The Korean Journal of Applied Statistics
2022, Vol. 35, No. 3, 445–455

DOI: https://doi.org/10.5351/KJAS.2022.35.3.445
Print ISSN1225-066X / Online ISSN 2383-5818

Introduction to variational Bayes for high-dimensional linear and logistic regression models

Insong Jang[a], Kyoungjae Lee[1,b]

[a]Department of Statistics, Inha University; [b]Department of Statistics, Sungkyunkwan University

Abstract

In this paper, we introduce existing Bayesian methods for high-dimensional sparse regression models and compare their performance in various simulation scenarios. Especially, we focus on the variational Bayes approach proposed by Ray and Szabó (2021), which enables scalable and accurate Bayesian inference. Based on simulated data sets from sparse high-dimensional linear regression models, we compare the variational Bayes approach with other Bayesian and frequentist methods. To check the practical performance of the variational Bayes in logistic regression models, a real data analysis is conducted using leukemia data set.

Keywords: variable selection, regression model, spike and slab prior, horseshoe prior

1. 서론

다음의 선형 회귀모형

$$Y = X\beta + \varsigma Z, \tag{1.1}$$

또는 로지스틱 회귀모형

$$P(Y_i = 1 \mid X_i) = \frac{e^{X_i^T\beta}}{1 + e^{X_i^T\beta}}, \quad i = 1, \ldots, n \tag{1.2}$$

을 고려하자. 위 식 (1.1)에서 $Y = (Y_1, Y_2, \ldots, Y_n)^T \in \mathbb{R}^n$는 종속변수, $X = (X_1, X_2, \ldots, X_n)^T \in \mathbb{R}^{n\times p}$는 설계 행렬 (design matrix), $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^p$는 회귀계수 벡터, ςZ는 오차항, $Z \sim N_n(0, I_n)$ 을 의미한다. 다음으로 식 (1.2)에서 $Y_i \in \{0, 1\}$는 이항 종속변수, $X_i \in \mathbb{R}^p$는 독립변수 벡터, $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^p$는 회귀계수 벡터이다. 본 논문에서는 자료의 개수 n보다 변수의 개수 p가 큰 고차원 상황에서, 대부분의 회귀계수 β_i가 0인 희박성(sparsity)을 가정하는 고차원 희박 회귀모형을 고려한다. 이때, 특히 추론의 불확실성을 고려하는 베이지안 방식의 회귀계수 추정과 변수선택에 관심이 있다.

고차원 상황에서 사용하는 대표적인 사전분포는 크게 두 가지로 나눌 수 있다. 첫 번째는 spike and slab 사전분포로, 신호와 잡음을 나타내는 두 가지 분포의 혼합 형태로 구성되어 있다 (George와 McCulloch, 1993; Ishwaran과 Rao, 2005). 이는 신호와 잡음에 다른 분포를 부여함으로써 변수선택을 자연스럽게 만드는 장점이 있는 반면, 총 2^p개의 가능한 모든 모형 공간을 탐험해야 하므로 고차원 상황에서는 계산적인 문제가

This work was supported the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A4A1018207).

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: leekjstat@gmail.com

Published 30 June 2022 / journal homepage: http://kjas.or.kr

발생하게 된다. 두 번째는 horseshoe 사전분포로 대표되는 연속 수축 사전분포(continuous shrinkage prior)이다 (Carvalho 등, 2010). 연속 수축 사전분포들은 각 회귀계수에 연속 사전분포를 고려함으로써, spike and slab 사전분포에 비해 모형 공간을 획기적으로 줄이는 효과가 있다. 하지만 회귀계수가 정확하게 0의 값을 가질 확률을 0으로 만들기 때문에, 변수선택 자체가 목적일 때에는 사후분포 표본들을 이용한 추가적인 작업이 필요하다.

위의 두 가지 사전분포를 이용한 사후분포 추론은, 보통 사후분포로부터 얻은 표본들에 기반해서 이루어진다. 하지만 대용량 자료를 분석할 때에는 하나의 사후분포 표본을 얻는 데에 오랜 시간이 걸릴 수 있고 다수의 표본들이 필요하기 때문에, 다양한 계산적인 문제가 발생하게 된다. 이러한 문제를 해결하기 위한 하나의 대안으로, 변분 베이즈 방법(variational Bayes method)이 사용되고 있다. 변분 베이즈 방법은 다루기 쉬운 분포를 사용하여 사후분포를 근사하는 방법으로, 사후분포와의 쿨백-라이블러 발산(Kullback–Leibler divergence)을 최소화하는 분포를 찾아서 근사 분포로 사용한다. 즉, 사후분포 추론 문제를 최적화 문제로 바꾸기 때문에 기존 베이지안 방법들에 비해 속도가 매우 빠르며, 기존 베이지안 방법들을 적용하기 힘든 대용량 자료를 추론할 때에도 사용이 가능하다는 장점이 있다.

최근 Ray와 Szabó (2021)와 Ray 등 (2020)은 고차원 선형 및 로지스틱 회귀분석에서 적용 가능한 변분 베이즈 방법을 제안하였다. 이때, Gaussian slab을 사용하는 경우 추정된 계수가 과도하게 수축되어 성능 저하를 유발할 수 있기 때문에, 이 연구들에서는 Laplace slab을 사용한 spike and slab 사전분포를 기반으로 변분 베이즈 방법으로 사후분포를 근사하였다. 본 논문에서는 모의실험과 실제 자료 분석을 통해 Ray와 Szabó (2021)가 제안한 변분 베이즈 방법과 다른 변수선택 방법들의 추론 성능 및 계산 속도를 비교하는 연구를 수행하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 고차원 상황에서 적용 가능한 기존의 대표적인 베이지안 변수선택 방법론과 변분 베이즈 방법을 소개한다. 3장에서는 다양한 실험 세팅에서 시뮬레이션 자료를 생성하여 고차원 선형 회귀분석을 수행하고 변분 베이즈 방법과 다른 변수선택 방법론 간의 성능을 비교한다. 다음으로 4장에서는 실제 자료를 사용하여 고차원 로지스틱 회귀분석을 수행하고 변분 베이즈 방법과 다른 베이지안 방법론 간의 성능을 비교한다. 마지막으로 5장에서는 자료 분석 결과를 정리하여 기술하고 논문을 마무리한다.

