Variational Inference Notes

Bufei Guo

November 26, 2018

1 Introduction and Notation

Variational inference could make a fast approximation on the statistic model especially when the model parameters are complicated. When one is doing Gibbs sampling the ultimate goal is to attain the true value by simultaneously maximizing the likelihood functions. In the context of variational inference, instead of maximizing the log-likelihood of $\mathbf{y}$, we try to maximize the lower bound of log-likelihood. Let q be some arbitrary density function over the parameter space Θ.Then:

$$\begin{aligned}
\log p(\mathbf{y}) &= \log p(\mathbf{y}) \int q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log\{\frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})/q(\boldsymbol{\beta}_p, \mathbf{z})}{p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})/q(\boldsymbol{\beta}_p, \mathbf{z})}\} d\boldsymbol{\beta}_p d\mathbf{z} \\
&= \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log\{\frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})}{q(\boldsymbol{\beta}_p, \mathbf{z})}\} d\boldsymbol{\beta}_p d\mathbf{z} + \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log\{\frac{q(\boldsymbol{\beta}_p, \mathbf{z})}{p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})}\} d\boldsymbol{\beta}_p d\mathbf{z} \\
&\geq \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log\{\frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})}{q(\boldsymbol{\beta}_p, \mathbf{z})}\} d\boldsymbol{\beta}_p d\mathbf{z}
\end{aligned} \tag{1}$$

Where $\boldsymbol{\theta} = (\boldsymbol{\beta}_p, \mathbf{z})$, $q(\cdot)$is some other distribution used to approximate the real distribution $p(\cdot)$. This inequality holds because the second integral $\int q(\boldsymbol{\beta}_p, \mathbf{z}) \log\{\frac{q(\boldsymbol{\beta}_p, \mathbf{z})}{p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})}\} d\boldsymbol{\beta}_p d\mathbf{z}$
is the *Kullback-Leibler* distance between $q(\boldsymbol{\beta}_p, \mathbf{z})$ and $p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})$, which is always non-negative, with the equality if and only if $q(\boldsymbol{\beta}_p, \mathbf{z}) = p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})$, and then the estimation will coincide with the estimation given by Gibbs sampling.
Here we will use "naive mean approximation" to calculate $q(\cdot)$, by making assumption that the parameters of interest are independent. One can also use other approximation like "generalized mean field", by dividing the parameters of interest into groups, with dependence inside each group. Then the lower bound can be written in the form of:

$$\begin{aligned}
\int q(\boldsymbol{\beta}_p, \mathbf{z}) \log\{\frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})}{q(\boldsymbol{\beta}_p, \mathbf{z})}\} d\boldsymbol{\beta}_p d\mathbf{z} &= \int q(\boldsymbol{\beta}_p) q(\mathbf{z}) \log\{\frac{p(\mathbf{y}|\boldsymbol{\beta}_p, \mathbf{z}) p(\boldsymbol{\beta}_p) p(\mathbf{z})}{q(\boldsymbol{\beta}_p) q(\mathbf{z})}\} d\boldsymbol{\beta}_p d\mathbf{z} \\
&= \int q(\boldsymbol{\beta}_p) q(\mathbf{z}) \log\{p(\mathbf{y}|\boldsymbol{\beta}_p, \mathbf{z})\} d\boldsymbol{\beta}_p d\mathbf{z} + \int q(\boldsymbol{\beta}_p) \log\{\frac{p(\boldsymbol{\beta}_p)}{q(\boldsymbol{\beta}_p)}\} d\boldsymbol{\beta}_p \\
&+ \int q(\mathbf{z}) \log\{\frac{p(\mathbf{z})}{q(\mathbf{z})}\} d\mathbf{z}
\end{aligned} \tag{2}$$

The form of $q(\cdot)$ is chosen from the density functions over Θ. Maximizing this lower bound with respect to each parameter and ignore the terms that do not contain the parameter of interest, the $q^*(\cdot)$ is then:

$$q(\theta_i) \propto \exp(\mathbb{E}_{\boldsymbol{\theta}-i}(\log p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\theta}_{-i})))$$

where $\boldsymbol{\theta}_{-i}$ represents the vector of variables exclude the ith variable θ_i.

