An Algorithm for Bayesian Variable Selection in High-dimensional Generalized Linear Models

Vitara Pungpapong
Department of Statistics, Faculty of Commerce and Accountancy
Chulalongkorn Unversity, Bangkok, THAILAND
email: vitara@cbs.chula.ac.th

Abstract

Inspired by analysis of genomic data, the primary quest is to identify associations between studied traits and genetic markers where number of markers is typically much larger than sample size. Bayesian variable selection methods with Markov chain Monte Carlo (MCMC) are extensively applied to analyze such high-dimensional data. However, MCMC is often slow to converge with large number of candidate predictors. In this study, we examine the empirical Bayes variable selection with a sparse prior on the unknown coefficients. An iterated conditional modes/medians (ICM/M) algorithm is proposed for implementation by iteratively minimizing a conditional loss function in high-dimensional linear regression model. Attention is then directed to extend the algorithm to a generalized linear model. The performances of our approach are evaluated through simulation study.

Keywords: Bayesian inference, high-dimensional data, sparse variables, generalized linear models

1. Introduction

Linear and nonlinear models have been extensively used to identify associations between response and explanatory variables. The advent of high throughput technologies enables the collection of massive and complex data in biomedical science. High-dimensional data (i.e. number of predictors is larger than sample size) have posed a challenge in selecting variables for such models.

Tibshirani (1996) proposed lasso by putting an ℓ_1-penalty on likelihood function producing sparse coefficients. Lasso cannot be applied to only a linear model but has also extended to generalized linear models (GLMs). Methods to determine the entire path of lasso for GLMs were discussed in Efron et. al. (2004), Park and Hastie (2007), Friedman et. al. (2010).

Johnstone and Silverman (2004) proposed empirical Bayes thresholding by putting a mixture prior of an atom of probability at zero and a heavy-tailed density. Pungpapong et. al. (2012) presented a contribution of empirical Bayes thresholding to select variables in linear regression framework. An iterated conditional modes/medians (ICM/M) is introduced for fast and easy-to-implement algorithm.

The aim of this study is to generalize empirical Bayes variable selection (Pungpapong et. al., 2012) to more class beyond linear model. The details of ICM/M algorithm for GLMs are also described. The ability in selecting variables is demonstrated in simulation study.

2. Methods

Empirical Bayes variable selection in linear model

Consider a normal linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n). \tag{1}$$

$\mathbf{Y}$ is a $n \times 1$ random vector, $\mathbf{X}$ is a $n \times p$ matrix of p predictors, and I_n is a $n \times n$ identity matrix. $\mathbf{Y}$ is further assumed to be centralized and $\mathbf{X}$ to be standardized. In order to perform variable selection, Pungpapong et. al. introduced the following independent mixture prior distribution to model the sparsity of β_j:

$$\beta_j | \sigma \sim (1-\omega)\delta_0(\beta_j) + \omega\gamma_\alpha(\beta_j | \sigma), \tag{2}$$

where $\delta_0(.)$ is a Dirac delta function at zero and $\gamma_\alpha(.|\sigma)$ is a probability function. Under this prior distribution, each β_j is zero with probability $(1-\omega)$ and β_j is drawn from the nonzero part of prior $\gamma_\alpha(.|\sigma)$ with probability ω. Johnstone and Silverman suggested a heavy-tailed density such as Laplace distribution for $\gamma_\alpha(.|\sigma)$

$$\gamma_\alpha(\beta_j|\sigma) = \frac{\alpha\sqrt{n-1}}{2\sigma}\exp\left(-\frac{\alpha\sqrt{n-1}}{\sigma}|\beta_j|\right), \tag{3}$$

where $\alpha > 0$ is a scale parameter. The default value of $\alpha = 0.5$ is used throughout the paper as suggested by Johnstone and Silverman. Jeffrey's prior (Jeffreys, 1946) is taken on σ^2.