2. 고차원 회귀모형에 대한 사전분포 및 변분 베이즈 방법

2.1. Spike and slab 사전분포

Ishwaran과 Rao (2005)에 의해 처음 제안된 spike and slab 사전분포는, 분산이 작은 spike 부분과 큰 slab 부분의 혼합 모형으로 이루어진다. 이때, spike 부분에 해당하는 경우 잡음으로 간주하여 회귀계수를 0으로 만들고, slab 부분에 해당하는 경우 유의한 신호로 간주하여 0이 아닌 값을 추정하는 방식으로 변수선택 문제에 광범위하게 사용되고 있다. Spike and slab 사전분포는 spike 부분의 분포 타입에 따라, 크게 (1) 연속 spike and slab 사전분포 또는 (2) 불연속 spike and slab 사전분포로 구분할 수 있다.

먼저, 회귀계수에 대한 연속 spike and slab 사전분포는 다음과 같이 정의된다:

$$\begin{aligned} \beta_j \mid \gamma_j &\overset{\text{ind}}{\sim} (1-\gamma_j)N(0,\nu_0) + \gamma_j N(0,\nu_1), \\ \gamma_j &\overset{\text{iid}}{\sim} \text{Ber}(\theta), \quad j = 1,\ldots,p. \end{aligned} \tag{2.1}$$

여기서 γ_j는 j번째 회귀계수가 유의한 신호인 경우 1, 그렇지 않은 경우 0의 값을 갖는 지시변수이다. 연속 spike and slab 사전분포라는 이름은, 위 (2.1) 사전분포의 slab과 spike가 모두 연속형 분포로 이루어져 있음에 기인한다. 일반적으로 ν_0 값은 매우 작은 값으로, ν_1 값은 큰 값으로 설정하며, 정규분포 대신 t-분포 또는 Laplace 분포 등을 사용하기도 한다. Spike and slab 사전분포 (2.1)과 모형 (1.1)로부터 유도되는 γ와 β의 완전

조건부 사후분포는 다음과 같다:

$$\begin{aligned}\beta \mid \gamma, Y &\sim N_p\left(\left(\varsigma^{-2}X^TX + D_\gamma^{-1}\right)^{-1}X^TY\varsigma^{-2},\ \left(\varsigma^{-2}X^TX + D_\gamma^{-1}\right)^{-1}\right),\\ \gamma_j \mid \beta, Y &\sim \text{Ber}\left(\frac{a_j}{a_j+b_j}\right), \quad j=1,\ldots,p.\end{aligned} \tag{2.2}$$

여기서 $a_j = f(\beta_j|\gamma_j=1)\theta$, $b_j = f(\beta_j|\gamma_j=0)(1-\theta)$, $D_\gamma = \text{diag}((1-\gamma_1)\nu_0+\gamma_1\nu_1,\ldots,(1-\gamma_p)\nu_0+\gamma_p\nu_1)$이고, $f(\beta_j|\gamma_j=1)$와 $f(\beta_j|\gamma_j=0)$는 각각 $N(0,\nu_1)$과 $N(0,\nu_0)$ 확률밀도함수의 β_j에서의 값이다. 이후 3장의 모의실험에서 spike and slab 사전분포에 대한 Markov chain Monte Carlo (MCMC) 추론을 진행할 때, 위와 같은 연속형 spike and slab 사전분포 (2.1)를 이용하였다. 또한 변수선택을 위해, MCMC 표본으로부터 계산된 포함 확률(inclusion probability)이 분계점(threshold)보다 높은 경우 변수가 유의하다고 판단하였으며, 분계점으로 0.5를 사용하였다.

한편, 회귀계수에 대한 불연속 spike and slab 사전분포는 다음과 같이 정의된다:

$$\begin{aligned}\beta_j \mid \gamma_j &\overset{\text{ind}}{\sim} (1-\gamma_j)\delta_0(\beta_j) + \gamma_j N\left(0, \varsigma^2\nu_1\right),\\ \gamma_j &\overset{\text{iid}}{\sim} \text{Ber}(\theta), \quad j=1,\ldots,p.\end{aligned} \tag{2.3}$$

여기서 $\delta_0(\cdot)$는 0에서 확률 1을 갖는 point mass를 의미한다. 불연속 spike and slab 사전분포라는 이름은, spike 부분의 분포가 불연속 분포임에 기인한다. 불연속 spike and slab 사전분포 (2.3)와 모형 (1.1)로부터 유도되는 사후분포의 추론은 다음과 같은 분포들을 이용하게 된다:

$$\begin{aligned}\beta_\gamma \mid \gamma, Y &\sim N_{|\gamma|}\left(\left(X_\gamma^TX_\gamma + \nu_1^{-1}I_{|\gamma|}\right)^{-1}X_\gamma^TY,\ \left(X_\gamma^TX_\gamma + \nu_1^{-1}I_{|\gamma|}\right)^{-1}\right),\\ \gamma_j \mid \gamma_{-j}, Y &\sim \text{Ber}\left(\frac{q_j}{1+q_j}\right), \quad j=1,\ldots,p.\end{aligned} \tag{2.4}$$

