Big Data Bayesian Linear Regression and Variable Selection by Normal-Inverse-Gamma Summation

Mickey Warner

12 March 2018

Review of the paper by Hang Qian (2017)

Linear Regression with Big Data

With n independent observations and k covariates, fitting the typical linear regressional model

$$y|X,\beta,\sigma^2 \sim N_n\left(X\beta,\sigma^2 I\right) \tag{1}$$

can be problematic when n is so large that we cannot load all the data into memory to perform standard computations.

Need a way to break up the data and perform computations on separate processors.

Normal-Inverse-Gamma (NIG) prior

If β and σ^2 are defined in the following way

$$\begin{aligned} \beta|\sigma^2 &\sim N_k(\mu, \sigma^2\Lambda^{-1}) \\ \sigma^2 &\sim IG(a, b) \end{aligned} \tag{2}$$

then the joint density function is given by

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-(a+k/2+1)} e^{-\frac{1}{\sigma^2}\left[b+\frac{1}{2}(\beta-\mu)^\top\Lambda^{-1}(\beta-\mu)\right]} \tag{3}$$

and we write $(\beta, \sigma^2) \sim NIG(\mu, \Lambda, a, b)$. The NIG distribution is a conjugate prior to the linear model.

A non-informative prior is $NIG(0_k, 0_{k\times k}, -k/2, 0)$.

NIG posterior

The posterior is given by

$$\beta, \sigma^2 | X, y \sim NIG(\overline{\mu}, \overline{\Lambda}, \overline{a}, \bar{b}) \tag{4}$$

where

$$\begin{aligned} \overline{\mu} &= (\Lambda + X^\top X)^{-1}(\Lambda\mu + X^{)} \\ \overline{\Lambda} &= \Lambda + X^\top X \\ \overline{a} &= a + \frac{n}{2} \\ \bar{b} &= b + \frac{1}{2}y^\top y + \frac{1}{2}\mu^\top \Lambda\mu - \frac{1}{2}\overline{\mu}^\top \overline{\Lambda}\overline{\mu} \end{aligned} \tag{5}$$

NIG summation

Consider the k-dimensional distributions $NIG(\mu_1, \Lambda_1, a_1, b_1)$ and $NIG(\mu_2, \Lambda_2, a_2, b_2)$. If a distribution $NIG(\mu, \Lambda, a, b)$ satisfies

$$\begin{aligned} \mu &= (\Lambda_1 + \Lambda_2)^{-1}(\Lambda_1\mu_1 + \Lambda_2\mu_2) \\ \Lambda &= \Lambda_1 + \Lambda_2 \\ a &= a_1 + a_2 + \frac{k}{2} \\ b &= b_1 + b_2 + \frac{1}{2}(\mu_1 - \mu_2)^\top(\Lambda_1^{-1} + \Lambda_2^{-1})^{-1}(\mu_1 - \mu_2) \end{aligned} \tag{6}$$

then it is said to be the sum of two NIG distributions

$$NIG(\mu, \Lambda, a, b) = NIG(\mu_1, \Lambda_1, a_1, b_1) + NIG(\mu_2, \Lambda_2, a_2, b_2) \tag{7}$$

Algorithm

Partition the data into m subsets

$$(X_1, y_1), \ldots, (X_m, y_m),$$

where X_i is $n_i \times k$, and y_i is $n_i \times 1$, and $n_1 + \cdots + n_m = n$.

These should be constructed so that $X_i^\top X_i$, $X_i^\top y_i$, and $y_i^\top y_i$ can be computed in memory.

Compute the NIG posterior (4) for each subset using (5), under a non-informative prior. Combine the results with (6) and (7), then add any prior information. The result is the posterior as if we used all of the data.

Simulation study

We simulate from the model

$$y_i \sim N(x_i^\top \beta, \sigma^2)$$

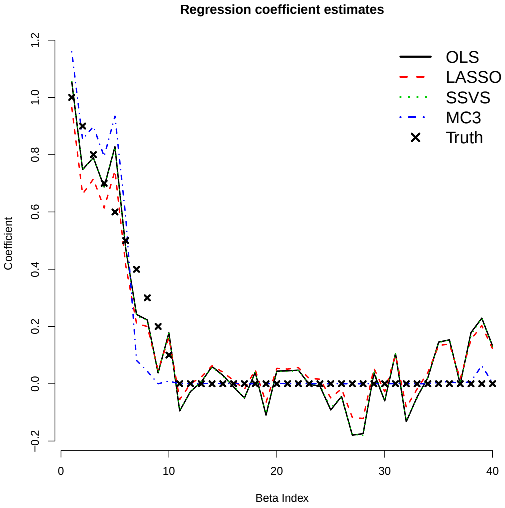
where $\beta = (1, 0.9, \ldots, 0.1, 0, \ldots, 0)^\top$, $\sigma = 10$ and the x_i's are from a zero-mean multivariate normal with correlation 0.99 for all variables.

Data are simulated with $n = 100,000$ and $k = 40$, but the method can easily handle much larger n.

Simulation study, continued

Comparisons are made between four models:

1. Standard linear model (OLS)
2. LASSO with penalty $\lambda = 10$
3. SSVS
4. MCMC model composition (MC^3)


Regression coefficient estimates
Coefficient
Beta Index
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
10
20
30
40
OLS
LASSO
SSVS
MC3
Truth

	OLS	LASSO	SSVS	MC^3
MSE	0.0206	0.0230	0.0203	0.0111

Table: MSE for β.

i	1	2	3	4	5
SSVS	1.00	1.00	1.00	1.00	1.00
MC^3	1.00	1.00	1.00	1.00	1.00

i	6	7	8	9	10
SSVS	0.75	0.07	0.05	0.00	0.02
MC^3	1.00	0.25	0.14	0.00	0.04

Table: Posterior $E(\gamma_i|y)$.