Statistical Computing for Scientists and Engineers

Homework 4

Jiale Shi

Oct/29/2018

1 Accept-Reject

Generate samples of a standard normal distribution, $f(x) \sim N(0,1)$, using the accept-reject method with a double-exponential proposal distribution, $g(x|\alpha) = (\alpha/2)\exp(-\alpha|x|)$.

(a) Derive the upper bound for the likelihood ratio, $M = f(x)/g(x)$ and show that the ideal acceptance rate is obtained when $\alpha = 1$

Solution: a standard normal distribution, $f(x) \sim N(0,1)$

$$f(x|0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{1}$$

$$M = \frac{f(x)}{g(x)} = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}{(\alpha/2)\exp(-\alpha|x|)} = \frac{\sqrt{2}}{\alpha\sqrt{\pi}} \exp\left\{-\frac{|x|^2}{2} + \alpha|x|\right\} \tag{2}$$

The ratio M is max at $|x| = \alpha$.

$$M = \frac{\sqrt{2}}{\alpha\sqrt{\pi}} \exp\left\{\frac{\alpha^2}{2}\right\} \tag{3}$$

$$\begin{aligned} \frac{\partial M}{\partial \alpha} &= \frac{\sqrt{2}}{\sqrt{\pi}} \exp\left\{\frac{\alpha^2}{2}\right\}(1 - \frac{1}{\alpha^2}) = 0 \\ \alpha = 1, M' &= \frac{\sqrt{2}}{\sqrt{\pi}} \exp\left\{\frac{1}{2}\right\} \end{aligned} \tag{4}$$

(b) Implement the accept-reject method and plot the true PDF and the proposal distribution for $\alpha = 1$ super-imposed on to the normalized histogram of your samples.

Solution:

$\alpha = 1$ and $M' = \frac{\sqrt{2}}{\sqrt{\pi}} \exp\left\{\frac{1}{2}\right\}$