여기서 $\gamma_{-j} = (\gamma_1,\ldots,\gamma_{j-1},\gamma_j=0,\gamma_{j+1},\ldots,\gamma_p)$, $\gamma_{+j} = (\gamma_1,\ldots,\gamma_{j-1},\gamma_j=1,\gamma_{j+1},\ldots,\gamma_p)$, $|\gamma| = \sum_{j=1}^p\gamma_j$이고, $X_\gamma \in \mathbb{R}^{n\times|\gamma|}$는 γ의 성분이 1인 열만 추출한 X의 부분 행렬이며, q_j는 다음과 같은 정의된다:

$$\begin{aligned}q_j = \frac{\theta}{1-\theta}&\left\{\frac{\det\left(\nu_1X_{\gamma_{-j}}^TX_{\gamma_{-j}} + I_{|\gamma_{-j}|}\right)}{\det\left(\nu_1X_{\gamma_{+j}}^TX_{\gamma_{+j}} + I_{|\gamma_{+j}|}\right)}\right\}^{\frac{1}{2}}\\ &\times\ \exp\left[-\frac{1}{2\varsigma^2}\left\{Y^TX_{\gamma_{+j}}\left(X_{\gamma_{+j}}^TX_{\gamma_{+j}} + \nu_1^{-1}I_{|\gamma_{+j}|}\right)X_{\gamma_{+j}}^TY - Y^TX_{\gamma_{-j}}\left(X_{\gamma_{-j}}^TX_{\gamma_{-j}} + \nu_1^{-1}I_{|\gamma_{-j}|}\right)X_{\gamma_{-j}}^TY\right\}\right].\end{aligned}$$

이때 식 (2.4)는 (2.2)와 비교했을 때, γ_j의 조건부 사후분포에서 β가 제외되었다는 점에서 차이가 있다. 불연속 spike and slab 사전분포 기반의 추론에서는 γ에 따라 β의 차원이 달라지기 때문에, Markov chain의 detailed balance 조건을 만족시키기 위해 해당 과정이 반드시 필요하다.

2.2. Horseshoe 사전분포

앞서 소개한 spike and slab의 경우 모수의 차원이 커질수록 탐험해야 하는 모형 공간이 급격하게 증가하여 계산을 수행하는데 긴 시간이 걸리고, MCMC 표본이 잘 mixing 되지 않는 문제가 발생한다. 이를 해결하기 위해 계산적으로 이점이 있는 수축 사전분포(shrinkage prior)가 고차원 상황에서 주목받게 되었고, 그 중 대표적인 수축 사전분포는 Carvalho 등 (2010)이 제안한 horseshoe 사전분포이다. 본 연구에서는 β_j에 대하여 다음과

같은 horseshoe 사전분포를 고려한다:

$$\begin{aligned} \beta_j \mid \lambda_j, \tau, \varsigma^2 &\sim N\left(0, \lambda_j^2 \tau^2 \varsigma^2\right), \\ \lambda_j &\sim C^+(0,1), \\ \tau &\sim C^+(0,1). \end{aligned} \tag{2.5}$$

이때, $C^+(0,1)$은 half-Cauchy 분포를 나타내며, τ의 분포로 임의의 고정된 값이나 truncated Cauchy 분포를 사용하기도 한다. 모형 (1.1)의 오차항 분산 ς^2에는 제프리스 사전분포를 사용한다.

Horseshoe 사전분포 (2.5)와 모형 (1.1)로부터 계산되는 β, τ 그리고 Λ의 완전 조건부 사후분포는 다음과 같다 (Makalic과 Schmidt, 2016):

$$\begin{aligned} \beta \mid \text{rest} &\sim N_p\left(A^{-1}X^TY,\ \varsigma^2 A^{-1}\right), \\ \varsigma^2 \mid \text{rest} &\sim \text{IG}\left(\frac{n+p}{2},\ \frac{1}{2}(Y-X\beta)^T(Y-X\beta) + \frac{1}{2\tau^2}\beta^T\Lambda^{-1}\beta\right), \\ \lambda_j^2 \mid \text{rest} &\sim \text{IG}\left(1,\ \frac{1}{\nu_j} + \frac{\beta_j^2}{2\varsigma^2\tau^2}\right), \quad j = 1, \ldots, p, \\ \tau^2 \mid \text{rest} &\sim \text{IG}\left(\frac{p+1}{2},\ \frac{1}{\xi} + \frac{1}{2\varsigma^2}\beta^T\Lambda^{-1}\beta\right), \\ \nu_j \mid \text{rest} &\sim \text{IG}\left(1,\ 1 + \frac{1}{\lambda_j^2}\right), \quad j = 1, \ldots, p, \\ \xi \mid \text{rest} &\sim \text{IG}\left(1,\ 1 + \frac{1}{\tau^2}\right). \end{aligned}$$

위의 식에서 잠재 변수 ξ와 ν_j를 도입하여 τ^2과 λ_j^2의 완전 조건부 사후분포를 유도하였으며, 이때, $A = (X^TX + 1/\tau^2\Lambda^{-1})$, $\Lambda = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2)$ 이다. 본 연구에서는 β의 사후분포 표본에 기반한 95% 신용 구간(credible interval)이 0을 포함하는지의 여부를 확인하여, 신용 구간이 0을 포함하지 않으면 유의한 변수로 선택하였다.

2.3. 변분 베이즈 방법

변분 베이즈 방법은 mean-field 근사를 이용하여 사후분포를 근사하는 방법으로, 다루기 쉬운 분포 중 쿨백-라이블러 발산을 최소화하는 분포를 사후분포의 근사 분포로 사용한다. 계산이 복잡하거나 어려운 사후분포를 근사할 수 있으며, 모형의 모수가 서로 독립이므로 계산이 용이해져 계산 시간을 단축할 수 있다.

Ray와 Szabó (2021)는 아래의 식 (2.6)과 같이 정규분포 $N(\mu_i, \sigma_i^2)$ slab과 0에서의 Dirac measure δ_0 spike로 구성된 mean-field variational family로 사후분포를 근사하는 변분 베이즈 방법을 제안하였다:

$$P_{MF} = \left\{ P_{\mu,\sigma,\gamma} = \bigotimes_{j=1}^{p} \left[\gamma_j N\left(\mu_j, \sigma_j^2\right) + (1-\gamma_j)\delta_0\right] :\ \mu_j \in \mathbb{R},\ \sigma_j > 0,\ \gamma_j \in [0,1] \right\}. \tag{2.6}$$

여기서 $\mu = (\mu_1, \ldots, \mu_p)^T$, $\sigma = (\sigma_1, \ldots, \sigma_p)^T$, $\gamma = (\gamma_1, \ldots, \gamma_p)^T$이다. 이때, 사후분포의 계산을 위한 사전분포는 Laplace slab을 사용한 spike and slab에 이진 잠재변수 z_j를 도입하여 아래의 식 (2.7)과 같이 선택한다:

$$\begin{aligned} \beta_j \mid z_j &\overset{\text{ind}}{\sim} z_j \text{Lap}(\lambda) + (1-z_j)\delta_0, \\ z_j \mid w &\overset{\text{iid}}{\sim} \text{Bernoulli}(w), \\ w &\sim \text{Beta}(a_0, b_0). \end{aligned} \tag{2.7}$$

이 방법은 성능 향상을 위해 slab의 분포로 Laplace 분포를 사용하여 추정된 계수가 과도하게 수축되는 것을 방지하였고, γ_j를 이진 변수가 아닌 구간 $[0,1]$ 사이의 포함 확률로 대체함으로써 모형 공간을 p차원으로 축소시켜 계산 시간을 극적으로 단축시켰다.