When the information of structural relationship among predictors is available, it is more sensible to incorporate such information in the prior. Pungpapong et. al. (2012) introduced an indicator variable $\tau = (\tau_1, \dots, \tau_p)^t$ where $\tau_j = 1_{\{\beta_j \neq 0\}}$ and an underlying structure among τ can be represented by an undirected graph $G = (V, E)$ comprising a set V of vertices and a set E of edges. Specifically, the prior distribution is set to be dependent to τ,

$$\beta_j | \tau_j \sim (1-\tau_j)\delta_0(\beta_j) + \tau_j\gamma_\alpha(\beta_j|\sigma). \tag{4}$$

The following Ising model (Onsagar, 1943) is then employed to model a priori information on τ,

$$P(\tau) = \frac{1}{Z(a,b)}\exp\left\{a\sum_j \tau_j + b\sum_{<j,k>\in E}\tau_j\tau_k\right\}, \tag{5}$$

where a and b are two parameters and $Z(a, b)$ is a normalizing constant.

ICM/M Algorithm

Pungpapong et. al. (2012) presented an iterated conditional modes/medians (ICM/M) algorithm for fast computation of empirical Bayes variable selection. Data-driven optimal values for hyperparameters and auxiliary parameters are obtained as the modes of their full conditional distribution functions. Each regression coefficient is obtained as the median of its full conditional distribution function. With the presented Bayesian framework, the use of conditional medians can enforce the variable selection and estimation simultaneously. The iterative procedure for updating regression coefficients and other parameters is carried out until convergence.

The ICM/M algorithm for empirical Bayes variable selection with independent prior on coefficients has the following steps:

1. Obtain initial estimates of all parameters denoted as $\hat{\beta}$, $\hat{\sigma}^2$.

2. Estimate hyperparameter ω with the value of $\hat{\omega}$ as the mode of its full conditional distribution function.
$$\hat{\omega} = \text{mode}(\omega|\mathbf{Y}, \mathbf{X}, \hat{\beta}, \hat{\sigma}^2)$$
3. Update each regression coefficient $\hat{\beta}_j$ as the median of its full conditional distribution function.
$$\hat{\beta}_j \; = \text{median}\left(\beta_j \middle| \mathbf{Y}, \mathbf{X}, \hat{\beta}_{-j}, \hat{\sigma}^2, \hat{\omega}\right), \qquad j = 1, \dots, p,$$
where $\hat{\beta}_{-j} = (\hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p)$.
4. Update $\hat{\sigma}^2$ as the mode of its full conditional distribution function.
$$\hat{\sigma}^2 = \text{mode}(\sigma^2|\mathbf{Y}, \mathbf{X}, \hat{\beta}, \hat{\omega})$$
5. Iterate between steps 2-4 until convergence.

With Ising prior for structured predictors, the ICM/M algorithm is implemented involving steps as follows:

1. Obtain initial estimates of all parameters denoted as $\hat{\beta}$, $\hat{\sigma}^2$.
2. Obtain the indicator variable $\hat{\tau} = \left(\hat{\tau}_1, \dots, \hat{\tau}_p\right)^t$ where $\hat{\tau}_j = 1_{\{\hat{\beta}_j \neq 0\}}$.
3. With current value of $\hat{\tau}$, estimate hyperparameters (a, b) with the values of $(\hat{a}, \hat{b})$ as the mode of its pseudo-likelihood function.
$$\left(\hat{a}, \hat{b}\right) = \text{mode}\left\{\prod_{i=1}^{p} P\left(\hat{\tau}_j \middle| \hat{\tau}_{-j}; a, b\right)\right\}$$
$$= \text{mode}\left\{\prod_{i=1}^{p} P(\hat{\tau}_j|\{\hat{\tau}_k : <\hat{\tau}_j, \hat{\tau}_k> \in E\}; a, b)\right\}$$
The general setup of ICM/M algorithm suggests that $(\hat{a}, \hat{b})$ are the modes of their full conditional distribution function which is the prior likelihood. The reason for using the pseudo-likelihood function instead is to avoid computational difficulty in computing an unknown normalizing constant in Ising model.
4. Update each regression coefficient $\hat{\beta}_j$ as the median of its full conditional distribution function.
$$\hat{\beta}_j \; = \text{median}\left(\beta_j \middle| \mathbf{Y}, \mathbf{X}, \hat{\beta}_{-j}, \hat{\sigma}^2, \hat{a}, \hat{b}\right), \qquad j = 1, \dots, p,$$
where $\hat{\beta}_{-j} = (\hat{\beta}_1, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_p)$.
5. Update $\hat{\sigma}^2$ as the mode of its full conditional distribution function.
$$\hat{\sigma}^2 = \text{mode}(\sigma^2|\mathbf{Y}, \mathbf{X}, \hat{\beta}, \hat{a}, \hat{b})$$
6. Iterate between steps 2-5 until convergence.

