On the Computational Complexity of High-Dimensional Bayesian Variable Selection

Yun Yang, Martin J. Wainwright and Michael I. Jordan

Group meeting· Jiapeng Liu· Dec 1, 2015

Outline

Introduction

Model

Main Results

Counter Example

Simulations

Introduction

Consider the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^{*} + \boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n) \tag{1}$$

The response vector $\boldsymbol{Y} \in \mathbb{R}^n$, a design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\beta^* \in \mathbb{R}^p$ is the unknown regression vector and $p \gg n$.

- Our goal:
  To recover the support set of β^*, or more generally, a subset of covariates with absolute regression weights above some threshold.
- One can take a Bayesian point of view on high-dimensional regression.

Why Bayesian Variable Selection

- A subset of possible models instead of a single model.
- The marginal probability of including each covariate.

How to do Bayesian Variable Selection

- First, imposing a prior over the binary indicator vector $\gamma \in \{0, 1\}^p$.
- Then using posterior probability $\pi(\gamma|\boldsymbol{Y})$ to perform variable selection.

- By the Bayes' Theorem,

$$\pi(\gamma|Y) = \frac{P(Y|\gamma)\pi(\gamma)}{P(Y)}$$

Usually, $P(Y)$ is difficult to calculate. Therefore, the most widely used tool for fitting Bayesian models are sampling techniques based on Markov chain Monte Carlo (MCMC).

One Central Problem

$Mixing\ time$ of the Markov Chain:

- $Mixing\ time$ – the number of iterations required to converge.
- $Rapid\ Mixing$ – mixing time grows at most polynomially in the problem parameters.
- $Slowly\ Mixing$ – mixing time grows exponentially in the problem parameters.

Contributions of This Paper

- Showing that a Bayesian approach can achieve $variable - selection\ consistency$ under relatively mild conditions on the design matrix.
- Providing a set of conditions that guarantee BOTH $variable - selection\ consistency$ and $rapid\ mixing$ of a particular Metropolis-Hastings algorithm.
- Providing a counter-example to illustrate that variable-selection consistency DOES NOT imply rapid mixing.

Metropolis-Hastings random walk

- Gibbs samplers
- Metropolis-Hastings random walks
  - Step1: Current state γ;
  - Step2: Choose γ' with probability $S(\gamma, \gamma')$ in $\mathcal{N}(\gamma)$;
  - Step3: Move to γ' with probability $R(\gamma, \gamma')$, stay in γ with probability $1 - R(\gamma, \gamma')$;

$$R(\gamma, \gamma') := \min\{1, \frac{\pi_n(\gamma' | \boldsymbol{Y}) S(\gamma', \gamma)}{\pi_n(\gamma | \boldsymbol{Y}) S(\gamma, \gamma')}\}$$

Metropolis-Hastings random walk

- Hamming distance: $d_H(\gamma, \gamma') = \Sigma_{j=1}^{p} \mathbb{I}(\gamma_j \neq \gamma'_j)$
- $\mathcal{N}_1(\gamma) := \{\gamma' \mid d_H(\gamma, \gamma') = 1\}$
  $\mathcal{N}_2(\gamma) := \{\gamma' \mid d_H(\gamma, \gamma') = 2$ and
  $\exists (k, l) \in S(\gamma) \times S^c(\gamma)$ s.t. $\gamma'_k = 1 - \gamma_k$ and $\gamma'_l = 1 - \gamma_l\}$
  Here $S(\gamma) = \{j \in [p] \mid \gamma_j = 1\}$
- $P(\mathcal{N}_1) = P(\mathcal{N}_2) = 0.5,\ S(\gamma, \cdot)$ are uniform distribution:
- The transition matrix:

$$\boldsymbol{P}_{MH}(\gamma, \gamma') = \begin{cases} \frac{1}{2p} \min\{1, \frac{\pi_n(\gamma'|\boldsymbol{Y})}{\pi_n(\gamma|\boldsymbol{Y})}\} & \text{if } \gamma' \in \mathcal{N}_1 \\ \frac{1}{2|S(\gamma)||S^c(\gamma)|} \min\{1, \frac{\pi_n(\gamma'|\boldsymbol{Y})}{\pi_n(\gamma|\boldsymbol{Y})}\} & \text{if } \gamma' \in \mathcal{N}_2 \\ 0 & \text{if } d_H(\gamma, \gamma') > 2 \\ 1 - \Sigma_{\gamma' \neq \gamma} \boldsymbol{P}_{MH}(\gamma, \gamma') & \text{if } \gamma' = \gamma \end{cases}$$