이후 근사 분포의 추론을 위해, 아래와 같이 사후분포 $\Pi(\cdot \mid Y)$와의 쿨백-라이블러 발산을 최소화하는 분포 $\tilde{\Pi}$를 찾는다:

$$\tilde{\Pi} = \underset{P_{\mu,\sigma,\gamma} \in P_{MF}}{\operatorname{argmin}} \ \mathrm{KL}(P_{\mu,\sigma,\gamma} \,\|\, \Pi(\cdot \mid Y)).$$

위의 최적화 문제를 풀기 위해, Ray와 Szabó (2021)에서는 coordinate ascent variational inference (CAVI) 알고리듬을 사용하였는데, 나머지 모수를 고정한 상태에서 쿨백-라이블러 발산을 최소화하는 변분 베이즈 사후분포의 모수 $\mu_j, \sigma_j, \gamma_j$를 순차적으로 업데이트한다.

먼저 μ_j와 σ_j의 경우, 다음의 두 식

$$\begin{aligned} f_j(\mu_j \mid \sigma, \mu_{-j}, \gamma, z_j = 1) = \mu_j \sum_{k \neq j} \left(X^T X\right)_{jk} \gamma_k \mu_k + \frac{1}{2} \left(X^T X\right)_{jj} \mu_j^2 - \left(Y^T X\right)_j \mu_j \\ + \lambda \sigma_j \sqrt{\frac{2}{\pi}} e^{\frac{-\mu_j^2}{2\sigma_j^2}} + \lambda \mu_j \left(1 - 2\Phi\left(-\frac{\mu_j}{\sigma_j}\right)\right), \end{aligned} \tag{2.8}$$

$$g_j(\sigma_j \mid \sigma_{-j}, \mu, \gamma, z_j = 1) = \frac{1}{2} \left(X^T X\right)_{jj} \sigma_j^2 - \log \sigma_j + \lambda \mu_j \sigma_j \sqrt{\frac{2}{\pi}} e^{\frac{-\mu_j^2}{2\sigma_j^2}} + \lambda \mu_j \left(1 - \Phi\left(\frac{\mu_j}{\sigma_j}\right)\right) \tag{2.9}$$

가 최대값이 되는 μ_j와 σ_j를 사용하여 쿨백-라이블러 발산을 최소화할 수 있다. 이때 (2.8)과 (2.9)를 최대로 만드는 값들은 닫힌 형식으로 구할 수 없으므로, R 내장함수인 `optimize`를 사용하여 수치적으로 계산한다. 다음으로 γ_j는 주어진 나머지 모수 μ, σ, γ_{-j}와 아래 식 (2.10)으로부터 $\mathrm{logit}^{-1}(\Gamma_i(\mu, \sigma, \gamma_{-i}))$를 계산한 값을 사용하여 쿨백-라이블러 발산을 최소화할 수 있다:

$$\begin{aligned} \log \frac{\gamma_j}{1-\gamma_j} &= \log \frac{a_0}{b_0} + \log \sqrt{\frac{\pi}{2}} \sigma_j \lambda + \left(Y^T X\right)_j \mu_j + \frac{1}{2} - \mu_j \sum_{k \neq j} \left(X^T X\right)_{jk} \gamma_k \mu_k \\ &\quad - \frac{1}{2} \left(X^T X\right)_{jj} \left(\sigma_j^2 + \mu_j^2\right) - \lambda \sigma_j \sqrt{\frac{2}{\pi}} e^{\frac{-\mu_j^2}{2\sigma_j^2}} - \lambda \mu_j \left(1 - 2\Phi\left(-\frac{\mu_j}{\sigma_j}\right)\right) \\ &=: \Gamma_i(\mu, \sigma, \gamma_{-j}). \end{aligned} \tag{2.10}$$

여기서 Φ는 표준정규분포의 누적분포함수를 의미한다.

또한 초기값 설정 및 모수 업데이트 순서에 민감하다는 기존 CAVI 알고리듬의 문제를 해결하기 위해, Ray와 Szabó (2021)에서는 평균 벡터 μ의 초기값을 계산한 다음, 추정치의 절댓값을 기준으로 계수를 내림차순으로 정렬하고 해당 순서로 모수를 업데이트하는 방법을 사용하였다. 본 연구에서도 동일한 알고리듬을 사용하였고, 평균 백터 μ의 초기값으로 ridge 회귀 추정량 $\hat{\mu}^{(0)} = (X^T X + I_p)^{-1} X^T Y$를 사용하였다. 상술한 CAVI 알고리듬을 구체적으로 제시하기 위해, Ray와 Szabó (2021)에 서술되어 있는 알고리듬 설명을 Algorithm 1에 옮겨두었다. 이때 임의의 p차원 벡터 $\mu = (\mu_1, \ldots, \mu_p)^T$에 대해, $\mathrm{order}(|\mu|) = (a_1, \ldots, a_p)^T$는 μ의 절댓값 기준 내림차순 순서로, $|\mu_{a_1}| \geq \cdots \geq |\mu_{a_p}|$를 만족한다. 업데이트 전후의 포함확률 $\gamma_{\mathrm{old},j}$와 γ_j의 최대 엔트로피 변화 $\Delta_H := \max_{j=1,\ldots,p} |H(\gamma_j) - H(\gamma_{old,j})|$가 미리 정해둔 작은 임계값 ϵ 아래로 떨어질 때까지 알고리듬을 반복한다. 여기서 $H(p) = -p \log p - (1-p)\log(1-p),\ p \in (0,1)$이다.

