HW1 — CSE8803: IUQ

Ashish Dhiman — ashish.dhiman9@gatech.edu

May 7, 2023

Contents

1 Question 1: Monte Carlo Markov Chain

1.1 Task 1: Definitions and Proofs

Key items in context of MCMC

- **Proposal Distribution**: The proposal distribution determines the transition probabilities between the states of the Markov chain. The proposal distribution in MCMC methods plays a crucial role in determining how the Markov chain transitions between 2 states. It defines the probabilities of moving from the current state to a proposed new state, and is selected to facilitate efficient sampling across the state space.

- **Acceptance Ratio**: Acceptance Ratio or AR is defined as the total number of steps (or samples) that have been accepted from all the samples generated. It provides us a rough tradeoff between exploration and convergence of chain. If AR is very high, we might arrive at a Target distribution which is not sufficiently explored, while a low AR would imply that we need long chains to arrive at sufficient number of samples.

- **Acceptance Probability**: In MCMC, a new state is proposed by a transition kernel from the current state of the chain, and then the new state is either accepted or rejected based on an acceptance probability.

$$A(\theta^{'}, \theta) = \min(1, \frac{\pi(\theta^{'})p(\theta|\theta^{'})}{\pi(\theta)p(\theta^{'}|\theta)})$$

Proof of detailed balance equation for MALA

In case of MALA, we use gradient in the proposal distribution to take more informed steps compared to the random walk in Metropolis Hastings.

$$p(\theta'|\theta) \propto \exp(|\theta' - \theta - \sigma^2 \nabla_\theta \log \pi(\theta)/2|^2$$

Acceptance Probability is

$$A(\theta'|\theta) = min(1, \frac{\pi(\theta')g(\theta|\theta')}{\pi(\theta)g(\theta'|\theta)})$$

We can verify detailed balance equation using:

$$\pi(\theta) * p(\theta'|\theta) = \pi(\theta) * (g(\theta'|\theta) * A(\theta'|\theta))$$

Here p is transition probability and g is proposal

$$\implies \pi(\theta) * p(\theta'|\theta) = \pi(\theta) * (g(\theta'|\theta) * A(\theta'|\theta))$$

Hence on re-arranging and assuming detailed balance equation is true we get

$$\pi(\theta) * g(\theta'|\theta) * A(\theta'|\theta) = \pi(\theta') * g(\theta|\theta') * A(\theta|\theta')$$

$$\implies \frac{A(\theta'|\theta)}{A(\theta|\theta')} = \frac{\pi(\theta') * g(\theta|\theta')}{\pi(\theta) * g(\theta'|\theta)} = r(\text{ Let})$$

Now for every $A(\theta'|\theta) = min(1, r)$ above equation holds

Because if $r \leq 1 \implies A(\theta|\theta')$ and vice versa

(1)

Thus detailed balance equation is satisfied for MALA as long as Acceptance probablity is chosen as defined above.

Proof of detailed balance equation for HMC

Proposal in HMC works in two steps, first at location x_0 draw a random momentum vector $\omega \sim e^{-V(\omega)}$ where V is Kinetic energy and is given as $p(\omega) \propto \exp(-\frac{1}{2}\omega^T M^{-1}\omega)$. The next step, which is deterministic is to apply Hamiltonian dynamics until time T at location x_0 and velocity ω with Hamiltonian energy $H(x_0, \omega) = \log \pi(x) + V(\omega)$.

Acceptance Probability is then defined as:

$$A(x_0, \omega, x_T, \omega') = min(1, \frac{-H(x_T, \omega')}{-H(x_0, \omega)})$$

For verifying detailed balance equation we want to show :

$$\pi(x)p(y|x) = \pi(y)p(x|y)$$

where *pi* is the stationary distribution while, $p(x|y)$ is transition probability from x to y.

