Bayesian Methods for High-Dimensional Variable Selection

Faming Liang

University of Florida
Overview

- MCMC, SAMC, and Parallel MCMC
- Bayesian variable selection for high-dimensional GLMs
- Bayesian variable selection for high-dimensional nonlinear systems
- Revisit of Computational Strategies: split-and-merge
Bayesian variable selection for high-dimensional GLMs
The problem

Let $D^n = \{(x_i, y_i) : x_i \in \mathbb{R}^{P_n}, y_i \in \mathbb{R}, i = 1, \ldots, n\}$ denote a data set of $n$ pairs of $P_n$ predictors and a response, where $P_n$ can increase with the sample size $n$. Suppose that the data can be modeled by a generalized linear model (GLM):

$$\eta = g(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{P_n} x_{P_n},$$

and $Y$ has the density given by

$$f(y|\theta) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}.$$  

Under the situation of small-$n$-large-$P$, variable selection can pose a great challenge on the existing statistical methods.
The penalized likelihood approach is to find a model, i.e., a subset of \( \{x_1, \ldots, x_{P_n}\} \), to minimize a penalized likelihood function of the form

\[
- \sum_{i=1}^{n} \log f(y_i|\theta) + p_\lambda(\beta),
\]  

(3)

where \( p_\lambda(\cdot) \) is the penalty function, \( \lambda \) is a tunable scale parameter, and \( \beta = (\beta_0, \beta_1, \ldots, \beta_{P_n}) \) is the vector of regression coefficients.
Penalized Likelihood Approach

Different choices of $p_\lambda(\cdot)$ lead to different methods:

- Lasso (Tibshirani, 1996): $L_1$ norm.
- elastic net (Zou and Hastie, 2005): A linear combination of $L_1$ and $L_2$ norms.
- SCAD (Fan and Li, 2001), MCP (Zhang, 2010): Concave penalty norms.
- rLasso (Song and Liang, 2015)

A nice property shared by these penalty norms is that they are singular at zero and thus can induce sparsity of the model by increasing $\lambda$. In practice, $\lambda$ can be determined through a cross-validation procedure.
Penalized Likelihood Approach: Marginal utility-based

Sure independence screening (SIS) (Fan and Lv, 2008; Fan and Song, 2010): SIS first ranks predictors according to their marginal utility, namely, each predictor is used independently as a predictor to decide its usefulness for predicting the response, then selects a subset of predictors with the marginal utility exceeding a predefined threshold, and finally refine the model using Lasso or SCAD.

Fan and Song (2010) suggested to measure the marginal utility for GLMs using marginal regression coefficients or marginal likelihood ratios (with respect to the null model).

Iterated SIS (ISIS) is an iterative version of SIS: It considers information from those already chosen predictors in the previous SIS steps when measuring the marginal utility of remaining predictors, and has, in general, an improved performance over SIS.
Penalized Likelihood Approach: Information theory-based

Extended BIC criterion:

\[ p_\lambda(\beta) = \frac{|\beta|}{2} \log n + \gamma |\beta| \log P_n, \]  

(4)

where $|\beta|$ denotes the model size, and $\gamma > 0$ is a tunable parameter. Under certain regularity conditions, Chen and Chen (2011) showed that EBIC is consistent if $P_n = O(n^\kappa)$ and $\gamma > 1 - \frac{1}{2\kappa}$, where the consistency means the causal features can be identified with probability 1 by minimizing (3) as the sample size $n \to \infty$.

Average (E)-BIC criterion (Xue, Luo and Liang, 2017)
Bayesian Methods

- Bayesian Lasso: A Laplace-like prior is used for Bayesian variable selection (Park and Casella, 2008)
- non-local prior (Johnson and Rossell, 2012)
- Posterior Consistency: Jiang (2006, 2007) and Liang, Song and Yu (2013) showed that the true density (2) can be consistently estimated based on the models sampled from the posterior distribution through a Bayesian variable selection procedure as $n \to \infty$. This property is called posterior consistency.
The prior

Let \( \xi_n = \{x_1^*, \ldots, x_{k_n}^*\} \subset \{x_1, x_2, \ldots, x_{P_n}\} \) denote a subset model of the GLM, where \( x_i^*, i = 1, \ldots, k_n \), denote the predictors included in the subset model.

