integer in $(1, \kappa^*)$. Then, in the first iteration of ISIS, the top κ_1 ranking covariates from SIS are selected. In the second iteration, the screening statistic of the k th covariate is the MLE of the regression coefficient associated with the k th covariate with a working model that contains the k th covariate and the κ_1 covariates determined in the first iteration. For the second iteration with a given κ_1 , the computational complexity of the k th screening statistic is $O(n(\kappa_1 + 2)^2)$. This is because there are $\kappa_1 + 2$ unknown parameters in one working model including intercept, in the view of using iteratively re-weighted least squares and hence the total complexity for calculating ISIS screening statistics is $O(pn(\kappa_1 + 2)^2)$. To save computational burden, in the following simulation studies, we set $\kappa_1 = 4$ if no further specification is provided. The other reason of choosing $\kappa_1 = 4$ is that there are 4 active covariates in above models and thus, ideally, the ISIS has chance to catch all active covariates at the first iteration.

4.2. Simulation Results

For each configuration of the covariance structures and mean models, 500 datasets were simulated. We implemented the SIS, ISIS, DC, KF, and Fabs to screen variables. Simulation results were summarized in Tables 1 – 3. In Table 1, the mean model M1 was considered. As expected, the covariance structure among covariates did affect the screening results. Under V_I , the five screening statistics had comparable performance. Over 95% chance, the first $38 \approx 200/\log(200)$ important covariates included all active covariates. This chance reduced to over 75% when the covariance structures were V_M and V_A . When applying V_C , only the ISIS and Fabs retained over 75% chance and the rest screening approaches became worse. We observed that when the mean model is M1, the ISIS and Fabs can defeat the effect due to rich covariance structures (V_C is much richer than V_I). Results shown in Table 2 was similar to Table 1, especially under V_C . Additionally, both the model-free approaches, KF and DC, also suffered from the rich covariances. This may be due to that the detection power is deteriorated when the response is dichotomous. When the underlying model is nonlinear, M3, as shown in Table 3, both KF and DC showed their superiority over the model-based approaches, SIS, ISIS, and Fabs. However, the smallest number required to include all active covariates was, with high chance, too large to yield a successful screening, especially under V_I , V_M , and V_C covariance structures.

In general, under M1, SIS, ISIS and Fabs outperformed KF and DC because the model is actually linear. Under some special combinations of the model and covariance structure, e.g., model M2 with $V_C(r)$, the SIS had high chance to miss some active covariates. The ISIS and Fabs greatly alleviated this deficiency at least under the considered scenarios. The performances of aforementioned model-free approaches were also affected by the covariance structure among covariates. Their corresponding screening results were not promising under the nonlinear model M3 although they did outperform the model-based approaches. Lastly, the ISIS approach outperformed Fabs except under the scenario with M2 and V_C .

Up to now, the simulation results showed that the ISIS has satisfactory screening performance for linear models with binary response. However, the ISIS might not perform well for smaller sample size. For example, in our case data study (Section 5, when $n = 72$ and $p = 7000$), the behavior of the ISIS is not as good as the cases of large sample size, say 200, as in previous simulations. We observed that the ISIS for logistic regressions is computationally unstable because, frequently, the MLEs do not exist or are not unique for logistic regression coefficients. This phenomenon had been studied and explained in [14]. The non-existence or non-unique MLE does not only significantly hurt the performance of the ISIS but also prolongs the computing time greatly. One way to obtain a finite estimate is to use Firth's approach which is implemented in the R package "logistf" [37].

To further illustrate this weakness, we consider the cases of $n = 72$ (as in our case study) and set $p = 1000, 3000$, and 7000. For the linear part, we only consider M1 and M2 since the considered approaches were designed for linear models. Applied covariance models were V_I and V_C because results with V_M and V_A should be similar to the one with V_I when using M1 and M2. Four screening methods, Fabs, SIS, ISIS, and ISIS_F (using Firth's approach to replacing the regular logistic regression), are used to screen covariates. Results are shown in

Tables 5 – 7. We observed that, as expected, the $ISIS_F$ outperformed $ISIS$ but the improvement is very limited. Moreover, the SIS outperformed $ISIS_F$ greatly, and $Fabs$ and SIS performed similarly under V_I and $Fabs$ is much better than SIS under V_C . This simulation result suggests that, when the sample size is small, $Fabs$ can be a computationally stable alternative to the logistic regression MLE in terms of screening.