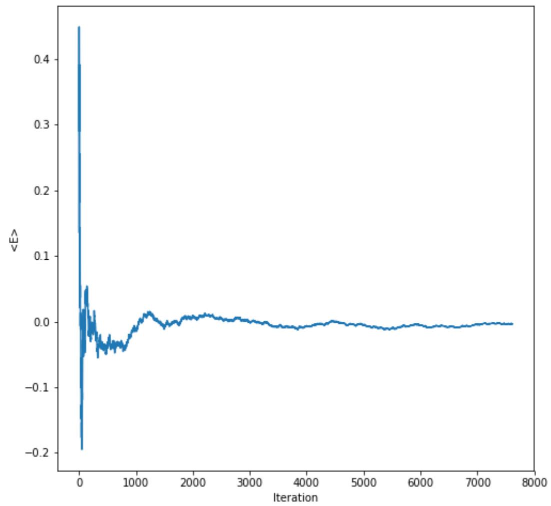

Figure 1: $< E >$- iteration for $\alpha = 1$

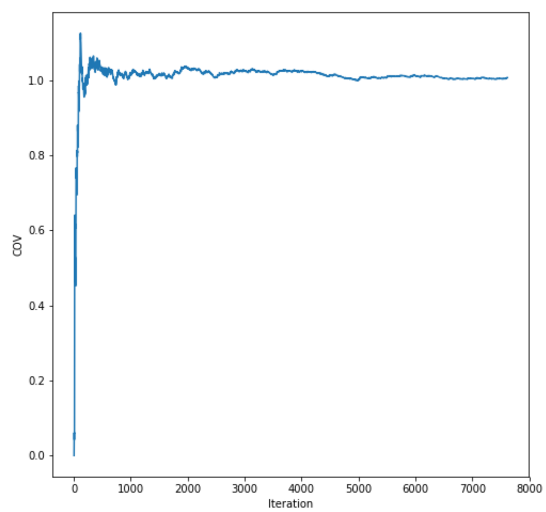

Figure 2: COV- iteration for $\alpha = 1$

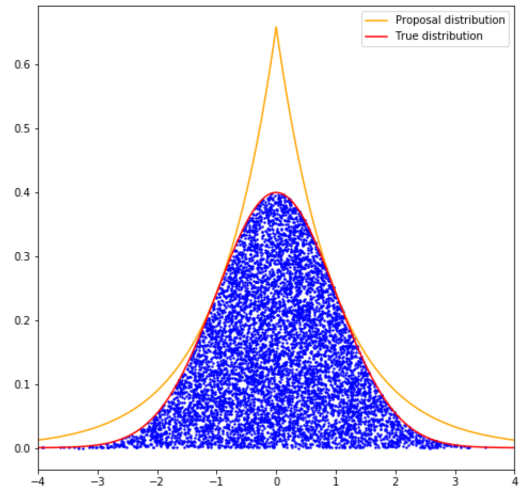

Figure 3: the true pdf and the proposal distribution for $\alpha = 1$

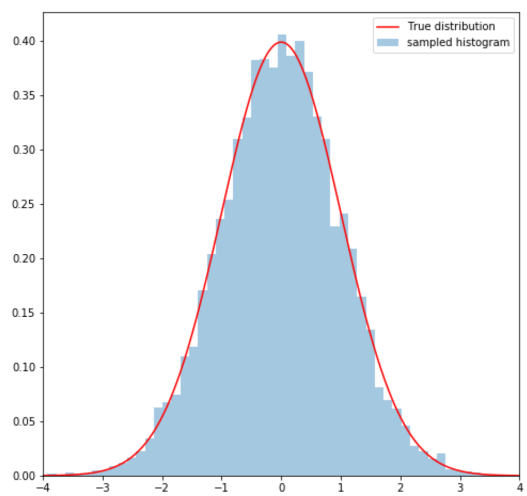

Figure 4: histogram of samples for $\alpha = 1$

(c) Repeat part (b) but now use a sub-optimal proposal distribution with $\alpha = 2$, plot both distributions and your histogram. How do the acceptance rates compare?

Solution:

$\alpha = 2$ and $M' = \frac{\sqrt{2}}{2\sqrt{\pi}} \exp\{2\}$

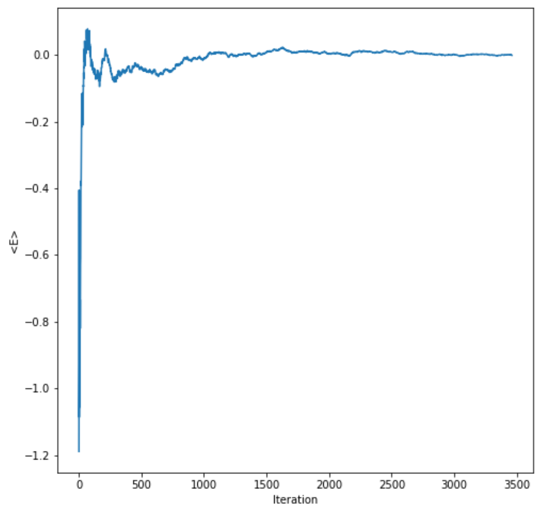

Figure 5: $< E >$- iteration for $\alpha = 1$

By comparing Figure 1 and Figure 3, it is easy to figure out that the acceptance rates ($\alpha = 2$) is smaller than that acceptance rates ($\alpha = 1$)