Extension to GLMs

Generalized linear model (GLM) is an extension of the ordinary linear model. Each component Y_i is assumed to be independently distributed with mean $\mu_i = E[Y_i]$ and variance $Var[Y_i]$. Y_i is further assumed to have a distribution in the exponential family taken the form
$$f_{Y_i}(y_i|\theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}. \tag{6}$$
$a(.)$, $b(.)$, and $c(.)$ are functions which vary according to distributions. $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ is known as the canonical parameter. Here we assume a fixed

known dispersion parameter as $a(\phi) = \phi = 1$. A link function $g(.)$ is introduced to connect the linear predictor $\eta_i = \mathbf{X}_i\beta$ to the mean μ_i. That is, $g(\mu_i) = \eta_i = \mathbf{X}_i\beta$.

A quadratic approximation to the likelihood is employed to construct pseudodata and pseudovariances. The basic idea is to approximate the generalized linear model by a linear model. Such a procedure is equivalent to an iteratively reweighted least squares (IRLS) algorithm. Particularly, based on the current parameter estimates at k-th iteration, $\hat{\beta}^{(k)}$, pseudodata $\mathbf{Z} = (Z_1, \dots, Z_n)^t$ and pseudovariances $\Sigma = \text{diag}\{\sigma_i^2\}$ are constructed as

$$\mathbf{Z}^{(k)} = \hat{\eta}^{(k)} + (\mathbf{Y} - \hat{\mu}^{(k)})\left(\frac{\partial\eta}{\partial\mu}\right)_{(k)} \tag{7}$$

and

$$\Sigma^{(k)} = \text{diag}\{\sigma_i^2\} = \text{diag}\left\{Var[Y_i]\left(\frac{\partial\eta}{\partial\mu}\right)^2_{(k)}\right\}, \tag{8}$$

where $\hat{\eta}^{(k)} = \mathbf{X}_i\hat{\beta}^{(k)}$ and the derivative is evaluated at $\hat{\mu}^{(k)}$. The approximated distribution of pseudodata $\mathbf{Z}^{(k)}$ is $N(\mathbf{X}\hat{\beta}^{(k)}, \Sigma^{(k)})$.

Borrowing the idea of IRLS, pseudodata is treated as a working response. After obtaining pseudodata and pseudovariances, the ICM/M algorithm is used to update regression coefficients and other parameters. The procedure consists of the outer and inner loop. The outer loop is taken place to update pseudodata and pseudovariances based on current estimates. The inner loop is where ICM/M is employed to loop through update all parameters. This simple-to-implement algorithm achieves fast computation even in high-dimensional data analysis.

3. Results

To demonstrate the proposed methods, we consider four different cases of large p small n datasets. Our simulations focus on binary and survival data which are particularly interested in analysis of genomic data. Logistic regression GLM was employed to analyze binary data and Cox proportional hazards model (Cox, 1972) was carried out to deal with survival data. In the first three cases, we fixed p = 1,000 and n = 250. The simulations were run 100 times for each setup.

Case 1: Binary data: high/mild correlations in covariates

The binary response was simulated with the probability following logistic regression model:

$$P(Y_i = 1|X_i) = \frac{\exp(\beta_0 + X_i\beta)}{1 + \exp(\beta_0 + X_i\beta)}.$$

The intercept term is zero and the non-zero regression coefficients are $\beta_1 = \cdots = \beta_5$= 10 and $\beta_{11} = \cdots = \beta_{15}$ = -5. The covariates are partitioned into ten blocks, with each block including 100 covariates and they are serially correlated at the same level of ρ. The values of ρ are {0, 0.5, 0.9}.