5. Case Study: Discriminating Leukemia Types

We applied the aforementioned screening approaches to the leukemia dataset [38] which consists of the gene expression data of 7128 genes for 72 patients. Among the 72 patients, 45 were diagnosed with acute lymphoblastic leukemia, and the other 27 with acute myeloid leukemia. The goal of the original study was to classify the two leukemia types based on the expressions of the 7128 genes, even though only a handful of genes are expected to make difference. After the screening given in [39], 3571 genes remained for further analysis. In the literature, this dataset has been used to evaluate the performance of different variable selection methods, mostly based on mis-classification rates. Because the focus of the paper is a variable screening method instead of a variable selection method, a more appropriate metric should be based on whether the active genes are included in the set of top ranking genes. Given that the active genes are unknown, we used the list of important genes reported in [24, 25] as our positive controls and treated these genes are “active genes”. These two papers identified 5 and 6 important genes, respectively, to establish accurate classification rules with biological explanations based on these important genes. Both studies utilized the data augmentation approach [40] and the stochastic search variable selection [41] for gene selection but with different MCMC sampling strategies.

We considered SIS , $ISIS$, KF , DC , and $Fabs$ approaches and compared the resulting ranks of the “active genes” obtained from different screening methods. For $ISIS$, as suggested by [5], we selected $\kappa_1 = 8$ top ranking genes from the results of SIS and then included these 8 genes in the working model to perform the second iteration of the screening. The results, shown in Table 8, suggest that the $Fabs$ and SIS performed similarly but the $ISIS$ perform extremely poorly. Specifically, the $Fabs$ had the smallest $\kappa_0 = 57$, closely followed by the SIS ($\kappa_0 = 63$). For the model-free approaches, DC and KS need to take more than 100 genes to make sure the screening results which cover all active genes. The $ISIS$ required to include the top 6906 important genes in order to cover all of the positive-control genes. This result is not surprising because, as we have concluded in Section 4, the density function of the smallest numbers required to include all active covariates tends to have a very long right tail.

Last, we reveal the computing time among $Fabs$, SIS and $ISIS$ when performing the screening task with the full data: the SIS took 13 seconds, the $ISIS$ took 33 seconds, and the proposed approach took less than 1 second.

6. Concluding Remarks

In this article, we introduce a Bayesian screening method to efficiently remove inactive covariates and retain all of the active covariates in ultrahigh dimensional space. Specifically, we propose a generalized g-prior for binary response regression under a high dimensional

setting. Then with the probit and logit links, we derive the posterior means of the regression coefficients in closed forms, respectively. Using these posterior means as screening statistics, we are able to perform joint screening while reducing a significant amount of the computing time in the screening stage. Moreover, unlike ISIS whose performance depends highly on the choices of screening parameters, the proposed Fabs involves no tuning parameter, which makes it attractive in practice. Last, the Fabs works only for $n < p$. When $n > p$ cases, we suggest using the Bayesian methods as discussed in the review paper of [42] to select important covariates.

The proposed screening statistic, the Fabs, is a generalized ridge regression coefficient estimation when treating the binary responses as continuous values. According to simulations, if only the order of the magnitudes of regression coefficients are of interest, the Fabs can be reliable and computationally efficient especially when the sample size is small.

We also notice that the sure screening property of the SIS requires the condition that the largest eigenvalue of $V_K(r)$ should be smaller than cn^τ for some $\tau \geq 0$ and $c > 0$. [43, 44] proved that the covariance matrix of a stationary series satisfies this condition. So V_F , V_M or V_A guarantees the sure screening property of the SIS because the eigenvalue condition is held. However, the largest eigenvalue of $V_C(r)$ is $O(p)$ so the SIS would fail under $V_C(r)$ and $\log(p) = O(n^\xi)$ for $0 < r < 1$ and $0 < \xi < 1$.

Last, we address the computational complexity of the Fabs. As demonstrate in Appendix C, the major burden of the Fabs is the SVD of the design matrix. According to [45], the computational complexity of one SVD is $O(p^2n)$ when $p > n$. On the other hand, for the SIS, the computational complexity of the i th screening statistic is $O(2^2n)$ when using iteratively re-weighted least squares and hence the total complexity for calculating p screening statistics is $O(pn)$ as long as the algorithm always converges in a finite step. Additionally, consider Bayesian joint-screening procedures with MCMC algorithms and take the component-wise Gibbs sampler proposed in [46] as an illustration. A single round of Gibbs sampling requires $O(pn)$ flops so this approach cannot be faster than the Fabs when the required MCMC steps are larger than $O(p)$. To sum up, according to our experience, using Bayesian variable selection approaches reviewed in [42] to screen out unimportant covariates is computationally cumbersome, in general, and, therefore, we advocate using Fabs for binary response variable screening when there is no evidence of nonlinear relationships.