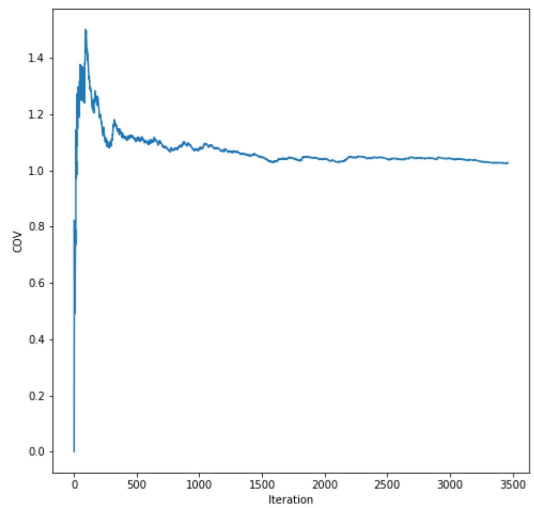

Figure 6: COV- iteration for $\alpha = 1$

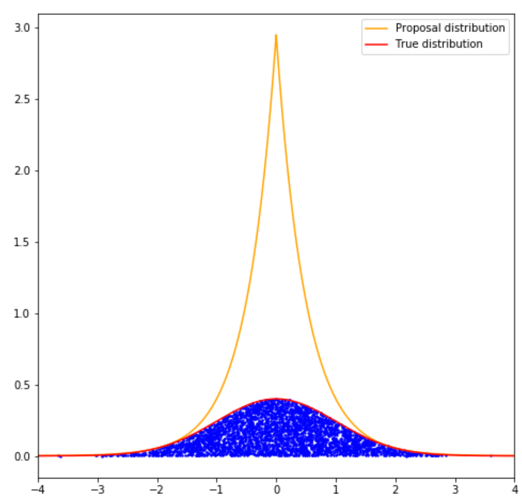

Figure 7: the true pdf and the proposal distribution for $\alpha = 1$

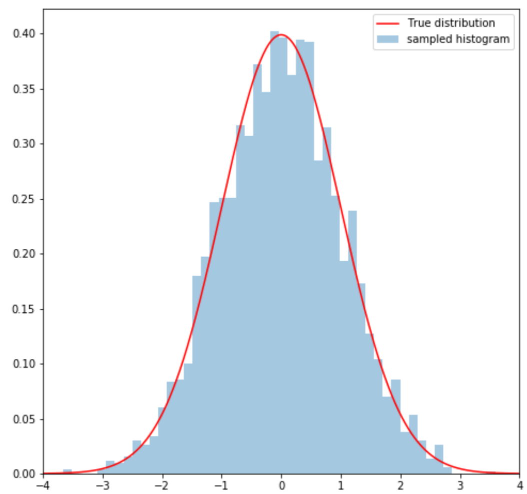

Figure 8: histogram of samples for $\alpha = 1$

2 Independent Metropolis-Hastings

Traditionally in the Metropolis-Hastings algorithm the arbitrary proposal distribution is conditioned on the current state of the chain. Namely, one draws samples from $x' \sim q(x'|x_t)$ where x_t indicates the state of the chain. Consider a proposal distribution that is independent of the chain's current state $q(x')$. When such a distribution is used, this is referred to as the *Independent Metropolis-Hasting algorithm.*

Prove that the Independent Metropolis-Hastings accepts more than the Accept-Reject method when both have identical target $(f(x'))$ and proposal $(g(x'))$ distributions.

Solution:

for Accept-Reject method:

$$\begin{aligned} prob_{AJ} &= \frac{f(x')}{M \cdot g(x')} \\ M &= sup\frac{f(x)}{g(x)} \geq \frac{f(x)}{g(x)} \end{aligned} \tag{5}$$

Since $M \cdot g(x') \geq f(x')$, $M \cdot g(x')$ curve is above $f(x')$ curve, $prob_{AJ} \leq 1$

for Metropolis-Hasting algorithm

$$
\begin{aligned}
&prob_{MH} = \min[1, \frac{\frac{f(x')}{g(x')}}{\frac{f(x_t)}{g(x_t)}}] \\
&if 1 \leq \frac{\frac{f(x')}{g(x')}}{\frac{f(x_t)}{g(x_t)}} \rightarrow prob_{MH} = 1 \geq prob_{AJ} \\
&if 1 \geq \frac{\frac{f(x')}{g(x')}}{\frac{f(x_t)}{g(x_t)}} \rightarrow prob_{MH} = \frac{\frac{f(x')}{g(x')}}{\frac{f(x_t)}{g(x_t)}} \geq \frac{f(x')}{M \cdot g(x')} = prob_{AJ}
\end{aligned} \tag{6}
$$

The prob of Accept-Reject method is smaller than that of Metropolis-Hasting algorithm. Therefore,the Independent Metropolis-Hastings accepts more than the Accept-Reject method when both have identical target ($f(x')$) and proposal ($g(x')$) distributions.

3 Accept-Reject & Metropolis-Hastings

(a) Implement the accept-reject algorithm to calculate the mean of a gamma distribution $\mathcal{G}(4.3, 6.2)$ using a $\mathcal{G}(4, 7)$ candidate. Draw the true density function on top of the sample histogram and plot the convergence.

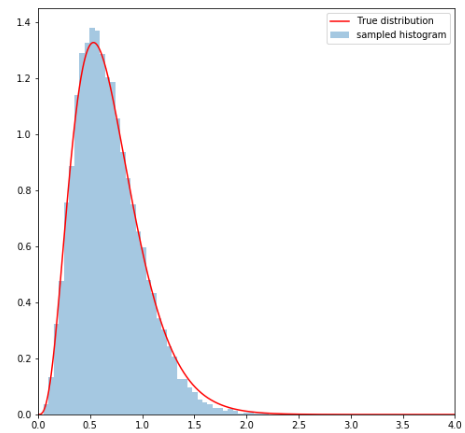

Figure 9: the true pdf and histogram of samples for accept-reject

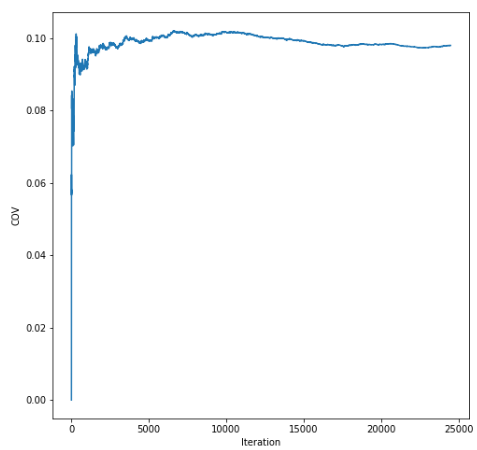

Figure 10: COV for accept-reject

(b) Implement the Metropolis-Hastings algorithm to calculate the mean of a gamma distribution $\mathcal{G}(4.3, 6.2)$ using the following candidate densities:

A gamma $\mathcal{G}(4, 7)$ candidate distribution.

A gamma $\mathcal{G}(5, 6)$ candidate distribution.

For both candidate distributions draw the true and candidate density functions on top of the sampled histogram. Plot the convergence using each candidate distribution on the same axis . How do the means compare?