Model Description

Bayesian hierarchical model $\mathbb{M}_\gamma$:

$$\text{Linear Model:} \quad \boldsymbol{Y} = \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim \mathcal{N}(0, \phi^{-1}\boldsymbol{I}_n) \tag{2a}$$

$$\text{Precision Prior:} \quad \pi(\phi) \propto \frac{1}{\phi} \tag{2b}$$

$$\text{Regression prior:} \quad (\boldsymbol{\beta}_\gamma|\gamma) \sim \mathcal{N}(0, g\phi^{-1}(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}) \tag{2c}$$

$$\text{Sparsity prior:} \quad \pi(\gamma) \propto \left(\frac{1}{p}\right)^{\kappa|\gamma|} \mathbb{I}[|\gamma| \leq s_0] \tag{2d}$$

- Integer $s_0 \leq p$
- The regression prior (2c) is Zellner's g-prior, which leads to an especially simple form of the marginal likelihood function.

Model Description

- Suppose $\boldsymbol{Y} \in \mathbb{R}^n$ is generated from $\boldsymbol{Y} = \boldsymbol{X}\beta^* + \omega$, where $\omega \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, β^* and σ_0^2 are unknown.
- $S := \{j \in [p] \mid |\beta_j^*| \geq C_\beta(\sigma_0, n, p)\}$ is the subset we want to recover.
- γ^* is the indicator vector that selects the influential covariates, let $s^* := |\gamma^*|$.

Several assumptions

Assumption A (Conditions on β^*)

The true regression vector has components $\beta^* = (\beta^*_S, \beta^*_{S^c})$ that satisfy the bounds

$$\begin{aligned} \text{Full } \beta^* \text{ condition:} \quad & \left\| \frac{1}{\sqrt{n}} \boldsymbol{X} \beta^* \right\|_2^2 \leq g \sigma_0^2 \frac{\log p}{n} \\ \text{Off-support } S^c \text{ condition:} \quad & \left\| \frac{1}{\sqrt{n}} \boldsymbol{X}_{S^c} \beta^*_{S^c} \right\|_2^2 \leq \widetilde{L} \sigma_0^2 \frac{\log p}{n} \end{aligned} \qquad (3)$$

for some universal constant $\widetilde{L}$

- In the simplest case, $\beta^*_{S^c} = 0$, so that the off-support condition holds trivially.

Several assumptions

Assumption B (Conditions on the design matrix)

The design matrix has been normalized so that $\|X_j\|_2^2 = n$ for all $j = 1, ..., p$; moreover, letting $Z \sim \mathcal{N}(0, \boldsymbol{I}_n)$, there exist constants $\nu > 0$ and $L < \infty$ such that

Lower restricted eigenvalue (RE(s)): $\min_{|\gamma| \leq s} \lambda_{min} \left(\frac{1}{n} \boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma \right) \geq \nu$,

Sparse projection condition (SI(s)):

$$\mathbb{E}_Z \left[\max_{|\gamma| \leq s} \max_{k \in [p] \setminus \gamma} \frac{1}{n} |\langle (I - \Phi_\gamma) \boldsymbol{X}_k, Z \rangle| \right] \leq \frac{1}{2} \sqrt{L \nu \text{log} p} \tag{4}$$

where Φ_γ denotes projection onto the span of $\{X_j, j \in \gamma\}$.

- We set integer $s = Ks^*$ or $s = s_0$.
- The sparse projection condition can always be satisfied by choosing $L = O(s_0)$.

