Stacks : Robust Regression

Sri Krishna Priya Dhulipala

November 28, 2017

Abstract

In this assignment I applied MCMC based and Variational inference for Gaussian likelihood based Bayesian Linear Regression on Stacks data.

1 Model Equation

The parameters of the following linear regression model on stack loss data are estimated using two inference techniques: Regression using MCMC and ridge regression using VI

$$m_i = b_0 + b_1 \times z_{1i} + b_2 \times z_{2i} + b_3 \times z_{3i} \tag{1}$$

$$y_i \sim Normal(m_i, \beta^-1) \tag{2}$$

2 Bayesian Linear Regression with MCMC

2.1 Gibbs Sampling

1. Likelihood function of the model is given by

$$p(Y \backslash B, \sigma^2) = (2\pi\sigma^2)^{-T/2} exp(\frac{-(Y - XB)'(Y - XB)}{2\sigma^2}), \tag{3}$$

2. Used a normally distributed prior for coefficients $B = [b_0, b_1, b_2, b_3]$ given by $P(B) \sim N(B_0, \Sigma_0)$. B_0 and Σ_0 are chosen as indicated in the assignment.

$$P(B) = (2\pi)^{-K/2} |\Sigma_0|^{-\frac{1}{2}} exp[-0.5(B - B_0)' \Sigma_0^{-1} (B - B_0)], \tag{4}$$

3. And for error variance we use an inverse gamma prior as gamma distribution is conjugate to normal distribution

$$p(\sigma^2) \sim \Gamma^{-1}(\frac{T_0}{2}, \frac{\theta_0}{2}), \tag{5}$$

4. The full conditional for B given σ^2 is then

$$p(B\backslash\sigma^2, Y) = (2\pi)^{-K/2}|\Sigma_0|^{-\frac{1}{2}}exp[-0.5(B-B_0)^{'}\Sigma_0^{-1}(B-B_0)]\times \\ (2\pi\sigma^2)^{-T/2}exp(\frac{-(Y-BX)^{'}(Y-BX)}{2\sigma^2}), \quad (6)$$

This is a gaussian distribution with the mean and the variance given by

$$M^* = (\Sigma_0^{-1} + \frac{1}{\sigma^2}X_t^{'}X_t)^{-1}(\Sigma_0^{-1}B_0 + \frac{1}{\sigma^2}X_t^{'}Y_t), \quad (7)$$

$$V^* = (\Sigma_0^{-1} + \frac{1}{\sigma^2}X_t^{'}X_t)^{-1}, \quad (8)$$

5. The full conditional of σ^2 given B^1 is then an inverse gamma distribution given by

$$p(\sigma^2\backslash B, Y) \sim \Gamma^{-1}(\frac{T_1}{2}, \frac{\theta_1}{2}), \quad (9)$$

$T_1 = T_0 + T$
$\theta_1 = \theta_0 + (Y_t - B^1X_t)^1(Y_t - B^1X_t)$

6. Repeat the above two steps M times to obtain $B^1...B^M$ and $(\sigma^2)^1...(\sigma^2)^M$

7. The mean of the last 10,000 values of B and σ^2 from the above iterations are used as the estimates. First 1000 samples were discarded as burn-in.

3 Variational Inference based Bayesian Linear Regression

3.1 Ridge Regression

In order to introduce shrinkage on the coefficients $w = [b_0, b_1, b_2, b_3]$, we use a normal prior with zero mean and inverse variance α as shown below:

$$p(w/\alpha) = \prod_{m=1}^{M} N(w_m|0, \alpha^{-1}) \quad (10)$$

This α is modeled using a gamma prior with a and b as the shape and scale parameters respectively. This prior is selected because it is conjugate to the Gaussian.

$$p(\alpha/a, b) = Gamma(\alpha|a_0, b_0) \quad (11)$$

Like in the previous case, noise inverse variance β is also modelled by a gamma distributions given below:

$$p(\beta/c, d) = Gamma(\beta|c_0, d_0) \quad (12)$$

Assuming posterior independence between weights w and variance parameters α and β, we have

$$p(w, \alpha, \beta | t; a, b, c, d) \approx q(w, \alpha, \beta) = q(w)q(\alpha)q(\beta) \tag{13}$$

The parameters of the variational distributions i.e mean and variance for $q(w)$ which is normal and shape and scale parameters for $q(\alpha)$ and $q(\beta)$ which are gamma distributions can be derived to be :

$$w_N = V_N X^T Y \tag{14}$$

$$V_N^{-1} = \alpha I + X^T X \tag{15}$$

$$a_N = a_0 + D/2 \tag{16}$$

$$b_N = b_0 + (\beta/2) w^T w + Tr(V_N)/2 \tag{17}$$

$$c_N = c_0 + N/2 \tag{18}$$

$$d_N = d_0 + 1/2(\|Y - XW\|^2 + \alpha w^T w) \tag{19}$$

The lower bound given below is maximized by iterating over the updates for V_N, w_N, a_N, b_N, c_N and d_N until $L(Q)$ reaches a plateau.

$$\begin{aligned} L(Q) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_n \left(\frac{a_N}{b_N}(y_n - w_N^T x_n)^2 + x_n^T V_N x_n\right) + \frac{1}{2}\ln|V_N| \\ + \frac{D}{2} - \ln\Gamma(a_0) + a_0 \ln b_0 - b_0 \frac{a_N}{b_N} + \ln\Gamma(a_N) - a_N \ln b_N + a_N \\ - \ln\Gamma(c_0) + c_0 \ln d_0 + \ln\Gamma(c_N) - c_N \ln d_N \end{aligned} \tag{20}$$

4 Results

We observe that convergence is obtained much faster using variational inference compared to Gibbs Sampling.