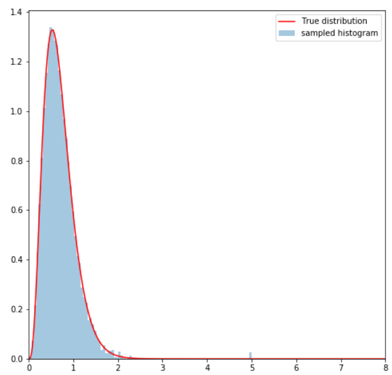

Figure 11: $\mathcal{G}(4, 7)$

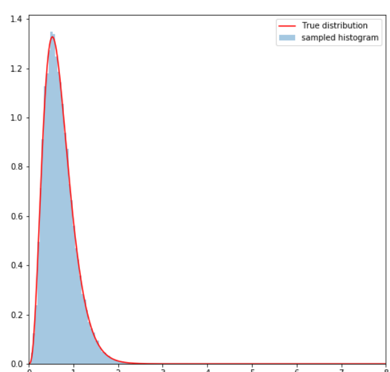

Figure 12: $\mathcal{G}(5, 6)$

$\mathcal{G}(5, 6)$ convergences much quickly than $\mathcal{G}(4, 7)$. And we can see it from COV-iteration and $< E >$-iteration figures.

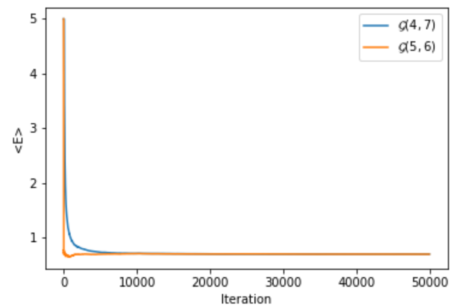

Figure 13: $< E >$ comparation for different proposal functions

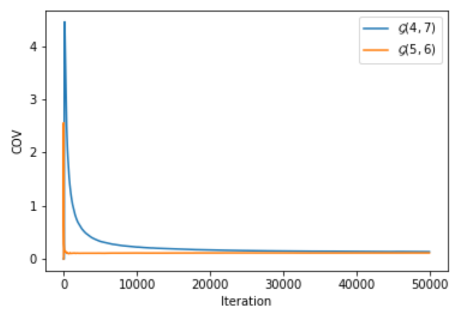

Figure 14: COV comparation for different proposal functions

4 Gibbs & Metropolis-Hastings

Consider sampling from a 2D Gaussian. Suppose $x \sim \mathcal{N}(\mu, \Sigma)$ where $\mu = (1, 1)$ and $\Sigma = (1, -0.5; -0.5, 1)$.

(a) Derive the full conditional $p(x_1|x_2)$ and $p(x_2|x_1)$. Implement the Gibbs algorithm for this case and plot the 1D marginals $p(x_1)$ and $p(x_2)$ as well as (superimposed) the computed histograms.

Derive the full conditional distribution from Bishop-Pattern Recognition and Machine Learning(2.81 and 2.82)

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b) \tag{7}$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \tag{8}$$

Therefore,

$$\begin{aligned} \mu_{x_1|x_2} &= 1 - 0.5(x_2 - 1) \\ \Sigma_{x_1|x_2} &= 0.75 \\ x_1|x_2 &\sim \mathcal{N}(1 - 0.5(x_2 - 1), 0.75) \\ \mu_{x_2|x_1} &= 1 - 0.5(x_1 - 1) \\ \Sigma_{x_2|x_1} &= 0.75 \\ x_2|x_1 &\sim \mathcal{N}(1 - 0.5(x_1 - 1), 0.75) \end{aligned} \tag{9}$$

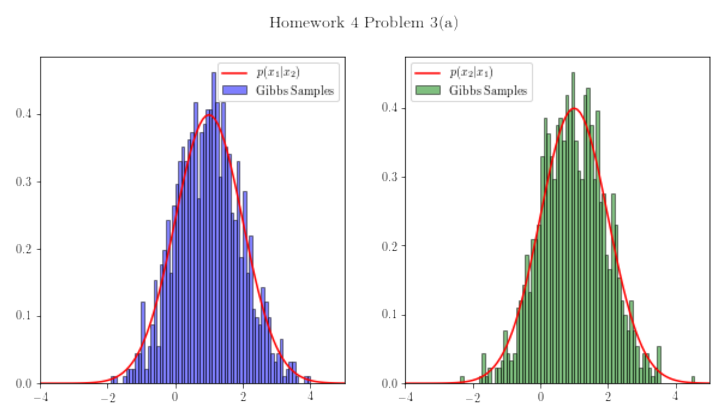

Figure 15:

(b) Let us now consider block-wise Metropolis Hastings. For our proposal distribution, $q(x)$ let us use a normal centered at the previous state/sample of the Markov chain/sampler, i.e: $q(x|x^{(t-1)}) \sim N(x^{(t-1)}, I)$, where I is a 2D identity matrix. Show the 2D target distribution and its sampled approximation.