Acknowledgments

Funding

Chang's research is partially supported by Ministry of Science and Technology (MOST) of Taiwan under Grant 103-2118-M-006-004-. Tzeng's research is supported by NIH P01 CA142538. Chen's research is partially supported by Ministry of Science and Technology (MOST) of Taiwan 103-2118-M-006-002-MY2 and the Mathematics Division of the National Center for Theoretical Sciences in Taiwan.

Appendix A. Proof of Proposition 2.1

Consider the probit regression likelihood $\prod_{i=1}^n \Phi(y_i x_i^T \beta)$ with the prior distribution $\phi(\beta; \beta_0, D)$. The corresponding moment generating function of the posterior distribution, $M_\beta(t)$, can be written as

$$\begin{aligned} M_\beta(t) &\propto \int_{\mathbb{R}^p} e^{\beta^T t} \left[\prod_{i=1}^n \Phi(y_i x_i^T \beta) \right] \phi(\beta; \beta_0, D) d\beta \\ &= e^{\beta_0^T t + t^T D t / 2} \int_{\mathbb{R}^p} \left[\prod_{i=1}^n \Phi(y_i x_i^T \beta) \right] \phi(\beta; \beta_0 + D t, D) d\beta \\ &= e^{\beta_0^T t + t^T D t / 2} \int_{\mathbb{R}^p} \phi(s; 0, I_p) \prod_{i=1}^n \Phi(y_i x_i^T [\beta_0 + D t + D^{1/2} S]) ds \end{aligned} \quad (A1)$$

where $s = D^{-1/2}(\beta - \beta_0 - D t)$. According to Arnold and Beaver (2000), the integral of the last equation above can be expressed as

$$\Pr \left(\bigcap_{i=1}^n [y_i x_i^T (\beta_0 + D t + D^{1/2} S) > u_i] \right) = \Pr \left([Y X D^{1/2}, -I_n] U > -Y X [\beta_0 + D t] \right)$$

where $U^T = (s^T, u_1, \dots, u_n)$, $s^T = (s_1, \dots, s_p)$ and s_i 's and u_i 's independently follow standard normal distribution. Therefore, because $U \sim N(0, I_{n+p})$ and

$$[Y X D^{1/2}, -I_n] [Y X D^{1/2}, -I_n]^T = I_n + Y X D X^T Y$$

is symmetric and positive definite, the random vector $z = [Y X D^{1/2}, -I_n] U$ follows $N(0, I_n + Y X D X^T Y)$. Let $\Delta = I_n + Y X D X^T Y$. Then,

$$\begin{aligned} M_\beta(t) &\propto e^{\beta_0^T t + t^T D t / 2} \\ &\quad \times \int_{-y_1 x_1^T (\beta_0 + D t)}^\infty \cdots \int_{-y_n x_n^T (\beta_0 + D t)}^\infty \phi(z; 0, \Delta) dz_n \cdots dz_1 \\ &= e^{\beta_0^T t + t^T D t / 2} \int_0^\infty \cdots \int_0^\infty \phi(w; Y X (\beta_0 + D t), \Delta) dw_n \cdots dw_1 \\ &= e^{\beta_0^T t + t^T D t / 2} \int_{\mathbb{R}_+^n} \phi(w; Y X (\beta_0 + D t), \Delta) dw \end{aligned}$$

where $w = z + Y X (\beta_0 + D t)$.

Similarly, for the logistic regression, the likelihood function is $\prod_{i=1}^n L(y_i x_i^T \beta)$. Given prior distribution $\phi(\beta; \beta_0, D)$, the corresponding moment generating function of the posterior distribution is proportional to

$$e^{\beta_0^T t + t^T D t / 2} \int_{\mathbb{R}^p} \phi(s; 0, I_p) \prod_{i=1}^n L(y_i x_i^T [\beta_0 + D t + D^{1/2} S]) ds \quad (A2)$$