Ray 등 (2021)에서는 고차원 로지스틱 회귀모형에서의 변분 베이즈 방법을 수행하기 위한 알고리듬을 제안하였으며, 이는 변분 베이즈 사후분포 모수의 업데이트를 위한 식이 약간 변경되는 것을 제외하면 Algorithm

Algorithm 1 Variational Bayes for Laplace prior slabs

Initialize: $(\Delta_H, \sigma, \gamma)$, $\mu = \hat{\mu}^{(0)}$, $a = \text{order}(|\mu|)$
while $\Delta_H \geq \epsilon$ **do**
  for $j = 1$ to p **do**
    $i = a_j$
    $\mu_i = \text{argmax}_{\mu_i} f_i(\mu_i \mid \mu_{-i}, \sigma, \gamma, z_i = 1)$
    $\sigma_i = \text{argmax}_{\sigma_i} g_i(\sigma_i \mid \mu, \sigma_{-i}, \gamma, z_i = 1)$
    $\gamma_{old,i} = \gamma_i$, $\gamma_i = \text{logit}^{-1}(\Gamma_i(\mu, \sigma, \gamma_{-i}))$
  $\Delta_H = \max_i\{|H(\gamma_i) - H(\gamma_{old,i})|\}$
  $a = \text{order}(|\mu|)$

1과 동일한 과정으로 진행된다. 본 연구에서는 실제 자료 분석을 통해, Ray 등 (2021)에서 제안한 변분 베이즈 방법의 성능을 기존의 고차원 로지스틱 회귀분석 방법들과 비교하였다.

3. 시뮬레이션 자료 분석을 통한 비교연구

본 장에서는 다양한 모의실험 자료를 생성하여 고차원 선형 회귀분석을 수행하고, 빈도론 및 베이지안 변수선택 방법들의 성능을 비교한다. 고차원 선형 회귀모형 (1.1)을 고려하였으며, 설계 행렬의 각 성분을 표준정규분포 $N(0,1)$로부터 독립적으로 생성하였다. 비교분석을 위해, 다음과 같은 총 5개의 모의실험 상황을 고려하였다.

(1) $n = 100,\ p = 500,\ s = 10,\ \beta_{\text{signal,j}} = 10,\ \varsigma = 1$

(2) $n = 100,\ p = 500,\ s = 10,\ \beta_{\text{signal,j}} = 10,\ \varsigma = 5$

(3) $n = 100,\ p = 500,\ s = 10,\ \beta_{\text{signal,j}} \sim \text{Unif}(-10, 10),\ \varsigma = 1$

(4) $n = 100,\ p = 1000,\ s = 10,\ \beta_{\text{signal,j}} = 10,\ \varsigma = 1$

(5) $n = 100,\ p = 5000,\ s = 10,\ \beta_{\text{signal,j}} = 10,\ \varsigma = 1$

여기서 s는 0이 아닌 회귀계수의 개수를 의미하며, $\beta_{\text{signal,j}}$는 0이 아닌 j번째 회귀계수의 값을 의미한다. 각 상황에서, 회귀벡터 β는 $\beta = (\beta_{\text{signal,1}}, \ldots, \beta_{\text{signal,s}}, 0, \ldots, 0)^T \in \mathbb{R}^p$와 같이 생성하였다.

빈도론 변수선택 방법으로는 ridge, LASSO, elastic net을 사용하였으며, `glmnet` R 패키지의 `cv.glmnet` 함수를 사용하였다. 조절모수 λ의 경우 100개의 값 $(10^{-2}, \ldots, 10^{10})$ 중 10겹 교차검증을 이용하여 평균 교차검증 에러를 가장 최소화하는 λ를 선택하였으며, 조절모수 α의 경우 ridge는 0, LASSO는 1, elastic net은 $(0.0, 0.1, 0.2, \ldots, 1.0)$ 중 평균 교차검증 에러를 가장 최소화하는 α와 λ를 선택하였다.

베이지안 변수선택 방법으로는 spike and slab 사전분포, horseshoe 사전분포, 변분 베이즈 방법을 사용하였다. 먼저 spike and slab 사전분포는 `BoomSpikeSlab` R 패키지의 `lm.spike` 함수를 사용하였는데, 다른 설정들은 함수 내 기본 설정을 사용하였고, `niter = 5000`으로 설정하여 5,000개의 표본을 추출했다. 그 중 초기의 1,000개의 표본을 버리고 4,000개의 표본을 모형 성능 평가에 사용하였다. 다음으로 horseshoe 사전분포는 `horseshoe` R 패키지의 `horseshoe` 함수를 사용하였다. Spike and slab 사전분포와 마찬가지로 5,000개의 표본 중 1,000개를 버린 후 모형 성능 평가에 사용하였다. 마지막으로 변분 베이즈 방법은 `sparsevb` R 패키지의 `svb.fit` 함수를 사용하였다. 이때, Ray와 Szabó (2021)의 변분 베이즈 방법은 오차항의 분산 ς^2이 주어져 있다고 가정한다. 따라서 ς의 추정량 $\hat{\varsigma}$를 구한 뒤, 모형 (1.1)을 다음과 같이 변환하여 생각한다: $\tilde{Y} = \tilde{X}\beta + \tilde{Z}$. 여기서 $\tilde{Y} = Y/\hat{\varsigma}$, $\tilde{X} = X/\hat{\varsigma}$, $\tilde{Z} = (\varsigma/\hat{\varsigma})Z$, $Z \sim N_n(0, I_n)$이다. 이때, ς의 추정값 $\hat{\varsigma}$를 얻기 위해 `selectiveInference` R 패키지의 `estimateSigma` 함수를 사용하였다.