Let \( \beta_{\xi_n} \) be subject to the Gaussian prior:

\[
\pi(\beta_{\xi_n}) = \frac{1}{(2\pi\sigma_{\xi_n}^2)^{k_n/2}} \exp \left\{ -\frac{1}{2\sigma_{\xi_n}^2} \sum_{i=1}^{k_n} \beta_i^*^2 \right\},
\]

where \( \sigma_{\xi_n}^2 \) denotes a pre-specified variance and choose \( \sigma_{\xi_n}^2 \) for each model such that

\[
\log \pi(\beta_{\xi_n}) = O_p(1),
\]

i.e., the prior information of \( \beta_{\xi_n} \) can be ignored for sufficiently large \( n \).

Under condition \((A_2)\) (given latter),

\[
\sigma_{\xi_n}^2 = \frac{1}{2\pi} e^{C_0/k_n}, \quad \text{for some positive constant } C_0,
\]

ensures (6) to hold for sufficiently large \( n \).
Further, we let the model $\xi_n$ be subject to the prior

$$
\pi(\xi_n) = \nu_n^{k_n} (1 - \nu_n)^{p_n - k_n},
$$

(8)

i.e., each variable has a prior probability $\nu_n$, independent of other variables, to be selected for the subset model.
The posterior

Let the prior probability $\nu_n$ in (8) take a value in the form

$$\nu_n = \frac{1}{1 + P_n^\gamma \sqrt{2\pi}},$$

(9)

for some parameter $\gamma$, then we have the posterior,

$$\log \pi(\xi_n | D^n) \approx C + \log f(y_n | \hat{\beta}_{\xi_n}, \xi_n, X_n) - \frac{k_n}{2} \log(n) - k_n \gamma \log(P_n).$$

(10)

Maximizing this posterior probability is equivalent to minimizing the EBIC given by

$$EBIC = -2 \log f(y_n | \hat{\beta}_{\xi_n}, \xi_n, X_n) + k_n \log(n) + k_n \gamma \log(P_n).$$
To facilitate calculation of the MLE $\hat{\beta}_{\xi_n}$, in practice, one often needs to put an upper bound for the model size $k_n$, e.g., $k_n < K_n$, where $K_n$ may depend on the sample size $n$. With this bound, the prior for the model $\xi_n$ becomes

$$\pi(\xi_n) \propto \nu_n^{k_n} (1 - \nu_n)^{P_n - k_n} I[k_n < K_n],$$

and the posterior becomes

$$\log \pi(\xi_n|D^n) \approx \begin{cases} C + \log f(y_n|\hat{\beta}_{\xi_n}, \xi_n, X_n) - \frac{k_n}{2} \log(n) - k_n \gamma \log(P_n), & \text{if } k_n < K_n, \\ -\infty, & \text{otherwise.} \end{cases}$$

Since this posterior leads to sampling of subset models without shrinkage of regression coefficients, for which the Bayesian estimator of $\beta_{\xi_n}$ is reduced to the MLE of $\beta_{\xi_n}$, we call it Bayesian subset regression (BSR).
We assume the following conditions hold:

(A1) (Predictor Size) $P_n \succ n^\delta$ for some $\delta > 0$, where $b_n \succ a_n$ means $\lim_{n \to \infty} a_n/b_n = 0$.