Several assumptions

Assumption C (Choices of prior hyperparameters))

The noise hyperparameter g and sparsity penalty hyperparameter $\kappa > 0$ are chosen such that

$$\begin{aligned} & g \asymp p^{2\alpha} && \text{for some } \ \alpha > 0 \\ & \kappa + \alpha \geq C_1(L + \widetilde{L}) + 2 && \text{for some universal constant } \ C_1 > 0 \end{aligned} \tag{5}$$

Several assumptions

Assumption D (Sparsity control)

For a constant $C_0 > 4$, one of the two following conditions holds:
Version D(s^*): We set $s_0 := p$ in the sparsity prior (4), and the true sparsity s^* is bounded as

$$s^* \leq \frac{1}{8C_0K}\left\{\frac{n}{\mathsf{log}p} - 16\widetilde{L}\sigma_0^2\right\} \quad \text{for some constant } K \geq 4 + \alpha + c\widetilde{L}.$$

Version D(s_0): The sparsity parameter s_0 in the prior (4) satisfies the sandwich relation

$$(2\nu^{-2}\omega(X) + 1)s^* \ \leq \ s^* \leq \frac{1}{8C_0K}\left\{\frac{n}{\mathsf{log}p} - 16\widetilde{L}\sigma_0^2\right\} \tag{6}$$

Where $\omega(X) := \max_{\gamma \in \mathcal{M}} \|(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1} \boldsymbol{X}_\gamma^T \boldsymbol{X}_{\gamma^* \setminus \gamma}\|_{op}^2$

Main Results

Suffcient conditions for posterior consistency

Theorem 1 (Posterior concentration)

Suppose that Assumption A, B (with $s = Ks^* \ll s_0$), C and D(s^*)hold.
If the threshold C_β satisfies

$$C_\beta^2 \geq c_0(L + \widetilde{L} + \alpha + \kappa)\sigma_0^2 \frac{\log p}{n}, \tag{7}$$

then we have $\pi_n(\gamma^*|Y) \geq 1 - c_1 p^{-1}$ with probability at least $1 - c_2 p^{-c_3}$.

Main Results

Remark

- It is worth noting that the result of Theorem 1 covers two regimes:

$$\text{High SNR: } S = \{j \in [p] \mid \beta_j^* \neq 0\}, \min_{j \in S} |\beta_j^*|^2 \geq c_0(L + \alpha + \kappa)\sigma_0^2 \frac{\log p}{n} \tag{8a}$$

$$\text{Low SNR: } S = \emptyset \text{ and } \left\| \frac{1}{\sqrt{n}} \boldsymbol{X} \beta^* \right\|_2^2 \leq \left(\frac{\alpha + \kappa - 2}{C_1} - L \right) \sigma_0^2 \frac{\log p}{n} \tag{8b}$$

- The high SNR regime corresponds to $\widetilde{L} = 0$,
- The low SNR regime corresponds to $\widetilde{L} = \frac{\alpha+\kappa-2}{C_1} - L$.

Main Results

Corollary 1

Under the conditions of Theorem 1, with probability at least $1 - c_2 p^{-c_3}$:
(a) Under the high SNR condition (8a), we have $\pi_n(\gamma^*|Y) \geq 1 - c_1 p^{-1}$.
(b) Under the low SNR condition (8b), we have $\pi_n(\gamma_0|Y) \geq 1 - c_1 p^{-1}$.
Where γ_0 is the indicator vector of null model.

Main Results

Remark

- This does not cover the intermediate regime:

$$\left(\frac{\alpha+\kappa-2}{C_1}-L\right)\sigma_0^2\frac{\log p}{n}\leq|\beta_j^*|^2\leq c_0(L+\alpha+\kappa)\sigma_0^2\frac{\log p}{n} \quad (9)$$

- Theorem 1 still guarantees Bayesian variable selection consistency in this regime. However, the MCMC algorithm for sampling from the posterior can exhibit slow mixing.