(c) We now consider component-wise Metropolis Hastings approximation of the same problem. The proposal distribution $q(x)$ is now a univariate Normal

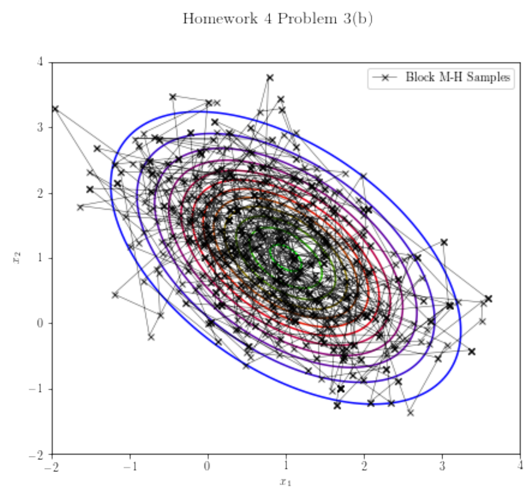

Figure 16:

distribution with unit variance in the direction of the i-th dimension to be sampled. Show the sampled and exact target distribution. Show your results and compare the convergence with that obtained with the block-wise, component-wise Metropolis-Hastings and Gibbs implementation.

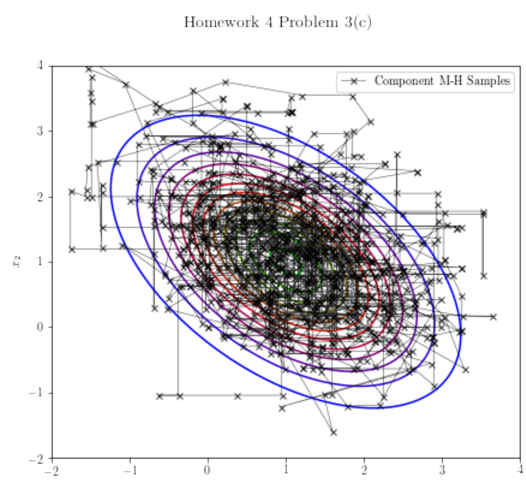

Figure 17:

Compare the convergence with that with that obtained with the block-wise, component-wise Metropolis-Hastings and Gibbs implementation.

From $< E >$ convergence and COV convergence, it is easy to find that the

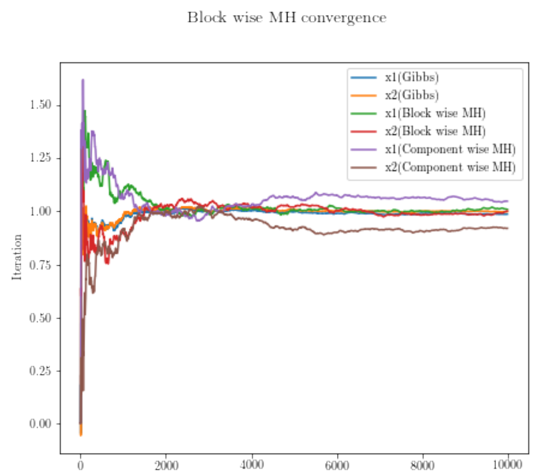

Figure 18: $< E >$ convergence

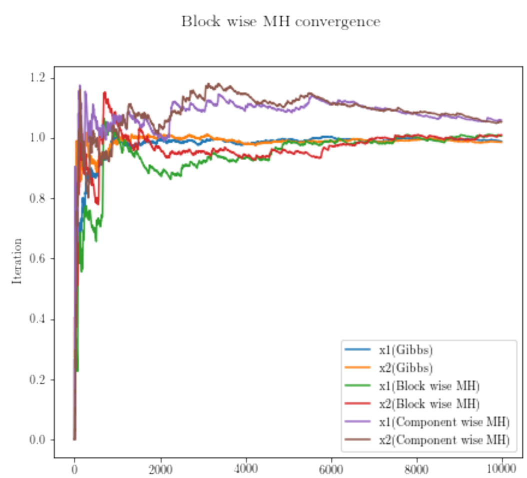

Figure 19: COV convergence

convergence rate: Gibbs is the fastest, Block-wise MH is the second, component-wise is the slowest.

5 Metropolis-Hastings

Consider the braking data of Tukey. It corresponds to breaking distances $y_{i,j}$ of cars driving at speeds x_i. It is thought that a good model for this dataset is quadratic model:

$$y_{i,j} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_{i,j} \tag{10}$$

where $\epsilon_{i,j} \sim N(0, \sigma^2), i = 1, ..., k$ and $j = 1, ..., n_i$ If we assume that $\epsilon_{i,j} \sim N(0, \sigma^2)$ are independent, then the likelihood function is

$$(\frac{1}{\sigma^2})^{N/2} e^{-\frac{1}{2\sigma^2} \sum_{i,j} (y_{i,j} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)} \tag{11}$$

We can view this likelihood as a posterior distribution of $\beta_0, \beta_1, \beta_2, \sigma^2$ and we can sample from it with a Metropolis-Hasting algorithm.