because of (A1). According to Andrews and Mallows (1974) and Stefanski (1990), $L(x_i) = \int_{\mathbb{R}^+} \Phi(x_i/\sigma_i) h(\sigma_i) d\sigma_i$ ($i = 1, \dots, n$) where $\sigma_i/2$ follows the Kolmogorov–Smirnov distribution and $h(\cdot)$ is its density function. Define $\boldsymbol{\sigma}^T = (\sigma_1, \dots, \sigma_n)$, and $h(\boldsymbol{\sigma}) = \prod_{i=1}^n h(\sigma_i)$. Then, (A2) can be represented as

$$\int_{\mathbb{R}_+^n} f(\boldsymbol{\sigma}) h(\boldsymbol{\sigma}) d\boldsymbol{\sigma}$$

where

$$f(\boldsymbol{\sigma}) = \int_{\mathbb{R}^p} \phi(\mathbf{w}; \mathbf{0}, \mathbf{I}_p) \prod_{i=1}^n \int_{\mathbb{R}_+} \Phi(y_i \mathbf{x}_i^T [\boldsymbol{\beta}_0 + \mathbf{D} \mathbf{t} + \mathbf{D}^{1/2} \mathbf{w}] / \sigma_i) d\mathbf{w}.$$

Following Arnold and Beaver (2000), the inner integral can be expressed as

$$\Pr \left(\bigcap_{i=1}^n \left[y_i \mathbf{x}_i^T (\boldsymbol{\beta}_0 + \mathbf{D} \mathbf{t} + \mathbf{D}^{1/2} \mathbf{s}) > \sigma_i u_i \right] \right) = \Pr \left(\left[\mathbf{Y} \mathbf{X} \mathbf{D}^{1/2}, -\sum \right] \mathbf{U} > -\mathbf{Y} \mathbf{X} [\boldsymbol{\beta}_0 + \mathbf{D} \mathbf{t}] \right)$$

where $\boldsymbol{\Sigma} = \text{diag}\{\sigma_i\}$. Consequently, the moment generating function is proportional to

$$e^{\boldsymbol{\beta}_0^T \mathbf{t} + \mathbf{t}^T \mathbf{D} \mathbf{t} / 2} \int_{\mathbb{R}_+^n} \int_{\mathbb{R}_+^n} \phi(\mathbf{w}; \mathbf{Y} \mathbf{X} (\boldsymbol{\beta}_0 + \mathbf{D} \mathbf{t}), \sum^2 + \mathbf{Y} \mathbf{X} \mathbf{D} \mathbf{X}^T \mathbf{Y}) h(\boldsymbol{\sigma}) d\boldsymbol{\sigma} d\mathbf{w}.$$

Appendix B. Proof of Proposition 3.2

We firstly consider the Bayesian probit regression model. When the generalized g-prior is applied, we have $\mathbf{X} \mathbf{D}_\varepsilon \mathbf{X}^T = \mathbf{I}_n$ and thus

$$\Pr \left(\left[\mathbf{Y} \mathbf{X} (g \mathbf{D}_\varepsilon)^{1/2}, -\mathbf{I}_n \right] \mathbf{U} > -\mathbf{Y} \mathbf{X} [\boldsymbol{\beta}_0 + (g \mathbf{D}_\varepsilon) \mathbf{t}] \right) = \prod_{i=1}^n \Phi \left(\frac{y_i \mathbf{x}_i^T [\boldsymbol{\beta}_0 + g \mathbf{D}_\varepsilon \mathbf{t}]}{\sqrt{1+g}} \right).$$

Consequently, the cumulant generating function of the posterior distribution is

$$S_\beta(\mathbf{t}) = \log(M_\beta(\mathbf{t})) = \mathbf{t}^T \boldsymbol{\beta}_0 + \frac{1}{2} \mathbf{t}^T (g \mathbf{D}_\varepsilon) \mathbf{t} + \sum_{i=1}^n \log \left(\Phi \left(\frac{y_i \mathbf{x}_i^T [\boldsymbol{\beta}_0 + g \mathbf{D}_\varepsilon \mathbf{t}]}{\sqrt{1+g}} \right) \right) + \log(C)$$

where C is the norming constant which is free of $\mathbf{t}$. Consequently, the posterior mean is

$$\boldsymbol{\beta}_0 + \frac{g}{\sqrt{1+g}} \mathbf{D}_\varepsilon \mathbf{X}^T \mathbf{Q} \mathbf{y} \quad \text{where} \quad \mathbf{Q} = \text{diag} \left\{ \frac{\phi(y_i \mathbf{x}_i^T \boldsymbol{\beta}_0 / \sqrt{1+g})}{\Phi(y_i \mathbf{x}_i^T \boldsymbol{\beta}_0 / \sqrt{1+g})} \right\}.$$