Case 2: Survival data: high/mild correlations in covariates

The covariates were simulated the same way as in Case 1. Survival times were generated from a Cox model and follow Weibull distribution with shape a parameter υ = 10 and a scale parameter λ = 1. Among 1,000 predictors, the failure rate is determined by a linear combination of 20 non-zero coefficients: $\beta_1 = \cdots = \beta_{10}$ = 5 and $\beta_{101} = \cdots = \beta_{110}$ = 2. The censoring times were generated randomly to achieve 50% of the observed event times.

Case 3: Survival data: Markov chain

Same as Case 2 except that the covariates were generated from $AR(1)$ with different value of ρ in {0, 0.5, 0.9} and the location of non-zero coefficients follows a Markov chain. Particularly, the indicator variables $\tau_1, \ldots, \tau_p$ form a Markov chain with the transition probabilities specified as follows:

$$P(\tau_{j+1} = 0 | \tau_j = 0) = 1 - P(\tau_{j+1} = 1 | \tau_j = 0) = 0.99,$$
$$P(\tau_{j+1} = 0 | \tau_j = 1) = 1 - P(\tau_{j+1} = 1 | \tau_j = 1) = 0.5.$$

The first indicator variable τ_1 was sampled from Bernoulli(0.5). The effect sizes of those non-zero coefficients were drawn from Uniform[0.5,5].

Case 4: Binary data: pathway information

Our simulation is based on sampling of 871 individuals from Parkinson's disease (PD) SNP data (dbGaP study accession number: phs000089.v3.p2). Here we consider 1,152 SNPs affiliating with 341 genes in PD-related metabolic pathways. The phenotype was simulated from logistic regression with 46 non-zero coefficients resided in 15 genes in PD pathway and the effect sizes were drawn from Uniform[1,10].

Here we evaluate the algorithm in terms of false positive and false negative rates. Two different sets of initial values $\hat{\beta}^{(0)}$ for EBVS were considered, one with true values and another one with lasso fit. We also compared the performance of EBVS with lasso. EBVS with Ising prior (EBVS$_\text{i}$) was applied to only Case 3 and Case 4 assuming known structural relationship among predictors.

Table 1: Median false positive rates (with standard errors in parentheses) across 100 datasets.

	ρ	Lasso	EBVS	EBVS	EBVS$_\text{i}$	EBVS$_\text{i}$
			(true)	(lasso)	(true)	(lasso)
	0.0	0.892(1e-3)	0.000(7e-4)	0.000(1e-3)	-	-
Case 1	0.5	0.872(2e-3)	0.000(0.00)	0.000(1e-3)	-	-
	0.9	0.843(3e-3)	0.000(8e-4)	0.000(4e-3)	-	-
	0.0	0.829(2e-3)	0.000(3e-3)	0.000(2e-3)	-	-
Case 2	0.5	0.839(1e-3)	0.000(2e-3)	0.000(7e-4)	-	-
	0.9	0.810(1e-3)	0.000(7e-4)	0.000(3e-3)	-	-
	0.0	0.835(2e-3)	0.000(3e-3)	0.000(0.00)	0.000(3e-3)	0.000(9e-4)
Case 3	0.5	0.839(1e-3)	0.000(3e-3)	0.000(9e-4)	0.000(2e-3)	0.000(3e-3)
	0.9	0.818(2e-3)	0.000(3e-3)	0.053(9e-3)	0.000(8e-4)	0.095(0.01)
Case 4	-	0.819(3e-3)	0.000(1e-3)	0.036(3e-3)	0.000(0.00)	0.038(5e-3)

Table 2: Median false negative rates (with standard errors in parentheses) across 100 datasets.