(A2) (Sparsity) $\lim_{n \to \infty} \sum_{j=1}^{P_n} |\beta_j| < \infty$. 
**Theorem 1.** Consider the GLM specified by (1) and (2), which satisfies the conditions $(A_1)$, $(A_2)$ and $|x_j| \leq 1$ for all $j$. Let $\epsilon_n$ be a sequence such that $\epsilon_n \in (0, 1]$ for each $n$ and $n\epsilon_n^2 \succ \log(P_n)$. Suppose the priors for the GLM are specified by (5) and (11) with the hyperparameter $\nu_n$ chosen in (9), and $\gamma$ and $K_n$ are chosen such that the following conditions hold:

$(B_1)$ $P_n \leq e^{C_1 n^\alpha}$ for some $C_1 > 0$ and some $\alpha \in (0, 1)$ for all large enough $n$;

$(B_2)$ $\Delta(r_n) = \inf_{\xi_n : k_n = r_n} \sum_{j : j \notin \xi_n} |\beta_j| \leq e^{-C_2 r_n}$ for some constant $C_2 > 0$, where $r_n = \lceil P_n \nu_n \rceil$ denotes the up-rounded expectation of the prior (8);

$(B_3)$ $C_2^{-1} \log(n) \leq r_n \leq K_n \prec n^\beta$ for some $\beta \in (0, q)$, where $q = \min(1 - \alpha, \delta)$ for logistic, probit and normal linear regressions, and $q = \min\{1 - \alpha, \delta, \alpha/4\}$ for Poisson regression and exponential regressions with log-linear link functions.

Then for some $c > 0$ and for all sufficiently large $n$,

$$P \left\{ \pi[d(\hat{f}, f) > \epsilon_n | D^n] \geq e^{-0.5c n \epsilon_n^2} \right\} \leq e^{-0.5c n \epsilon_n^2},$$

(13)

where $P\{\cdot\}$ denotes the probability measure for the data $D^n$, and $d(\cdot, \cdot)$ denotes the Hellinger distance between $\hat{f}$ and $f$. 

**Posterior Consistency**
Determination of $\gamma$

Theorem 1 suggests to choose $\gamma \in (1 - 1/\kappa, 1)$ for $\kappa > 1$ and $(0,1)$ otherwise. In practice, we set

$$\gamma = \inf \{ \tilde{\gamma} : \arg \max_{\|\xi_n\|} \pi(\|\xi_n\|D^n, \tilde{\gamma}) = \|\xi_{map, \tilde{\gamma}}\| \},$$

which is to find the minimum value of $\gamma$ for which the mode of $\pi(\|\xi_n\|D^n, \gamma)$ attains at the size of the MAP model.

Both Theorem 1 and Chen and Chen (2011) suggest that the MAP model is the model that is most likely true. However, the MAP model depends on the value of $\gamma$. If $\gamma$ is overly large, then the equality $\arg \max_{\|\xi_n\|} \pi(\|\xi_n\|D^n, \tilde{\gamma}) = \|\xi_{map, \tilde{\gamma}}\|$ always holds as, in this case, there will be only a single-size model sampled from the posterior, but the resulting MAP model will be likely a subset of the true model. On the other hand, if $\gamma$ is too small, then we have always $\arg \max_{\|\xi_n\|} \pi(\|\xi_n\|D^n, \tilde{\gamma}) > \|\xi_{map, \tilde{\gamma}}\|$. To balance between the two ends, we suggest the rule (14).
Let $M_*$ denote the set of causal features for a dataset $D^n$. Let $e_n = (e_1, e_2, \ldots, e_{P_n})$ denote the indicator vector of $P_n$ features; that is, $e_j = 1$ if $x_j \in M_*$ and 0 otherwise. Let $q_j$ denote the marginal inclusion probability of $x_j$, i.e.,

$$q_j = \sum_{\xi} e_{j|\xi} \pi(\xi|D^n),$$

where $e_{j|\xi}$ indicates whether $x_j$ is included in the model $\xi$. A conventional rule is to choose the features for which the marginal inclusion probability is greater than a threshold value $\hat{q}$; i.e., setting $\hat{M}_{\hat{q}} = \{j : q_j > \hat{q}, j = 1, \ldots, P_n\}$ as an estimator of $M_*$. 
Sure Variable Screening: identifiability condition