Secondly, for the logistic regression model, the moment generating function of the posterior distribution is proportional to $\exp\{\mathbf{t}^T \boldsymbol{\beta}_0 + \mathbf{t}^T \mathbf{D}_\varepsilon \mathbf{t} \times g/2\}$ times the factor

$$\int_{\mathbb{R}_+^n} \left[\int_{\mathbb{R}_+^n} \phi(\mathbf{w}; \mathbf{Y} \mathbf{X} (\beta_0 + g \mathbf{D}_\varepsilon \mathbf{t}), g \mathbf{I}_n + \sum^2) d\mathbf{w} \right] h(\sigma) d\sigma = \prod_{i=1}^n \left\{ \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \phi(w_i; \mu_i, \sigma_i^2 + g) dw_i h(\sigma_i) d\sigma_i \right\}$$

where $\mu_i = y_i x_i^T (\beta_0 + g \mathbf{D}_\varepsilon \mathbf{t})$. Moreover,

$$\begin{aligned} \int_{\mathbb{R}_+(\sigma_i)} \int_{\mathbb{R}_+(w_i)} \phi(w_i; \mu_i, \sigma_i^2 + g) h(\sigma_i) dw_i d\sigma_i &= \int_{\mathbb{R}_+(\sigma_i)} \int_{\mathbb{R}_+(w_i)} \int_{\mathbb{R}(y)} \phi(w_i; y, g) \phi(y; \mu_i; \sigma_i^2) h(\sigma_i) dy dw_i d\sigma_i = \int_{\mathbb{R}_+(w_i)} \int_{\mathbb{R}(x)} \left\{ \int_{\mathbb{R}_+(\sigma_i)} \phi(w_i; \mu_i, \sigma_i^2 + g) h(\sigma_i) d\sigma_i \right\} dx \\ &= \int_{\mathbb{R}_+(w_i)} \int_{\mathbb{R}(x)} \phi(w_i; x + \mu_i, g) l(x) dx dw_i \\ &= \int_{\mathbb{R}(x)} \left\{ \int_{-\infty}^{\infty} \phi(z; 0, g) dz \right\} l(x) dx \\ &= \int_{\mathbb{R}(x)} \Phi((x + \mu_i)/\sqrt{g}) l(x) dx. \end{aligned}$$

Therefore, the corresponding cumulant generating function is

$$\beta_0^T \mathbf{t} + \mathbf{t}^T \mathbf{D}_\varepsilon \mathbf{t} \times \frac{g}{2} + \log(C_L) \sum_{i=1}^n \log \left(\int_{\mathbb{R}(x)} \Phi \left(\frac{x + y_i x_i^T (\beta_0 + g \mathbf{D}_\varepsilon \mathbf{t})}{\sqrt{g}} \right) l(x) dx \right)$$

where C_L is free of $\mathbf{t}$. The moments of the posterior distribution are therefore tenable by taking the derivatives of above function with respect to $\mathbf{t}$.

Appendix C. R Code for computing Fabs

The Fabs statistic can be implemented easily if the computing language provides functions for singular value decomposition and matrix multiplication. For instance, in R language (The R core team, 2013), it requires only three lines:

```
## X: the design matrix containing the intercept
## y: the response vector with elements 1 or -1
Z <- svd(X) # singular value decomposition
Fabs <- Z$V %>% diag(1/Z$d) %>% t(Z$u) %>% y
Fabs <- Fabs[-1] # remove the intercept term
```

References

1. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B*. 1996; 58:267–288.
2. Zou H. The adaptive lasso and its oracle properties. *J Am Statist Assoc*. 2006; 101:1418–1429.
3. Fan J, Li R. Variable selection via non-concave penalized likelihood and its oracle properties. *J Am Statist Assoc*. 2001; 96:1348–1360.
4. Candès E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann Statist*. 2007; 35:2313–2044.
5. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J Royal Statist Assoc B*. 2008; 70:849–911.
6. Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann Statist*. 2010; 38:3567–3604.

