MATH 6450E: COMPUTER AGE STATISTICAL INFERENCE

Report of Project 1 — Topic 2

YE Rougang, Department of Mathematics

May 14, 2019

1 INTRODUCTION

This topic considers using Bayesian variable selection methods to analyze regression problems of large-scale data sets.

Specifically, consider a linear model that relates covariates $Z_1, \ldots, Z_m$ and variables $X_1, \ldots, X_p$ to the response Y :

$$Y = \sum_j Z_j \alpha_j + \sum_j X_j \beta_j + \epsilon, \tag{1.1}$$

where α_js are fixed effects, β_js are random effects and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Let γ_j be the variable indicating whether β_j is zero or not. We assume the following spike-slab prior:

$$\begin{cases} \beta_j \sim \mathcal{N}(0, \sigma_\beta^2), & \text{if } \gamma_j = 1 \\ \beta_j = 0, & \text{if } \gamma_j = 0 \end{cases}$$

where $Pr(\gamma_j = 1) = \pi$ and $Pr(\gamma_j = 0) = 1 - \pi$.

Thus $\boldsymbol{\gamma}$ is a Bernoulli distribution. Small values of π encourage sparse regression models, where a small proportion of the candidate variables X_i help predict the response Y. Denote hyperparameter vector $\boldsymbol{\theta} = \{\pi, \sigma_\beta^2, \sigma_\epsilon^2\}$. I will use the mean-field approximation to estimate the hyperparameter vector $\boldsymbol{\theta}$ and the posterior distribution of $\{\beta_j\}$.

2 METHOD

Consider a Bayesian model which involves observed variables X and latent variables Z. Idea of mean-field approximation starts from the decomposition of the log marginal probability of observed variables, that is,

$$\log p(X) = \mathscr{L}(q) + KL(q||p)$$

where

$$\mathscr{L}(q) = \int q(Z)\log\left\{\frac{p(X,Z)}{q(Z)}\right\}dZ,$$
$$KL(q||p) = -\int q(Z)\log\left\{\frac{p(Z|X)}{q(Z)}\right\}dZ.$$

$\mathscr{L}(q)$ is the variational lower bound and $KL(q||p)$ is the Kullback-Leibler divergence.Thus the lower bound obtains its maximum value when the KL divergence vanishes, which occurs when $q(Z)$ equals to the posterior distribution $p(Z|X)$.

Put it into the framework here, latent variables Z are β,γ. I restrict $q(\beta,\gamma)$ to be of the form

$$q(\beta,\gamma) = \prod_{j=1}^{p} q(\beta_j,\gamma_j). \tag{2.1}$$

The individual factors have the form

$$q(\beta_j,\gamma_j) = \begin{cases} \alpha_j N(\mu_j,\sigma_j^2) & \text{if } \gamma_j = 1 \\ (1-\alpha_j)\delta_0(\beta_j) & \text{if } \gamma_j = 0 \end{cases} \tag{2.2}$$

where δ_0 is the delta mass (or "spike") at 0. With probability α_j, the additive effect β_j is normal with mean μ_j and variance s_j^2 (the "slab"), and with probability 1-α_j, the variable has no effect on Y.

Using the same transformation mentioned in *varbvs: Fast Variable Selection for Large-scale Regression*, I analytically integrate out the fixed effects $\{\alpha_j, j = 1,2,\ldots,m\}$ in the linear model by using:

$$|\Sigma_0|^{1/2}P(y|X,Z,\beta,\sigma_\epsilon^2) = |Z^TZ|^{-1/2}P(\hat{y}|\hat{X},\beta,\sigma_\epsilon^2),$$

in which $P(y|X,Z,\beta,\sigma^2)$ is the multivariate normal likelihood of the linear regression model 1.1, while $P(\hat{y}|\hat{X},\beta,\sigma^2)$ is the likelihood given by linear regression $\hat{y} = \hat{X}\beta+\epsilon$, α is assigned a multivariate normal prior with zero mean and covariance Σ_0 such that $|\Sigma_0^{-1}|$ is close to zero and define $\hat{X} = X - Z(Z^TZ)^{-1}Z^TX$ and $\hat{y} = y - Z(Z^TZ)^{-1}Z^Ty$.

Then one can discard the linear effects of covariates Z by replacing all instances of X with $\hat{X}$ and y with $\hat{y}$, and by multiplying the likelihood by $|Z^TZ|^{-1/2}$. Thus, in the following calculation, I assume the simpler linear regression $y = X\beta+\epsilon$, replace X with $\hat{X}$ and y with $\hat{y}$. Multiplying by $|Z^TZ|^{-1/2}$ can generate the final solution.

By the "fully-factorized" class of approximating distributions 2.1 and 2.2, the variational lower bound can be derived as follows:

$$
\begin{aligned}
F(\theta, X, y) = \mathscr{L}(q) &= \int\int q(\beta,\gamma)\log\left\{\frac{p(y,\beta,\gamma)}{q(\beta,\gamma)}\right\} d\beta d\gamma \\
&= \int\int \prod_{j=1}^{p} q(\beta_j,\gamma_j)\left(\log p(y,\beta,\gamma) - \log q(\beta,\gamma)\right) d\beta d\gamma \\
&= -\frac{n}{2}\log(2\pi\sigma_\epsilon^2) - \frac{\|y - Xr\|_2^2}{2\sigma_\epsilon^2} - \frac{1}{2\sigma_\epsilon^2}\sum_{j=1}^{p}(X^TX)_{jj}(\alpha_j(s_j^2+\mu_j^2) - \alpha_j^2\mu_j^2)) \\
&- \sum_{j=1}^{p}\alpha_j\log(\frac{\alpha_j}{\pi}) - \sum_{j=1}^{p}(1-\alpha_j)\log(\frac{1-\alpha_j}{1-\pi}) + \sum_{j=1}^{p}\frac{\alpha_j}{2}\left[1 + \log\frac{s_j^2}{\sigma_\beta^2} - \frac{s_j^2+\mu_j^2}{\sigma_\beta^2}\right]
\end{aligned}
\tag{2.3}
$$