	ρ	Lasso	EBVS	EBVS	EBVS$_\text{i}$	EBVS$_\text{i}$
			(true)	(lasso)	(true)	(lasso)
	0.0	0.000(3e-5)	0.000(0.00)	0.002(9e-5)	-	-
Case 1	0.5	0.000(3e-5)	0.000(0.00)	0.003(9e-5)	-	-
	0.9	0.001(6e-5)	0.000(1e-5)	0.005(7e-5)	-	-
	0.0	0.000(1e-4)	0.000(0.00)	0.000(3e-4)	-	-
Case 2	0.5	0.000(0.00)	0.000(0.00)	0.000(0.00)	-	-
	0.9	0.000(3e-5)	0.000(0.00)	0.000(9e-5)	-	-
	0.0	0.000(9e-5)	0.000(0.00)	0.001(1e-4)	0.000(0.00)	0.001(1e-4)
Case 3	0.5	0.000(5e-5)	0.000(0.00)	0.001(1e-4)	0.000(0.00)	0.000(9e-5)
	0.9	0.000(1e-4)	0.000(0.00)	0.001(1e-4)	0.000(0.00)	0.001(1e-4)
Case 4	-	0.010(3e-4)	0.000(3e-4)	0.020(3e-4)	0.000(0.00)	0.017(3e-4)

As shown in Table 1 and Table 2, EBVS and EBVS$_\text{i}$ with true coefficients are the clear winners in all cases. However, true coefficients are unknown in practice. Lasso tends to select a large number of non-zero coefficients due to high false positive rates. EBVS and EBVS$_\text{i}$ with lasso fit as initial values can reduce false positive rates dramatically.

The median false positive rates for EBVS(lasso) are all zeros in both Case 1 and Case 2. While its false negative rates are not zero in Case 1, they are all zeros in Case 2 indicating that our method can pick up the correct model.

When comparing the performance between EBVS and EBVS$_i$, we see that they produce similar results in Case 3 with ρ = 0.00 and ρ = 0.05. When ρ = 0.09, EBVS$_i$ has higher median false positive rate. The median false positive rate for EBVS$_i$ is also slightly higher than that for EBVS in Case 4. However, EBVS$_i$ produces smaller false negative rates implying EBVS$_i$ has more power to identify true variables.

4. Discussion

In this study, we have introduced an algorithm for empirical Bayes variable selection in high-dimensional GLMs. We can extend ICM/M algorithm to fit GLMs by adopting pseudodata and pseudovariances as in IRLS. Our algorithm is easy to implement and achieves fast computation empirically.

Our results show that our method seems to work better with survival data than binary data in terms of false negative rate. This might be due to the fact that binary data is less informative than continuous data. We notice that EBVS$_i$ has more power to detect non-zero coefficients than EBVS when the structural information among predictors is available. However, it sacrifices with slightly higher false positive rate when correlations among covariates are high. We see that our method with true coefficients as starting values can select the true model for both EBVS and EBVS$_i$. Therefore, choices of initial values other than lasso should be further studied.

References

Cox, D.R. (1972) “Regression Models and Life-Tables”. *Journal of the Royal Statistical Society, Series B*, 34(2), 187-220.

Efron, B., Hastie, T., Johnstone, I.M., and Tibshirani, R. (2004) “Least angle regression”. *The Annals of Statistics*, 32, 407-499.

Friedman, J., Hastie, T., and Tibshirani, R. (2010) “Regularization paths for generalized linear models via coordinate descent”. *Journal of Statistical Software*,33, 1-22.

Jeffreys, H. (1946) “An invariant form for the prior probability in estimation problems”. *Proceedings of the Royal Society of London Series A*, 196, 453-461.

Johnstone, I.M., and Silverman, B.W. (2004) “Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequence” *The Annals of Statistics*, 32, 1594-1649.

Onsagar, L. (1944) “Crystal statistics. I.A two-dimensional model with an order-disorder transition”. *Physical Review*, 65(3-4), 117-149.

Park, M.Y., and Hastie, T. (2007). “L$_1$-regularization path algorithm for generalized linear models”. *Journal of the Royal Statistical Society Series B*, 69, 659-677.

Pungpapong, V., Zhang, M., and Zhang, D. (2012), “Empirical Bayes variable selection using iterated conditional modes/medians”. Manuscript submitted for publication.

Tibshirani, R. (1996) “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society Series B*, 58, 267-288.