7. Zhu LP, Li L, Li R, Zhu LX. Model-free feature screening for ultrahigh-dimensional data. *J Am Statist Assoc.* 2011; 106:1464–1475.
8. Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *J Am Statist.* 2012; 107:1129–1139.
9. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Statist.* 2007; 35:2769–2794.
10. Mai Q, Zou H. The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika.* 2013; 100:229–234.
11. Mai Q, Zou H. The fused Kolmogorov filter: a nonparametric model-free screening method. *Ann Statist.* 2015; 43:1471–1497.
12. Huang J, Horowitz J, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann Statist.* 2008; 36:587–613.
13. White H. Maximum likelihood estimation of misspecified models. *Econometrica.* 1982; 50:1–26.
14. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika.* 1984; 71:1–10.
15. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika.* 1993; 80:27–38.
16. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statist Med.* 2002; 21:2409–2419.
17. Chang SM, Chen RB, Chi Y. Bayesian variable selection for probit model with componentwise Gibbs samplers. *Commun Stat-Simul Comput.* 2016; 45:2752–2766.
18. Breslow N, Clayton D. Approximate inference in generalized linear mixed models. *J Am Statist Assoc.* 1993; 88:9–25.
19. Breslow N, Lin X. Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika.* 1995; 82:81–91.
20. Lin X, Breslow N. Bias correction in generalized linear mixed models with multiple components of dispersion. *J Am Statist Assoc.* 1996; 91:1007–1016.
21. Raudenbush S, Yang ML, Yosef M. Maximum likelihood for generalized linear models with nested random effects via high-order multivariate Laplace approximations. *J Compu Graph Statist.* 2000; 9:141–157.
22. Zellner A. Applications of Bayesian analysis in Econometrics. *Statistician.* 1983; 32:23–34.
23. Baragatti M, Pommeret D. 2012. A study of variable selection using g-prior distribution with ridge parameter. *J Comput Statist Data Anal.* 2012; 56:1920–1934.
24. Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. *Bioinformatics.* 2003; 19:90–97. [PubMed: 12499298]
25. Yang AJ, Song XY. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics.* 2010; 26:215–222. [PubMed: 19933163]
26. Andrews DF, Mallows CL. Scale mixtures of normal distributions. *J Royal Statist Asso B.* 1974; 36:99–102.
27. Stefanski LA. A normal scale mixture representation of the logistic distribution. *Statist Probabil Lett.* 1990; 11:69–70.
28. Feller W. On the Kolmogorov–Smirnov limit theorems for empirical distributions. *Ann Math Statist.* 1949; 19:1977–1989.
29. Amemiya T. Multivariate regression and simultaneous equations models when the dependent variables are truncated normal. *Econometrica.* 1974; 42:999–1012.
30. Leppard P, Tallis GM. Algorithm AS249: evaluation of the mean and covariance of the truncated multinormal distribution. *Appl Statist.* 1989; 38:543–55.
31. Tallis GM. The moment generating function for the truncated multi-normal distribution. *J Roy Stat Soc B.* 1961; 23:223–229.
32. Wihelm S, Manjunath BG. tmvtnorm: a package for the truncated multivariate normal distribution. *R J.* 2010:25–29.
33. Arnold BC, Beaver RJ. Hidden truncation models. *Sankhya.* 2000; 62:23–35.

34. Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g-priors for Bayesian variable selection. *J Am Statist Assoc.* 2008; 103:410–423.
35. R Core Team. *R: A Language and Environment for Statistical Computing.* 2013
36. Chang SM. Double shrinkage estimators for large sparse covariance matrices. *J Statist Compu Simu.* 2015; 85:1497–1511.
37. Heinze G, Ploner M, Dunkler D, Southworth H. logistf: Firths bias reduced logistic regression. *R J.* 2013
38. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Meirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.* 1999; 286:531–537. [PubMed: 10521349]
39. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Statist Assoc.* 2002; 97:77–87.
40. Albert JH, Chib S. Bayesian analysis of binary and polychotomous choice data. *J Amer Statist Assoc.* 1993; 88:669–679.
41. George EI, McCulloch RE. Variable selection via Gibbs sampling. *J Am Statist Assoc.* 1993; 7:338–373.
42. O’Hara RBO, Sillanpää MJ. A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* 2009; 4:85–118.
43. Bickel PJ, Levina E. Some theory for Fisher’s linear discriminant function, “naive Bayes”, and some alternatives when there are many more variables than observations. *Bernoulli.* 2004; 10:989–1010.
44. Bickel PJ, Levina E. Regularized estimation of large covariance matrices. *Ann Statist.* 2008; 36:199–227.
45. Trefethen, LN., David, B. *Numerical Linear Algebra.* Philadelphia: Society for Industrial and Applied Mathematics; 1997.
46. Chen RB, Chu CH, Lai TY, Wu Y. Stochastic matching pursuit for Bayesian variable selection. *Statist Comput.* 2011; 21:247–259.