(a) Obtain maximum likelihood estimate for β_0, β_1, β_2, σ^2

$$I(\beta_0, \beta_1, \beta_2, \sigma^2 | y, X) = (\frac{1}{\sigma^2})^{N/2} e^{-\frac{1}{2\sigma^2} \sum_{i,j} (y_{i,j} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)} \tag{12}$$

take the log of I

$$\log I(\beta_0, \beta_1, \beta_2, \sigma^2 | y, X) = (N/2) \log\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{i,j} (y_{i,j} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2) \tag{13}$$

for β_0, β_1, β_2:

$$\begin{aligned} \frac{\partial I}{\partial \beta_0} &= 0; \\ \frac{\partial I}{\partial \beta_1} &= 0; \\ \frac{\partial I}{\partial \beta_2} &= 0; \\ \frac{\partial I}{\partial \sigma} &= 0; \end{aligned} \tag{14}$$

then

$$\begin{aligned} \sum_{i,j} y_{i,j} &= \sum_i n_i (\beta_0 + \beta_1 x_i + \beta_2 x_i^2); \\ \sum_{i,j} y_{i,j} x_i &= \sum_i n_i (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) x_i; \\ \sum_{i,j} y_{i,j} x_i^2 &= \sum_i n_i (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) x_i^2; \\ \sigma^2 &= \frac{1}{N} \sum_{i,j} (y_{i,j} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2) \end{aligned} \tag{15}$$

We call

$$\begin{aligned} Y &= [y_{1,1}, ..., y_{1,n_1}, y_{2,1}, ..., y_{k,1}, ..., y_{k,n_k}] \\ X &= [x_1 I_{(n_1 \times 1)}, ..., x_k I_{(n_k \times 1)}] \end{aligned} \tag{16}$$

The solution of β satisfying those equation has a closed form:

$$\hat{\beta} = ([I, X, X^2]^T [I, X, X^2])^{-1} [I, X, X^2]^T Y \tag{17}$$

for σ^2, put estimated $\hat{\beta}$ into the likelihood expression of σ^2.

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i,j} (y_{i,j} - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2) \tag{18}$$

Use the braking data of Tukey, MLE of beta is $\hat{\beta} = [2.470.910.1]^T$; MLE of sigma is $\sigma^2 = 216.5$.

(b) Use the estimates to select a candidate distribution. Take normal for β_0, β_1, β_2, and inverted Gamma for σ^2.

The MLE estimated $\hat{\beta}$ can be used as the mean parameter of normal proposal density for β because it is unbiased estimator. As for the variance parameter in proposal density, we can rely on its covariance matrix approximation.

$$\mathbb{V}(\beta)|X, \sigma^2) = ([I, X, X^2]^T [I, X, X^2])^{-1} \hat{\sigma}^2 \tag{19}$$

The proposal density for β is then

$$\beta \sim N(\hat{\beta}, \mathbb{V}(\beta|X, \sigma^2)) \tag{20}$$

For the proposal density of parameters σ^2, according to Cochran's theorem:

$$\frac{N\hat{\sigma}^2}{\sigma^2} \sim \mathcal{X}^2_{N-3} = \mathcal{G}(\frac{N-3}{2}, 2) \to \frac{1}{\sigma^2} \sim \mathcal{G}(\frac{N-3}{2}, \frac{2}{N\hat{\sigma}^2}) \tag{21}$$

Therefore, the final proposal density for β, σ is then

$$\begin{aligned} p(\beta, \sigma^2) &= \mathcal{N}(\beta|\hat{\beta}, \mathbb{V}(\beta|X, \sigma^2)) \mathcal{IG}(\sigma^2|\frac{N-3}{2}, \frac{2}{N\hat{\sigma}^2}) \\ &= \mathcal{N}([2.47, 0.91, 0.1], \begin{bmatrix} 206.37 & -27.22 & 0.821 \\ -27.22 & 3.89 & -0.124 \\ 0.821 & -0.124 & 0.0041 \end{bmatrix}) \mathcal{IG}(23.5, 5405) \end{aligned} \tag{22}$$

For student T distribution

$$\mathbb{V}(\beta)|X, \sigma^2) = ([I, X, X^2]^T [I, X, X^2])^{-1} \hat{\sigma}^2 \frac{v}{v-2} \tag{23}$$

This covariance is bigger than the previous one.

$$\begin{aligned} p(\beta, \sigma^2) &= \mathcal{N}(\beta|\hat{\beta}, \mathbb{V}(\beta|X, \sigma^2)) \mathcal{IG}(\sigma^2|\frac{N-3}{2}, \frac{2}{N\hat{\sigma}^2}) \\ &= \mathcal{N}([2.47, 0.91, 0.1], \begin{bmatrix} 412.74 & -54.44 & 1.642 \\ -54.44 & 7.78 & -0.248 \\ 1.642 & -0.248 & 0.0082 \end{bmatrix}) \mathcal{IG}(23.5, 5405) \end{aligned} \tag{24}$$

(c) Make histogram of the posterior distributions of the parameters. Monitor convergence.