Table 1: Confusion matrix

		Predicted value	
		Positive	**Negative**
Actual value	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

모형의 성능을 비교하기 위해 다양한 평가 기준을 고려하였다. 우선 추정 및 예측 성능을 확인하기 위해 mean squared error (MSE)와 mean prediction error (MPE)를 측정하였다. 다음으로 변수선택 성능을 확인하기 위해 false negative (FN), false discovery rate (FDR), true positive rate (TPR), Matthews correlation coefficient (MCC)을 측정하였다. 마지막으로 계산 속도를 확인하기 위해 초 단위의 계산 시간을 측정하였다. 이때, ridge 회귀의 경우 변수선택이 불가능하므로 MSE, MPE 및 계산 시간만을 측정하였으며, spike and slab 사전분포와 변분 베이즈 방법의 경우 변수선택 성능 확인을 위한 평가값을 측정할 때 포함 확률이 0.5보다 큰 경우 회귀계수가 선택되었다고 판단하였다. Table 1에 FN을 비롯한 TP, TN, FP의 설명이 정리되어 있으며, 각 평가값들은 다음과 같이 정의된다:

$$\text{MSE} = \frac{1}{p}\left\|\hat{\beta}-\beta\right\|_2^2, \quad \text{MPE} = \frac{1}{n}\left\|X(\hat{\beta}-\beta)\right\|_2^2, \quad \text{FDR} = \frac{\text{FP}}{\text{TP}+\text{FP}},$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{MCC} = \frac{\text{TP}\times\text{TN - FP}\times FN}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}}.$$

Table 2는 시뮬레이션 자료 분석을 통해 측정한 결과를 정리한 표이다. 방법론 VB는 변분 베이즈 방법을 의미하고, 각 방법론을 총 50회씩 반복하여 각 평가값들의 평균과 표준편차로 결과를 나타내었다. 먼저 세팅 (1)에서의 분석 결과, 추정, 예측 및 변수선택에서 변분 베이즈 방법이 가장 좋은 성능을 보여준다. 계산 속도는 빈도론 방법인 LASSO에 비해 느리지만, 다른 베이지안 방법인 spike and slab 사전분포에 비해 약 12배, horseshoe 사전분포에 비해 약 34배 이상 빠른 속도를 보여준다. 오차항의 분산을 크게 한 세팅 (2)의 분석 결과에서도 세팅 (1)에서의 결과와 비슷한 경향을 보이는데, horseshoe 사전분포와 변분 베이즈 방법이 계산 속도를 제외한 추정, 예측 및 변수선택에서 가장 좋은 성능을 보여준다. 하지만 두 방법론의 추론 성능은 비슷한 반면, 변분 베이즈 방법의 계산속도가 약 33배 이상 빠른 것을 알 수 있다. 다음으로 0이 아닌 회귀계수의 값을 균일분포 Unif$(-10, 10)$에서 추출한 세팅 (3)의 분석 결과에서는 모든 방법론들이 추정 및 예측에서 비슷한 성능을 보여주었으나, 변수선택에서는 베이지안 방법론들이 우수한 성능을 보여줌을 확인하였다. 마찬가지로 변분 베이즈 방법이 다른 베이지안 방법론에 비해 계산 속도에서 큰 강점을 보여주는 것을 확인할 수 있다. 마지막으로 차원의 크기가 $p = 1000$인 세팅 (4)와 $p = 5000$인 세팅 (5)에서도 모든 방법론들이 비슷한 추정 및 예측 성능을 보였으나, horseshoe 사전분포와 변분 베이즈 방법이 좋은 변수선택 성능을 보여주고, 마찬가지로 비슷한 성능에 비해 변분 베이즈 방법이 horseshoe 사전분포에 비해 압도적으로 빠른 계산 속도를 보여준다. 이때 spike and slab 사전분포의 경우 변수선택 성능이 떨어지는 것을 확인할 수 있는데, 이는 변수의 개수 p가 커짐에 따라 모형 공간이 급격히 넓어져서 MCMC 표본들이 모형 공간을 충분히 탐험하지 못하는 것으로 해석할 수 있다.

4. 실제 자료 분석을 통한 비교연구

본 장에서는 실제 자료를 사용하여 고차원 로지스틱 회귀분석을 수행하고 변분 베이즈 방법과 다른 베이지안 변수선택 방법론 간의 성능을 비교한다. 분석에 사용한 자료는 (Golub 등, 1999)에 의해 연구된 백혈병 유전자 발현 자료로, 급성 골수성 백혈병(AML) 또는 급성 림프구성 백혈병(ALL)을 앓고 있는 72명의 환자($n = 72$)로부터 얻은 3,571개의 유전자 발현값($p = 3571$)으로 구성되어 있다.