Table 1

With mean model M1, percentiles of the smallest number required to include all active covariates in screening with $n = 200$ and $p = 1000$ under various covariance structures. The number of active covariates was 4 and the result was based on 500 replications. For each column, the boldface number is the largest number no greater than $38 \approx 200/\log(200)$.

	V_I						V_M					
	SIS	ISIS	KF	DC	Fabs	SIS	ISIS	KF	DC	Fabs	SIS	Fabs
5%	4	4	4	4	4	4	4	4	4	4	4	4
25%	4	4	4	4	4	4	4	4	4	4	4	4
50%	4	4	5	4	4	4	4	5	4	4	4	4
75%	4	4	8	5	5	5	5	9	5	5	5	5
95%	7	5	25	9	11	13	5	46	15	12		

	V_A						V_C					
	SIS	ISIS	KF	DC	Fabs	SIS	ISIS	KF	DC	Fabs	SIS	Fabs
5%	4	4	4	4	4	4	4	4	6	4	4	4
25%	4	4	6	4	4	4	5	18	7	4		
50%	6	5	13	7	5	16	5	56	24	5		
75%	11	5	38	16	10	90	6	201	113	9		
95%	98	5	152	91	47	490	27	19	597	46		

Table 2

With mean model M2, percentiles of the smallest number required to include all active covariates in screening with $n = 200$ and $p = 1000$ under various covariance structures. The number of active covariates was 4 and the result was based on 500 replications. For each column, the boldface number is the largest number no greater than $38 \approx 200/\log(200)$.

	V_I						V_M					
	SIS	ISIS	KF	DC	Fabs	SIS	ISIS	KF	DC	Fabs	SIS	Fabs
5%	4	4	4	4	4	4	4	4	4	4	4	4
25%	4	4	4	4	4	4	5	8	5	5	5	5
50%	4	4	6	4	4	7	5	17	8	8	8	8
75%	5	5	14	6	6	17	5	52	21	24	5	5
95%	17	5	79	29	33	131	6	219	143	154	6	6

	V_A						V_C					
	SIS	ISIS	KF	DC	Fabs	SIS	ISIS	KF	DC	Fabs	SIS	Fabs
5%	4	4	4	4	4	4	913	5	816	931	4	4
25%	5	5	8	5	4	1000	5	996	999	5	5	5
50%	10	5	24	13	5	1000	5	1000	1000	8	5	5
75%	43	5	87	54	9	1000	636	1000	1000	21	5	5
95%	257	5	464	331	48	1000	947	1000	1000	112	5	5

Table 3

With mean model M3, percentiles of the smallest number required to include all active covariates in screening with $n = 200$ and $p = 1000$ under various covariance structures. The number of active covariates was 4 and the result was based on 500 replications. For each column, the boldface number is the largest number no greater than $38 \approx 200/\log(200)$.

	V_I				V_M					
	SIS	ISIS	KF	DC	Fabs	SIS	ISIS	KF	DC	Fabs
5%	106	132	76	62	122	50	55	8	5	90
25%	394	455	297	240	457	316	304	40	14	383
50%	655	692	514	448	684	569	591	92	37	615
75%	832	835	714	690	866	800	828	186	93	822
95%	972	963	921	912	968	972	966	468	256	966

	V_A				V_C					
	SIS	ISIS	KF	DC	Fabs	SIS	ISIS	KF	DC	Fabs
5%	4	4	4	4	11	58	415	35	30	151
25%	21	324	6	5	117	226	710	125	126	425
50%	104	618	20	11	362	442	855	250	264	687
75%	405	822	73	31	677	640	943	417	456	863
95%	860	955	238	121	947	871	990	760	768	970

Table 4

Comparisons of ISIS and Fabs under various values of p . Percentiles of the smallest number required to include all active covariates in screening with $n = 200$ and $p = 2, 000 \sim 10, 000$ under covariance structures V_I and V_C . The number of active covariates was 4, the underlying mean model was M2 and the result was based on 500 replications.