Now for Hamiltonian dynamics it holds that if we start the dynamics with initial location x_T and momentum ω' , after time T we will come back to x_0 and momentum ω.

$$\begin{aligned}
\pi(x)p(y|x) &= \pi(x) * p(\omega) \\
&= \pi(x) * \frac{1}{K} * \exp(-\frac{1}{2}\omega^T M^{-1}\omega) \\
&\text{Also we have potential energy } U(x) = -\log \pi(x) \\
\implies \pi(x)p(y|x) &= \exp(-U(x)) * \frac{1}{K} * \exp(-\frac{1}{2}\omega^T M^{-1}\omega) \\
&= \frac{1}{K} * \exp(-[U(x) + \frac{1}{2}\omega^T M^{-1}\omega]) \\
&= \frac{1}{K} * \exp(-H(x, \omega)) \\
&= \frac{1}{K} * \exp(-H(y, \omega')) \quad \text{(Energy conserved)} \\
&= \frac{1}{K} * \exp(-[U(y) + \frac{1}{2}\omega'^T M^{-1}\omega']) \quad \text{(Energy conserved)} \\
&= \pi(y) * p(\omega') \\
&= \pi(y) * p(-\omega') \quad (\omega \text{ is from Gaussian}) \\
&= \pi(y) * p(y|x)
\end{aligned} \tag{2}$$

1.2 Task 2: MCMC/MALA/HMC for $\mathcal{N}(\mu, C)$

1.2.1 Density of samples

Density with Metropolis Hastings

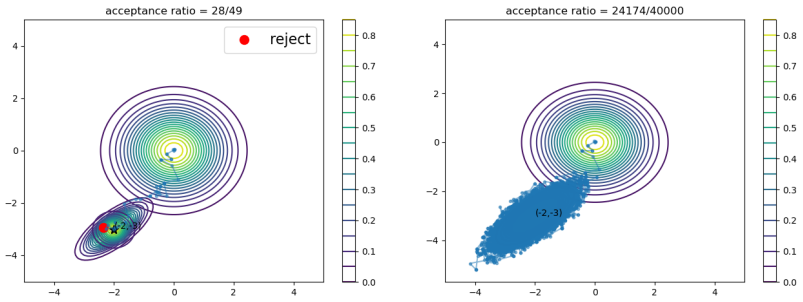

Figure 1: Convergence to Target distribution - MH

Density with MALA

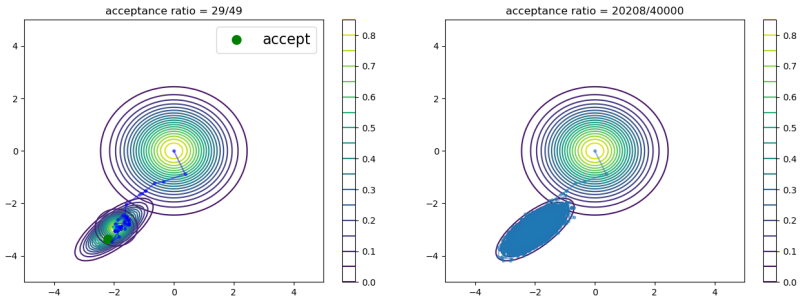

Figure 2: Convergence to Target distribution - MALA

Density with Hamiltonian

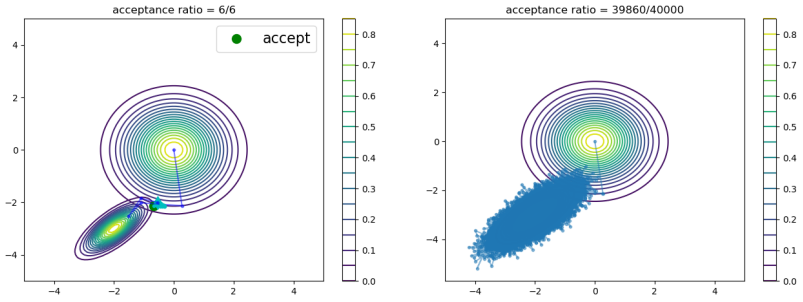

Figure 3: Convergence to Target distribution - HMC

I can notice the following observations:

- HMC has highest AR, and quickest convergence
- While MALA has similar AR wrt MH, MALA gives more localised convergence, while MH is more dispersed.