Table 2: Comparing methods in high-dimensional linear regression

Metric	Method	(1)	(2)	(3)	(4)	(5)
MSE	Ridge	0.0287 ± 0.0185	0.0261 ± 0.0148	0.0141 ± 0.0086	0.0125 ± 0.0090	0.0038 ± 0.0023
	LASSO	0.0022 ± 0.0006	0.0103 ± 0.0022	0.0022 ± 0.0005	0.0014 ± 0.0008	0.0008 ± 0.0012
	Elastic net	0.0024 ± 0.0008	0.0104 ± 0.0022	0.0022 ± 0.0005	0.0013 ± 0.0004	0.0010 ± 0.0014
	Spike and slab	0.0160 ± 0.0250	0.0212 ± 0.0249	0.0014 ± 0.0049	0.0178 ± 0.0133	0.0060 ± 0.0003
	Horseshoe	0.0011 ± 0.0003	0.0065 ± 0.0017	0.0011 ± 0.0003	0.0004 ± 0.0001	0.0024 ± 0.0032
	VB	0.0007 ± 0.0002	0.0052 ± 0.0020	0.0008 ± 0.0003	0.0012 ± 0.0042	0.0028 ± 0.0025
MPE	Ridge	0.4967 ± 0.6882	0.6435 ± 0.3625	0.2205 ± 0.2639	0.4494 ± 0.7169	1.1724 ± 1.2548
	LASSO	0.0762 ± 0.0120	0.3682 ± 0.0524	0.0773 ± 0.0119	0.0807 ± 0.0126	0.1878 ± 0.3336
	Elastic net	0.0785 ± 0.0137	0.3696 ± 0.0490	0.0778 ± 0.0124	0.0822 ± 0.0126	0.1749 ± 0.2632
	Spike and slab	0.6960 ± 1.0982	0.9570 ± 1.1179	0.0626 ± 0.2194	1.6434 ± 1.2388	2.8433 ± 0.2460
	Horseshoe	0.0702 ± 0.0198	0.4325 ± 0.1025	0.0723 ± 0.0200	0.0581 ± 0.0184	0.0716 ± 0.0880
	VB	0.0326 ± 0.0073	0.2330 ± 0.0612	0.0402 ± 0.0113	0.0375 ± 0.0227	0.2163 ± 0.4549
FN	LASSO	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.1800 ± 0.8965
	Elastic net	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.1600 ± 0.6503
	Spike and slab	2.2800 ± 4.1109	2.8800 ± 4.4568	0.2000 ± 1.4142	5.3600 ± 4.7154	9.5600 ± 0.7602
	Horseshoe	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	2.6400 ± 3.4390
	VB	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0400 ± 0.1979	1.4200 ± 2.1579
FDR	LASSO	0.7978 ± 0.0783	0.8036 ± 0.0600	0.8151 ± 0.0701	0.8535 ± 0.0532	0.9199 ± 0.0551
	Elastic net	0.8269 ± 0.0663	0.8134 ± 0.0566	0.8276 ± 0.0623	0.8613 ± 0.0536	0.9278 ± 0.0472
	Spike and slab	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	Horseshoe	0.0070 ± 0.0292	0.0338 ± 0.0628	0.0139 ± 0.0524	0.0070 ± 0.0292	0.2385 ± 0.3737
	VB	0.0000 ± 0.0000	0.0902 ± 0.1383	0.0481 ± 0.0833	0.0358 ± 0.1681	0.4386 ± 0.4261
TPR	LASSO	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9820 ± 0.0896
	Elastic net	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9840 ± 0.0650
	Spike and slab	0.7720 ± 0.4111	0.7120 ± 0.4457	0.9800 ± 0.1414	0.4640 ± 0.4715	0.0440 ± 0.0760
	Horseshoe	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.7360 ± 0.3439
	VB	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	0.9960 ± 0.0198	0.8580 ± 0.2158
MCC	LASSO	0.4215 ± 0.0918	0.4176 ± 0.0754	0.4012 ± 0.0866	0.3640 ± 0.0736	0.2614 ± 0.0769
	Elastic net	0.3850 ± 0.0903	0.4058 ± 0.0738	0.3864 ± 0.0794	0.3530 ± 0.0745	0.2504 ± 0.0726
	Spike and slab	0.7939 ± 0.3799	0.7339 ± 0.4203	0.9800 ± 0.1414	0.5166 ± 0.4482	0.1156 ± 0.1766
	Horseshoe	0.9963 ± 0.0155	0.9820 ± 0.0338	0.9924 ± 0.0291	0.9964 ± 0.0153	0.7393 ± 0.3541
	VB	1.0000 ± 0.0000	0.9494 ± 0.0823	0.9741 ± 0.0455	0.9734 ± 0.1272	0.6309 ± 0.3372
Runtime (sec)	Ridge	0.2240 ± 0.0255	0.2172 ± 0.0083	0.2206 ± 0.0181	0.4050 ± 0.0185	1.7994 ± 0.0606
	LASSO	0.0778 ± 0.0181	0.0804 ± 0.0070	0.0770 ± 0.0152	0.1092 ± 0.0195	0.4046 ± 0.0323
	Elastic net	1.0018 ± 0.0502	1.0428 ± 0.0373	0.9950 ± 0.0398	1.5230 ± 0.0481	5.6984 ± 0.1402
	Spike and slab	5.3910 ± 0.7461	5.2626 ± 0.7452	5.5324 ± 0.3202	10.2386 ± 1.6110	49.6352 ± 11.0167
	Horseshoe	15.2718 ± 0.3201	15.0646 ± 0.0593	15.1024 ± 0.0662	28.8188 ± 0.0843	138.8736 ± 2.3363
	VB	0.4500 ± 0.0481	0.4460 ± 0.0457	0.4342 ± 0.0433	0.8032 ± 0.1131	3.5904 ± 0.6979

변수선택 방법론으로는 expectation-maximization (EM) 알고리즘을 사용하여 이진 데이터에 대한 고차원 베이지안 변수선택 문제를 해결하는 variable selection for binary data using the EM algorithm (Binary EMVS) (McDermott 등, 2016) 방법, 고차원 특성을 가진 분류 문제에 적합한 두꺼운 꼬리를 갖는 사전분포에 기반한 베이지안 다항 로지스틱 회귀모형인 bayesian logistic regression with heavy-tailed priors (HTLR) (Li와 Yao, 2018) 방법, 마지막으로 Ray와 Szabó (2021)가 제안한 변분 베이즈 방법을 사용하였다. 분석은 R 패키지를 사용하여 수행하였으며, Binary EMVS, HTLR, 변분 베이즈 방법에 대하여 각각 `BinaryEMVS`, `HTLR`, `sparsevb` 패키지의 `BinomialEMVS`, `htlr`, `svb.fit` 함수를 사용하였다. HTLR의 경우 6,000개의 MCMC 표본을 추출하여 초기의 1,000개의 표본을 버리고 5,000개의 표본을 모형 성능 평가에 사용하였다. Leave-one-out 방식으로 모형을 적합하였으며, 모형 성능 비교를 위해 오분류율(error rate)과 초 단위의 계산 시간을 측정하였다.