V_I												
	ISIS ₄					Fabs						
	2000	4000	6000	8000	10000	2000	4000	6000	8000	10000		
5%	4	4	4	4	4	4	4	4	4	4	4	4
25%	4	4	4	4	4	4	4	4	4	4	4	4
50%	5	5	5	5	5	5	5	5	6	6	7	7
75%	5	5	5	5	5	8	10	13	15	23		
95%	5	6	7	9	8	49	69	108	111	150		

V_C												
	ISIS ₄					Fabs						
	2000	4000	6000	8000	10000	2000	4000	6000	8000	10000		
5%	5	5	5	5	5	4	4	4	4	4	4	4
25%	5	5	5	1100	1787	5	6	7	9	9		
50%	578	2026	3592	5226	6252	10	15	22	25	26		
75%	1461	3200	5147	6966	8588	26	53	81	102	119		
95%	1921	3884	5856	7829	9856	153	318	463	702	870		

Table 5

Small sample comparison. Percentiles of the smallest number required to include all active covariates in screening with $n = 72$ and $p = 1,000$ under models M1 and M2 and covariance structures V_I and V_C . The number of active covariates was 4, and the result was based on 500 replications.

	V_I					V_C				
	SIS	ISIS	ISIS _F	Fabs	SIS	ISIS	ISIS _F	Fabs	SIS	ISIS
M1	5%	8	8	8	9	27	22	21	18	18
	25%	22	44	30	24	116	217	212	52	52
	50%	50	158	106	53	330	740	740	124	124
	75%	130	417	312	142	744	951	955	272	272
	95%	426	860	779	453	970	994	996	633	633
M2	5%	7	7	7	6	438	18	25	17	17
	25%	24	57	36	25	922	841	844	67	67
	50%	70	198	140	73	991	936	939	158	158
	75%	167	470	337	171	999	975	976	326	326
	95%	454	862	804	475	1000	996	996	782	782

Table 6

Small sample comparison. Percentiles of the smallest number required to include all active covariates in screening with $n = 72$ and $p = 3,000$ under models M1 and M2 and covariance structures V_I and V_C . The number of active covariates was 4, and the result was based on 500 replications.

	V_I					V_C				
	SIS	ISIS	ISIS _F	Fabs	SIS	ISIS	ISIS _F	Fabs	SIS	Fabs
M1	5%	13	40	30	14	81	146	156	49	49
	25%	54	382	258	53	409	1676	1677	149	149
	50%	140	909	709	145	1066	2652	2642	350	350
	75%	387	1818	1592	360	2150	2926	2934	731	731
	95%	1198	2723	2661	1198	2903	2992	2992	2070	2070
M2	5%	18	57	29	19	1279	1948	1931	42	42
	25%	73	368	287	69	2828	2669	2693	174	174
	50%	172	963	688	172	2968	2887	2888	428	428
	75%	447	1845	1556	504	2995	2958	2955	1077	1077
	95%	1674	2756	2671	1795	3000	2990	2991	2338	2338

Table 7

Small sample comparison. Percentiles of the smallest number required to include all active covariates in screening with $n = 72$ and $p = 7,000$ under models M1 and M2 and covariance structures V_I and V_C . The number of active covariates was 4, and the result was based on 500 replications.

		V_I					V_C				
		SIS	ISIS	ISIS _F	Fabs	SIS	ISIS	ISIS _F	Fabs		
M1	5%	33	354	244	33	213	1372	1329	90		
	25%	110	1617	1280	118	836	5391	5253	302		
	50%	334	3276	2758	344	2252	6597	6561	795		
	75%	824	5126	4812	842	5154	6914	6916	1647		
	95%	2318	6586	6466	2358	6774	6989	6989	5009		
M2	5%	38	236	213	37	2780	4961	5146	99		
	25%	153	1328	1010	148	6452	6426	6462	406		
	50%	390	3010	2391	412	6926	6786	6782	932		
	75%	1156	4867	4244	1180	6989	6927	6923	2088		
	95%	3855	6548	6368	3784	7000	6985	6979	5014		

Table 8

Ranks of putative genes affecting leukemia types

Gene Name	YS	Lee	SIS	ISIS	Fabs	DC	KF
(Positive Controls)							
Zyxin	1	4	1	1	7	10	1
Cystatin C	2	1	5	5	11	5	3
leptin receptor overlapping transcript	3	–	63	2760	14	102	137
CD33 molecule	4	–	6	6	3	2	5
amyloid beta (A4) precursor-like protein 2	5	–	3	3	2	9	2
cathepsin A	6	–	26	2231	13	36	27
cystatin A (stefin A)	–	2	19	610	57	8	12
complement factor D (adipsin)	–	3	28	6906	15	3	9
spectrin, alpha, non-erythrocytic 1	–	5	13	58	19	16	40
Top κ_0 genes to cover positive control genes							
			63	6906	57	102	137

YS: Yang and Song (2010); Lee: Lee et al. (2003).