where $\| \, . \, \|_2$ is the Euclidean norm, r is a column vector with entries $r_i = \alpha_i\mu_i$.
The above result is calculated by the following details:

$$
\prod_{j=1}^{p} q(\beta_j,\gamma_j) = \prod_{j=1}^{p}\left[\frac{1}{\sqrt{2\pi}s_j}e^{-\frac{(\beta_j-\mu_j)^2}{2s_j^2}}\alpha_j^{\gamma_j}(1-\alpha_j)^{1-\gamma_j}\right] \tag{2.4}
$$

$$
p(y,\beta,\gamma) = \prod_{j=1}^{p} p(y_j,\beta_j,\gamma_j) \tag{2.5}
$$

$$
= \prod_{j=1}^{p}\left[\frac{1}{\sqrt{2\pi}\sigma_\epsilon}e^{-\frac{(y_j-\gamma_j\sum_{k=1}^{p}X_{jk}\beta_k)^2}{2\sigma_\epsilon^2}}\frac{1}{\sqrt{2\pi}\sigma_\beta}e^{-\frac{\beta_j^2}{2\sigma_\beta^2}}\pi^{\gamma_j}(1-\pi)^{1-\gamma_j}\right] \tag{2.6}
$$

For example, by $p(y,\beta,\gamma)$, using the trick of completing the square will produce the term $\frac{\|y-Xr\|_2^2}{2\sigma_\epsilon^2}$.

3 Result

As we need to maximize the variational lower bound or minimize the KL divergence, update for the free parameters are obtained by taking partial derivatives of the lower bound $F(\theta, X, y)$ 2.3, setting these partial derivatives to zero. That is,

$$
\mu_j = \frac{s_j^2}{\sigma_{beta}^2\sigma_\epsilon^2}\left((X^Ty)_j - \sum_{k\neq j}(X^TX)_{jk}\alpha_k\mu_k\right) \tag{3.1}
$$

	N	P	MFApprox	$\alpha = 10^{-10}$	$\alpha = 10^{-5}$	$\alpha = 0.001$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 5$
1	200	50	0.189341	0.292274	0.292268	0.291639	0.257792	0.205851	0.205851
2	200	200	0.167542	0.321091	0.313204	0.264296	0.207271	0.174361	0.174361
3	500	200	0.171335	0.257812	0.257802	0.256870	0.212600	0.180137	0.180137
4	500	500	0.175718	0.291719	0.289246	0.262307	0.221243	0.182700	0.182650
5	1000	300	0.170906	0.255629	0.255623	0.255006	0.220253	0.179568	0.179568
6	1000	1000	0.164817	0.264645	0.260156	0.240020	0.205532	0.170480	0.170444

Table 3.1: RMSE

$$s_j^2 = \frac{\sigma_\beta^2}{\frac{(X^T X)_{jj}}{\sigma_\epsilon^2} + \frac{1}{\sigma_\beta^2}} \tag{3.2}$$

α_j satisfies the following equation

$$\frac{\alpha_j}{1-\alpha_j} = \frac{\pi_j}{1-\pi_j} \frac{s_j}{\sigma_\beta} e^{\frac{\mu_j^2}{2s_j^2}} \tag{3.3}$$

Similarly, hyper parameters can be updated as follows, which is derived by corresponding gradients of the lower bound:

$$\sigma_\epsilon^2 = \frac{\|y - Xr\|_2^2 + \sum_{j=1}^p (X^T X)_{jj}(\alpha_j(s_j^2 + \mu_j^2) - \alpha_j^2 \mu_j^2))}{n} \tag{3.4}$$

$$\sigma_\beta^2 = \frac{\sum_{j=1}^p \alpha_j (s_j^2 + \mu_j^2)}{\sum_{j=1}^p \alpha_j} \tag{3.5}$$

With the above update formulas, I do simulations to compare estimations with the mean-field approximation and Lasso, which can select variables through adding the l_1 penalty term in the regression models. Changing different sizes of samples and features, I test different α values for Lasso, results of RMSE (root-mean-square error) is given in table 3.1.

N is the number of samples and P is the number of features. Thus, one can find that mean-field approximation performs better than Lasso with different α values from the point of RMSE among different combinations of N and P. Actually, Lasso with $\alpha > 0.1$ already turns a lot of coefficients of the regression model to 0.

4 Conclusion

As I find the formulas in *varbvs: Fast Variable Selection for Large-scale Regression* actually require $\beta \sim N(0, \sigma_\beta^2 * \sigma_\epsilon^2)$, which is different from the set up in our project, so I used mean-field approximation to derive the above variational lower bound, corresponding update expressions for hyper parameters $\boldsymbol{\theta}$ and the posterior distribution of random effects, β. Simulation studies are did to compare the estimated error of $\hat{\beta}$ through mean-field approximation and Lasso.