Table 3은 각 방법론을 사용하여 고차원 회귀분석을 수행한 결과를 정리한 표이다. 오분류율의 경우 Bi-

Table 3: Comparing methods in high-dimensional logistic regression

Method	Error rate	Runtime (sec)
Binary EMVS	0.0556	12.4883 ± 2.7501
HTLR	0.0694	15.8621 ± 0.1787
Variational Bayes	0.0556	2.0240 ± 0.0580

nary EMVS와 변분 베이즈 방법이 0.0556으로 가장 낮았으며, 이는 72개의 반응변수 중 4개의 값만을 잘못 분류한 경우로 좋은 분류 성능을 보여준다. HTLR의 경우 오분류율이 0.0694로 다른 방법론들에 비해 약간 높았지만, 세 개의 방법론들 모두 비슷하게 좋은 분류 성능을 보여준다. 하지만 계산 속도는 방법론 간 차이가 있는데, 변분 베이즈 방법이 Binary EMVS보다 평균적으로 약 6배 이상 빠른 계산 속도를 보여준다. 다만, `BinaryEMVS` 패키지는 순수하게 R 코드로 이루어진 반면 `sparsevb` 패키지의 경우 C++로 구현되어 있고, 함수의 최적화 정도에 차이가 있기 때문에 공정한 비교라고 보기 어렵다. 따라서 정확한 속도 비교를 위해서는 동일한 언어로 구현하는 것은 물론 각 함수를 최적화하는 작업이 필요할 것으로 보인다. 이 작업들은 본 논문의 주제를 넘어서므로, 향후 연구과제로 남겨두도록 한다.

5. 결론

본 논문에서는 Ray와 Szabó (2021)가 제안한 변분 베이즈 방법을 소개하고, 자료 분석을 통해 다른 변수선택 방법론들과의 성능을 비교하는 연구를 수행하였다. 모의실험 자료 분석을 통한 비교연구에서는 다양한 실험 세팅에서 변분 베이즈 방법이 다른 변수선택 방법론들에 비해 회귀계수 추정 및 변수선택에서 비교적 좋은 성능을 보여주는 것을 확인하였고, 계산 속도 측면에서도 대표적인 베이지안 변수선택 방법론인 spike and slab과 horseshoe 사전분포에 비해 압도적으로 빠른 속도를 보여줌을 확인하였다. 마찬가지로, 실제 자료 분석을 통한 비교연구에서도 다른 베이지안 변수선택 방법론들에 비해 변분 베이즈 방법이 비교적 좋은 분류 성능과 함께 가장 빠른 계산 속도를 보여주는 것을 확인하였다. 분석의 결과를 바탕으로 고차원 회귀분석에서 변분 베이즈 방법의 사용은 빠른 결과를 필요로 하는 경우 적절하며, 이를 통해 비교적 좋은 성능의 결과를 얻을 수 있을 것이라고 기대한다.

본 논문의 모의실험에서는 변수 간 상관관계가 없는 상황만을 고려하여 비교연구가 수행되었다. 논문 수정 중 한 심사위원의 코멘트를 통해 Ray와 Szabó (2021)의 보충자료에 제시된 모의실험 결과를 알게 되었는데, 그에 따르면 해당 설정이 계산 속도에는 영향을 미치지 않으며, 추정 및 변수선택 성능에서 여전히 효과적인 것을 확인할 수 있었다. 일반적으로, mean-field에 기반한 변분 베이즈 방법은 독립성을 가정하여 수행되기 때문에 변수 간 상관관계가 있는 설정에서 좋지 않은 성능을 보여줄 것이라 예상되는데, 위 결과는 직관에 반하는 다소 놀라운 결과라고 판단된다. 따라서, 이러한 경향이 실제로 변수 간에 상관관계가 존재하는 다양한 상황에서 일반적으로 관찰되는 것인지 조사하고, 독립성 가정이 없는 다른 베이지안 방법론들과의 비교연구를 수행하는 것은 향후 중요한 연구과제가 될 수 있을 것이다.

References

Carvalho C, Polson N, and Scott J (2010). The horseshoe estimator for sparse signals, *Biometrika*, **97**, 465–480.

George E and McCulloch R (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association*, **88**, 881–889.

Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, and Lander E (1999). Molecular classification of cancer: class discovery and class prediction

by gene expression monitoring, *Science*, **286**, 531–537.

Ishwaran H and Rao J (2005). Spike and slab variable selection: frequentist and Bayesian strategies, *The Annals of Statistics*, **33**, 730–773.

Li L and Yao W (2018). Fully Bayesian logistic regression with hyper-LASSO priors for high-dimensional feature selection, *Journal of Statistical Computation and Simulation*, **88**, 2827–2851.

Makalic E and Schmidt D (2016). A Simple Sampler for the Horseshoe Estimator, *IEEE Signal Processing Letters*, **23**, 179–182.

McDermott P, Snyder J, and Willison R (2016). Methods for Bayesian Variable Selection with Binary Response Data using the EM Algorithm, *arXiv: 1605.05429*.

Ray K, Szabó B, and Clara G (2020). Spike and slab variational Bayes for high dimensional logistic regression, *Advances in Neural Information Processing Systems*, **33**, 14423–14434.

Ray K and Szabó (2021). Variational Bayes for High-Dimensional Linear Regression With Sparse Priors, *Journal of the American Statistical Association*, To appear, 1–12.

Received January 29, 2022; Revised March 8, 2022; Accepted March 9, 2022

고차원 선형 및 로지스틱 회귀모형에 대한 변분 베이즈 방법 소개

장인송[a], 이경재[1,b]

[a]인하대학교 통계학과, [b]성균관대학교 통계학과

요 약

본 논문에서는 고차원 희소 회귀분석을 위한 기존의 베이지안 방법들을 소개하고, 다양한 모의실험 세팅에서 성능을 비교한다. 특히, 확장 가능하고 정확한 베이지안 추론을 가능하게 하는 변분 베이즈 방법(variational Bayes method) (Ray와 Szabó, 2021)에 중점을 둔다. 시뮬레이션 자료를 기반으로 한 희소 고차원 선형 회귀분석을 실시하고 변분 베이즈 방법의 성능을 다른 베이지안 및 빈도론 방법들과 비교한다. 로지스틱 회귀분석에서 변분 베이즈 방법의 실제 성능을 확인하기 위해 백혈병 유전자 발현 자료를 사용하여 실자료 분석을 수행한다.

주요용어: 변수선택, 회귀모형, spike and slab 사전분포, horseshoe 사전분포

[1]교신저자: (03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: leekjstat@skku.edu