Main Results

Suffcient conditions for rapid mixing

- For $\gamma \in \mathcal{M}$ and any subset $S \subseteq \mathcal{M}$, let $P(\gamma, S) = \Sigma_{\gamma' \in S} P(\gamma, \gamma')$.
- If γ is the initial state of the chain, then the total variation distance to the stationary distribution after t iterations is

$$\triangle_{\gamma}(t) = \|P^t(\gamma, \cdot) - \pi(\cdot)\|_{TV} := \max_{S \subset \mathcal{M}} |P^t(\gamma, S) - \pi(S)|.$$

Defination A (mixing time)

The $\epsilon - mixing$ time is given by

$$\tau_{\epsilon} := \max_{\gamma \in \mathcal{M}} \min\{t \in \mathbb{N} \mid \ \triangle_{\gamma}(t') \leq \epsilon \text{ for all } t' \geq t\} \qquad (10)$$

Main Results

Suffcient conditions for rapid mixing

Theorem 2 (Rapid mixing guarantee)

Suppose that $Assumption\ A,\ B\ (with\ s = s_0),\ C\ and\ D(s_0)$ hold. Then under either the $high\ SNR\ (8a)$ or $the\ low\ SNR\ (8b)$, there are universal constants c_1, c_2 s.t., for any $\epsilon \in (0,1)$, the $\epsilon - mixing$ time of the Metropolis-Hastings chain is upper bounded as

$$\tau_\epsilon \leq c_1 p s_0^2 (c_2 \alpha (n + s_0) \text{log} p + \text{log}(1/\epsilon) + 2) \tag{11}$$

with probability at least $1 - 4p^{-c_1}$.

- Theorem 2 does not characterize the intermediate regime in (9).
- Based on our simulations, we suspect that the Markov chain might be slowly mixing in this regime, but we do not have a proof of this statement

Main Results

Corollary 2

Under the conditions of Theorem 2, for any fixed iterate t such that:

$$t \geq c_1 p s_0^2 (c_2 \alpha (n + s_0) \text{log} p + \text{log} p + 2)$$

the iterate t from the MCMC algorithm matches γ^* with probability at least $1 - c_2 p^{-c_3}$.

Counter Example

- Suppose $p = n$, $g = p^{2\alpha}$ with $\alpha \geq 1$, let $\boldsymbol{Y} = \omega \sim \mathcal{N}(0, \boldsymbol{I}_n)$,
- Prior: untruncated distribution $\pi_n(\gamma) = Cp^{-\kappa|\gamma|}$,
- We can prove the mixing time of the Markov chain with transition probability in page 9 grows exponentially in n with probability at least 1/2 with respect to the randomness of ω.

Counter Example

Remark

- This example satisfies the conditions in Theorem 1, which imply Bayesian variable-selection consistency.
- The size constraint $|\gamma| < s_0$ is necessary for MCMC to mix rapidly.

Simulations

- Independent design:
  $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\beta^*, \sigma^2 \boldsymbol{I}_n)$ with $x_i \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_p)$ i.i.d.;
- $s^* = 10,\ s_0 = 100;$
- $\beta^* = SNR\sqrt{\frac{\sigma^2 \text{log} p}{n}}(2, -3, 2, 2, -3, 3, -2, 3, -2, 3, 0, ..., 0)^T \in \mathbb{R}^p$.

Simulations

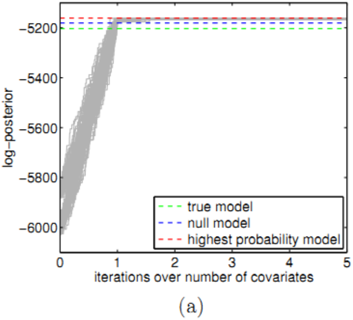

Figure 1. Log-posterior probability versus the number of iterations (divided by the number of covariates p) of 100 randomly initialized Markov chains with $n = 500$, $p = 1000$ and SNR $\in \{1, 3\}$ in the independent design. In all cases, each grey curve corresponds to one trajectory of the chain (100 chains in total). Half of the chains are initialized at perturbations of the null model and half the true model. (a) Weak signal case: SNR $= 1$. (b) Strong signal case: SNR $= 3$ (the posterior probability of the true model coincides with that of the highest probability model).

Future Direction

- It is interesting to investigate the mixing behavior of the MCMC algorithm when Bayesian variable selection fails;
- Another interesting direction is to consider the computational complexity of MCMC methods for models more complex than linear regression, for example, high-dimensional nonparametric additive regression;
- A third direction is to investigate whether the upper bound on mixing time provided in Theorem 2 is sharp up to constants.