Let $A_{\epsilon_n} = \{\xi : d(\hat{f}(y, x|\xi), f(y, x)) \leq \epsilon_n\}$, where $d(\cdot, \cdot)$ denotes the Hellinger distance between $\hat{f}$ and $f$. Define

$$
\rho_j(\epsilon_n) = \sum_{\xi \in A_{\epsilon_n}} |e_j - e_j|\xi|\pi(\xi|D^n),
$$

which measures the distance between the true model and the sampled models for feature $j$ in the $\epsilon_n$-neighborhood $A_{\epsilon_n}$.

$(C_1)$ (Identifiability of $M_\star$) For all $j = 1, 2, \ldots, P_n$, $\rho_j(\epsilon_n) \to 0$ as $n \to \infty$ and $\epsilon_n \to 0$. 
Sure Variable Screening: Model consistency

**Theorem 3.** Assume the conditions of Theorem 1 (i.e., \( A_1 \) and \( A_2 \)) and the condition \((C_1)\) hold.

(i) For any \( \delta_n > 0 \) and all sufficiently large \( n \),

\[
P \left( \max_{1 \leq j \leq P_n} |q_j - e_j| \geq 2\sqrt{\delta_n + e^{-0.5cn^2}} \right) \leq P_n e^{-0.5cn^2}.
\]

(ii) (Sure screening) For all sufficiently large \( n \),

\[
P(M_* \subset \hat{M}_{\hat{q}}) \geq 1 - s_n e^{-0.5cn^2},
\]

where \( s_n \) denotes the size of \( M_* \), for some choice of \( \hat{q} \in (0, 1) \) but preferably being away from 0 and 1.

(iii) (Consistency) For all sufficiently large \( n \),

\[
P(M_* = \hat{M}_{0.5}) \geq 1 - P_n e^{-0.5cn^2}.
\]
Lymph Example

This dataset consists of $n = 148$ samples with 100 node-negative cases (low risk for breast cancer) and 48 node-positive cases (high risk for breast cancer). After pre-screening, 4512 genes were selected for further study, which shows evidence of variation above the noise level. In addition, two clinical factors, estimated tumor size (in centimeters) and protein assay-based estrogen receptor status, were included as candidate predictors, with each coded as a binary variable, This brings the total number of predictors to $P_n = 4514$.

The subset data consists of only 34 genes, which corresponds to the set of significant genes identified by SVS at a FDR level of 0.01 based on the marginal inclusion probabilities produced by SAMC in a run on the full data with $\gamma = 0.85$. 
Table 1. Comparison of BSR, Lasso, elastic net, SIS and ISIS for the subset and full data of lymph: $r_{cv}(\%)$ refers to the minimum 1-CVMR (leave-one-out cross-validation misclassification rate), and MAP$_{\text{min}}$ refers to the minimum size MAP$_i$ model with zero 1-CVMR, where MAP$_i$ denotes the MAP model consisting of $i$ features.

<table>
<thead>
<tr>
<th>Data</th>
<th>Lasso</th>
<th>elastic net</th>
<th>SIS</th>
<th>ISIS</th>
<th>BSR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>\xi</td>
<td>r_{cv}(%)$</td>
<td>$</td>
<td>\xi</td>
</tr>
<tr>
<td>Sub</td>
<td>28 2.03</td>
<td>31 2.70</td>
<td>7 16.9</td>
<td>7 7.43</td>
<td>8 2.70</td>
</tr>
<tr>
<td>Full</td>
<td>23 15.5</td>
<td>51 16.9</td>
<td>7 17.6</td>
<td>7 10.8</td>
<td>8 0.68</td>
</tr>
</tbody>
</table>
Lymph Example