The coefficient values using both the methods and their respective training errors are given below.

1. For VI
   (a) RMSE value = 2.91873878072
   (b) b_0 = -38.95068316
   (c) b_1 = 0.71761505
   (d) b_2 = 1.29079727
   (e) b_3 = -0.16359841
2. for MCMC

(a) RMSE value = 3.16640285521

(b) b_0 = -17.53387862

(c) b_1 = 0.76258627

(d) b_2 = 1.19216922

(e) b_3 = -0.41832853

Below, the respective predictions are plotted:

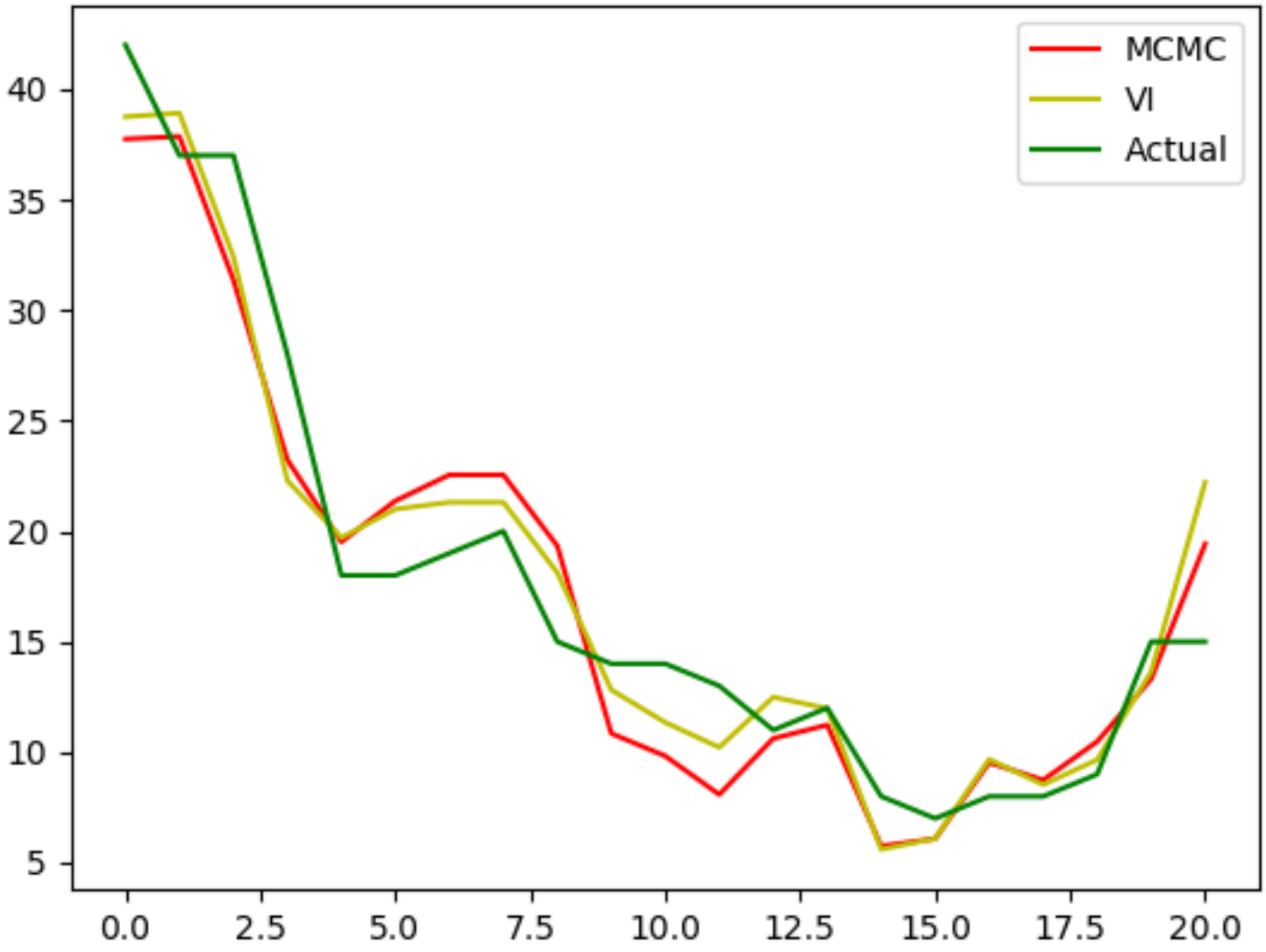

Figure 1: Predicted and Actual Y values for VI and MCMC