Robustness considerations could lead to using an error distribution with heavier tails. If we assume that $\epsilon_{i,j} \sim Gamma(0, \sigma^2)$ independent, then the likelihood function is

$$(\frac{1}{\sigma^2})^{N/2} \prod_{i,j} (1 + \frac{1}{v} \frac{(y_{i,j} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2}{\sigma^2})^{(v+1)/2} \tag{25}$$

where v is the degrees of freedom. For $v = 4$, use Metropolis-Hastings to sample β_0, β_1, β_2, σ^2 from the posterior distribution. Use either normal or Γ candidates for β_0, β_1, β_2 and inverted Gamma or half-Γ for σ^2.

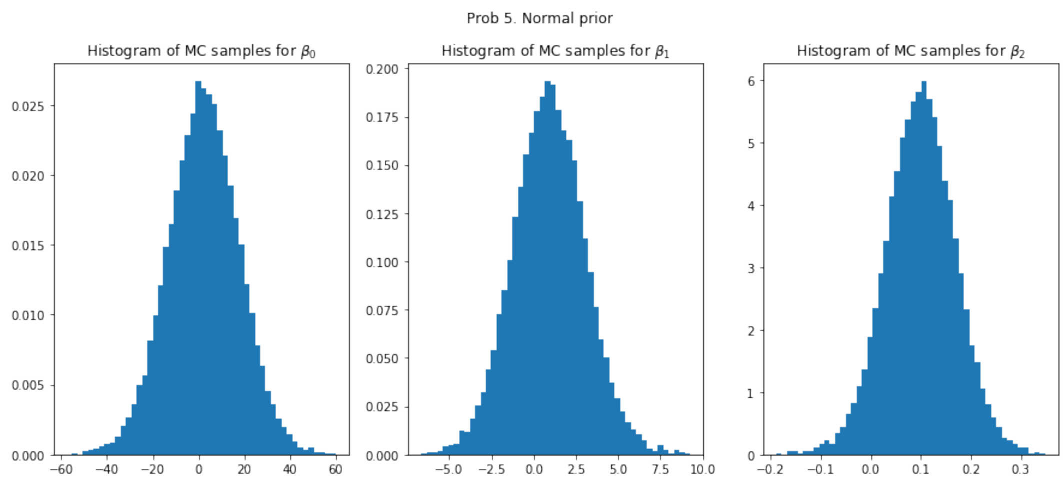

Figure 20: β Normal distribution

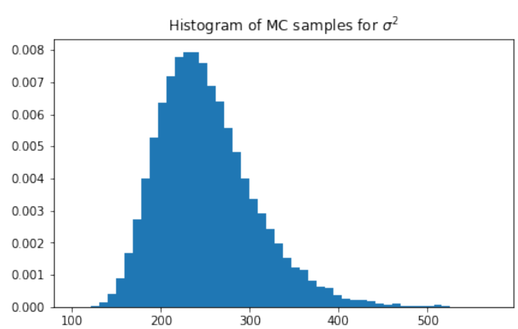

Figure 21: σ Normal distribution

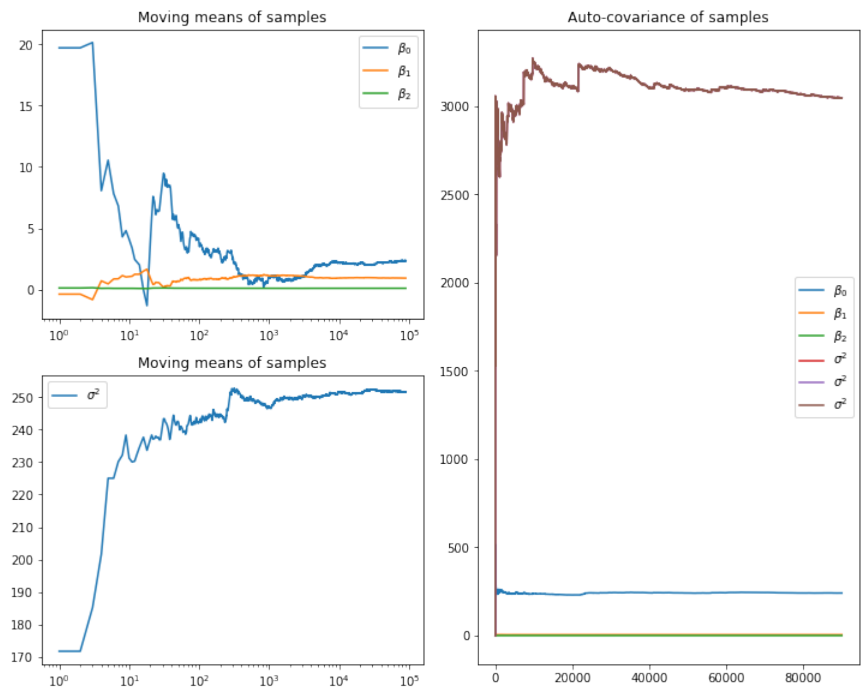

Figure 22: Normal distribution

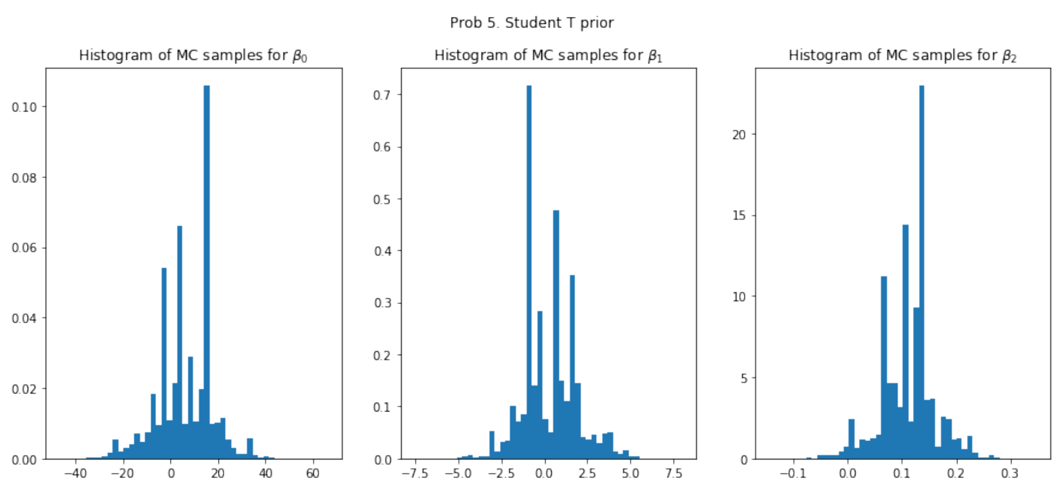

Figure 23: β Student T distribution

Normal distribution candidate is better than student T distribution candidate.

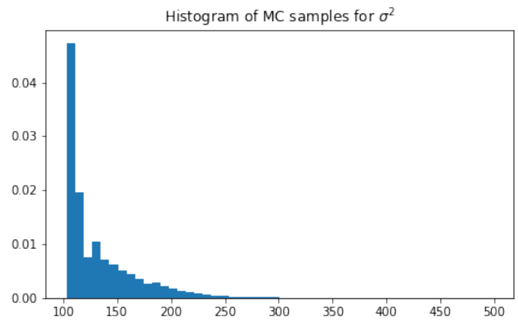

Figure 24: σ Student T distribution

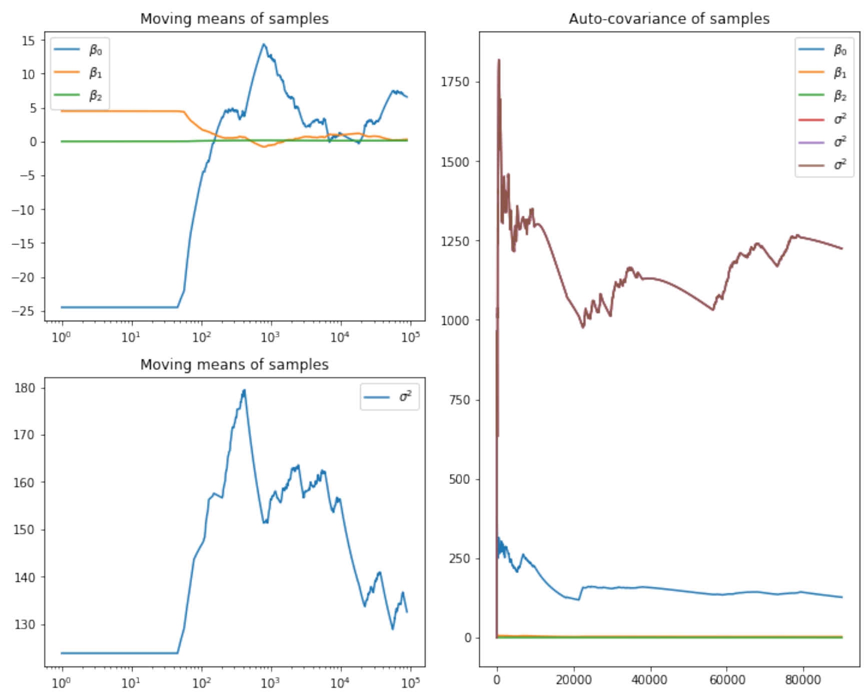

Figure 25: Student T distribution