Figure 1. Comparison of Lasso and BSR for the subset data (a) and the full data (b) of lymph: 1-CVMRs of the models selected by Lasso (while circles connected by solid line) and the MAP$_1$–MAP$_{16}$ models selected by BSR (black squares connected by dotted line).
Figure 2. Plots of $q$-values for the genes selected by BSR, SIS, ISIS, Lasso and elastic net. The model shown in (a) is one of the MAP models found by BSR with $\gamma = 0.85$ and it has a 1-CVMR of 2.7%. The models shown in (b), (c), (d) and (e) are selected by SIS, ISIS, Lasso and elastic net, and have 1-CVMRs 17.6%, 10.8%, 15.5% and 16.9%, respectively.
A Simulated Example

This example mimics a case-control genetic association study (GAS). The response variable $y$, which represents the disease status of a subject, takes value 1 for the case of disease and 0 for the control. The explanatory variables are generated as single nucleotide polymorphisms (SNPs) in the human genome. Each variable $x_{ij}$, the genotype of SNP $j$ of subject $i$, takes values 0, 1 or 2, where 0 stands for a homozygous site with major allele, 1 stands for a heterozygous site with minor allele, and 2 stands for a homozygous site with two copies of minor allele.

We independently generated 10 datasets with $n_1 = n_2 = 500$, $P_n = 10000$, $k = 8$, and $(\beta_1, \ldots, \beta_8) = (0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3)$. 
A Simulated Example (continued)

Table 2. Comparison of BSR with Lasso, elastic net, SIS and ISIS for the simulated data: \(^a\) results of SVS at a FDR level of 0.001; \(^b\) Average model size (over 10 datasets) produced by different methods with the standard deviation presented in the parentheses.

<table>
<thead>
<tr>
<th>Methods</th>
<th>BSR((\gamma = 0.85))</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>SVS(^a)</td>
<td>Lasso</td>
<td>Elastic net</td>
<td>SIS</td>
<td>ISIS</td>
</tr>
<tr>
<td>Size(^b)</td>
<td>8.2(0.25)</td>
<td>12.1(0.43)</td>
<td>44.9(20.48)</td>
<td>30.1(9.04)</td>
<td>36(0)</td>
<td>36(0)</td>
</tr>
<tr>
<td>fsr(%)</td>
<td>3.66</td>
<td>25.6</td>
<td>81.8</td>
<td>73.4</td>
<td>77.8</td>
<td>77.8</td>
</tr>
<tr>
<td>nsr(%)</td>
<td>12.22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3. Summary of four gene expression datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Publication</th>
<th>n</th>
<th>P</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lymph</td>
<td>Hans et al. (2007)</td>
<td>148</td>
<td>4514</td>
<td>positive/negative node</td>
</tr>
<tr>
<td>Colon</td>
<td>Alon et al. (1999)</td>
<td>62</td>
<td>2000</td>
<td>Tumor/normal tissue</td>
</tr>
<tr>
<td>Leukemia</td>
<td>Golub et al. (1999)</td>
<td>72</td>
<td>3571</td>
<td>Subtype of leukemia</td>
</tr>
<tr>
<td>Prostate</td>
<td>Singh et al. (2002)</td>
<td>102</td>
<td>6033</td>
<td>Tumor/normal tissue</td>
</tr>
</tbody>
</table>
Table 4. Posterior distribution $\pi(\|\xi\|D^n, \gamma = 0.99)$ obtained by BSR for colon, leukemia and prostate data.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>$\geq$ 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>0.001</td>
<td>0.387</td>
<td>0.201</td>
<td>0.260</td>
<td>0.118</td>
<td>0.029</td>
<td>0.004</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.023</td>
<td>0.561</td>
<td>0.337</td>
<td>0.070</td>
<td>0.007</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Prostate</td>
<td>0</td>
<td>0.005</td>
<td>0.098</td>
<td>0.681</td>
<td>0.191</td>
<td>0.024</td>
<td>0.002</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
<td>$&lt; 0.001$</td>
</tr>
</tbody>
</table>
Table 5. Comparison of BSR, Lasso, elastic net, SIS and ISIS for three gene expression data. $^a$ minimum 1-CVMR ($r_{cv}\%$) and the corresponding model size ($|\xi|$); $^b$ 1-CVMR ($r_{cv}\%$) and the corresponding model size ($|\xi|$); $^c$ zero 1-CVMR models among MAP$_i$’s sampled by SAMC.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lasso</th>
<th>elastic net</th>
<th>SIS</th>
<th>ISIS</th>
<th>BSR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$</td>
<td>\xi</td>
<td>$</td>
<td>$r_{cv}%$</td>
<td>$</td>
</tr>
<tr>
<td>Colon</td>
<td>15</td>
<td>12.9</td>
<td>26</td>
<td>11.3</td>
<td>3</td>
</tr>
<tr>
<td>Leukemia</td>
<td>16</td>
<td>4.17</td>
<td>37</td>
<td>2.78</td>
<td>4</td>
</tr>
<tr>
<td>Prostate</td>
<td>3</td>
<td>6.86</td>
<td>42</td>
<td>5.88</td>
<td>5</td>
</tr>
</tbody>
</table>
Figure 3. Comparison of 1-CVMRs produced by Lasso (while circles connected by solid line) and BSR (black squares connected by dotted line) for (a) colon data, (b) leukemia data, and (c) prostate data.
Table 6. Summary of CPU times of BSR: The CPU times are measured (in hours) on a Dell desktop of 3.0Ghz and 8GB RAM, and $E_{D^n}(|\xi|)$ denotes the posterior expectation of $|\xi|$.