2 Variational Inference on Asymmetric Laplace Distribution

Consider the model $y_i = \mathbf{x}_i'\boldsymbol{\beta}_p + \theta z_i + \tau\sqrt{\sigma z_i}u_i$, $i = 1, ..., n$. $\mathbf{x_i}$ and $\boldsymbol{\beta}_p$ are vectors in $\mathbb{R}^k$, the subscript p in $\boldsymbol{\beta}_p$ represents the quantile we are interested in. u_i's follow standard normal distribution and z is exponential distribution with parameter σ. θ and τ are constants with $\theta = \frac{1-2p}{p(1-p)}$ and $\tau^2 = \frac{2}{p(1-p)}$.
y_i follows a asymmetric Laplace distribution (ALD), $ALD(\mathbf{x}_i'\boldsymbol{\beta}_p, 1, p)$. The probability density function of ALD is:

$$f(x; \mu, \sigma, p) = \tfrac{p(1-p)}{\sigma} \exp(-\tfrac{(x-\mu)}{\sigma}[p - \boldsymbol{I}(x \leq \mu)])$$

where $0 \leq p \leq 1$ is the skew parameter, $\sigma > 0$ is the scale parameter, $\mu \in \mathbb{R}$ is the location parameter, and $\boldsymbol{I}(\cdot)$ is the indicator function.
In the context of regression $\mathbf{y}$ on $\mathbf{X} = (1, \mathbf{x}_1, \ldots, \mathbf{x}_n)$, the error part $\epsilon_i = \theta z_i + \tau\sqrt{\sigma z_i}u_i$ of regression also follows an asymmetric Laplace distribution with $f(\epsilon) = \frac{p(1-p)}{\sigma} \exp(-\rho_p(\frac{\epsilon}{\sigma}))$,$\rho_p(\cdot)$ is the loss function defined as $\rho_p(x) = x(p - \mathbf{I}(x < 0))$. We will first consider the case where, σ is 1 and this model simplifies to $f(\epsilon) = p(1-p)\exp(-\rho_p(\epsilon))$. Condition on $\boldsymbol{\beta}_p$ and $\mathbf{z}$, $\mathbf{y}$ follows a multivariate normal distribution $N_p \sim (\mathbf{X}\beta_p - \theta z, \tau^2 Z)$, Z is a diagonal matrix with its diagonal being $(z_1, \ldots, z_n)$:

$$f(\mathbf{y}|\boldsymbol{\beta}_p, \mathbf{z}) \propto (\Pi_{i=1}^n z_i^{-\frac{1}{2}}) \exp(-\Pi_{i=1}^n \tfrac{(y_i - x_i'\beta_p - \theta z_i)^2}{2\tau^2 z_i})$$

Algorithm:

- Initialize mean and variance $\boldsymbol{\mu}_q$ and Σ_q for $\boldsymbol{\beta}_p$
- while absolute change in lower bound $l \geq t$, t is the tolerance given
- Update parameters in distribution of $q(\mathbf{z})$. $q(z_i) \sim GIG(\frac{1}{2}, a_{q_i}, b_{q_i})$, where:

$$a_{q_i} = 2 + \frac{\theta^2}{\tau^2}$$

$$b_{q_i} = \frac{y_i^2 - 2y_i\mathbf{x}_i'\boldsymbol{\mu}_q + \mathbf{x}_i'(\boldsymbol{\mu}_q\boldsymbol{\mu}_q' + \Sigma_q)\mathbf{x}_i}{\tau^2}$$

- Update parameters in distribution of $q(\boldsymbol{\beta}_p)$. $q(\boldsymbol{\beta}_p) \sim N(\mu_q, \Sigma_q)$, where:

$$\Sigma_q = (\sum_{i=1}^n \frac{\mathbf{x}_i\mathbf{x}_i'}{\tau^2}\mathbb{E}(\frac{1}{z_i}) + \Sigma_{p0}^{-1})^{-1}$$

$$\mu_q = \Sigma_q(\sum_{i=1}^n \frac{\mathbf{x}_i y_i}{\tau^2}\mathbb{E}(\frac{1}{z_i}) - \sum_{i=1}^n \frac{\theta}{\tau^2}\mathbf{x}_i + \Sigma_{p0}^{-1}\boldsymbol{\mu}_{p0})$$

- Calculate lower bound l:

$$l = \mathbb{E}_{q(\mathbf{z}),q(\boldsymbol{\beta}_p)}(\mathbf{y}|\mathbf{z},\boldsymbol{\beta}_p) + \mathbb{E}_{q(\mathbf{z})}\log(p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})}log(q(\mathbf{z}))$$
$$+\mathbb{E}_{q(\boldsymbol{\beta}_p)}\log(p(\boldsymbol{\beta}_p)) - \mathbb{E}_{q(\boldsymbol{\beta}_p)}\log(q(\boldsymbol{\beta}_p))$$

3 Comparison

The result of variational inference is compared with that of Gibbs Sampling. Here I demonstrate it using the model where $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_p + \epsilon$, where $\boldsymbol{\beta}_p = (1,2,3)'$, $\mathbf{X} = (\mathbf{1},\mathbf{x_1},\mathbf{x_2})$ and $\epsilon \sim N(0, 0.6^2)$.
The predictive mean square error(MSE) under each methods is calculated to give a brief comparison. Here I used the estimation given by $rq(\cdot)$ function in package quantreg and $bayesQR(\cdot)$ in package bayesQR, where the Gibbs sampling will be applied.