| Dataset    | $n$ | $P_n$ | $\gamma$ | Iterations ($\times 10^6$) | $E_{D^n}(|\xi|)$ | CPU(h) |
|------------|-----|-------|----------|----------------------------|------------------|--------|
| Biliary stone | 715 | 1748  | 0.75     | 5.05                       | 1.9              | 2.0    |
| Lymph      | 148 | 4514  | 0.85     | 10.05                      | 8.5              | 12.7   |
| Colon      | 62  | 2000  | 0.99     | 5.05                       | 3.2              | 1.0    |
| Leukemia   | 72  | 3571  | 0.99     | 5.05                       | 2.5              | 1.7    |
| Prostate   | 102 | 6033  | 0.99     | 5.05                       | 4.1              | 1.5    |
After adjusting for the number of iterations, a linear regression analysis for the CPU time versus $n$, $P_n$ and $[E_{D^n}(|\xi|)]^3$ indicates that the CPU time mainly depends on $[E_{D^n}(|\xi|)]^3$, then $n$ and $P_n$ in the order of significance. Actually, a simple linear regression for the CPU time versus $[E_{D^n}(|\xi|)]^3$ has a $R^2$ of 95.1%. A cubic transformation for $E_{D^n}(|\xi|)$ is used here, because the CPU time for inverting a matrix is known in the cubic order of the matrix size.

This analysis shows that BSR (Bayesian subset regression) can be applied to high dimensional regression problems with reasonable CPU time.
Summary

- We propose a new prior setting for GLMs, which leads to a Bayesian subset regression with the MAP model coinciding with the minimum EBIC model. Under mild conditions, we establish consistency of the posterior.

- We propose a variable screening procedure based on the marginal inclusion probability and show that the procedure shares the same theoretical properties of sure screening and consistency with the SIS procedure proposed by Fan and Song (2010). Since the proposed procedure makes use of the joint information of all predictors, it generally outperforms SIS and its iterative extension ISIS.

- We make extensive comparisons of BSR with the popular penalized likelihood methods, including Lasso, elastic net, SIS and ISIS. Through a comparison study conducted on the subset and full data, we find that the performance of the penalized likelihood methods tend to deteriorate with the dimension $P_n$. Given the stable performance of BSR on the subset and full data, we conclude that BSR is more suitable than these penalized likelihood methods for high dimensional variable selection, although the latter are computationally more attractive. Our further results on the simulated data, real SNP data and four gene expression data make this conclusion more convincing.