1.2.2 Trace and Auto Corelation time

Trace for Metropolis Hastings

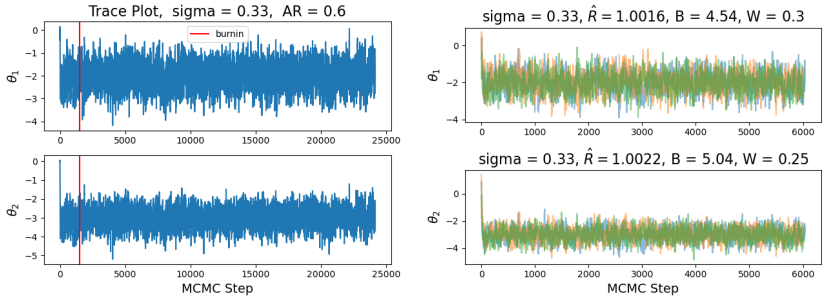

Figure 4: Trace for single chain (left), multiple chains (right)

Trace for MALA

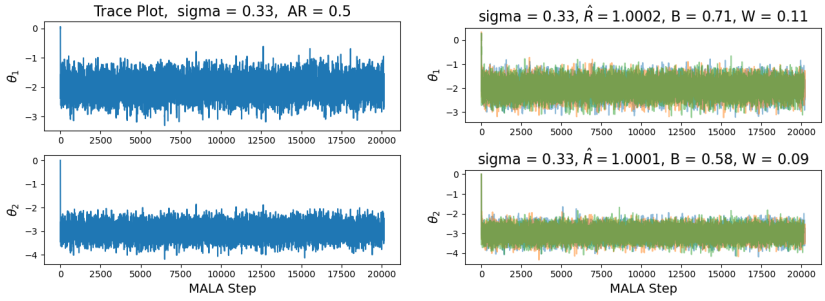

Figure 5: Trace for single chain (left), multiple chains (right)

Trace for Hamiltonian

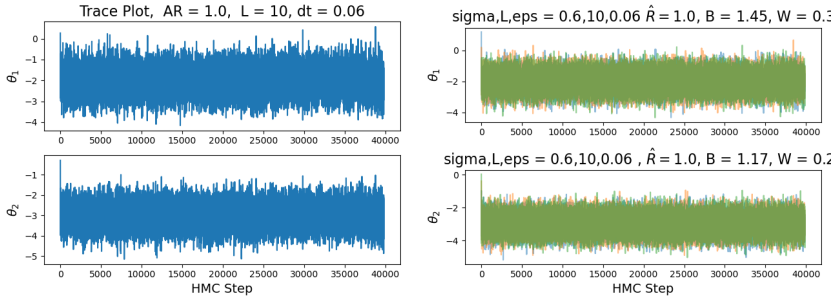

Figure 6: Trace for single chain (left), multiple chains (right)

We can see convergence of $\theta^T = (-2, -3)$, i.e. to Target distribution for all the methods.

ACT for Metropolis Hastings

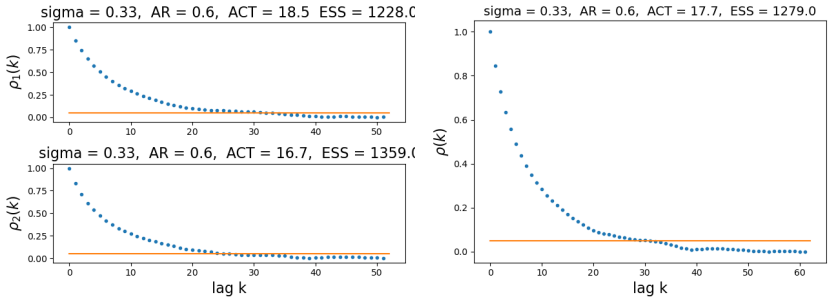

Figure 7: ACT: individual dimension(left), combined (right)

ACT for MALA

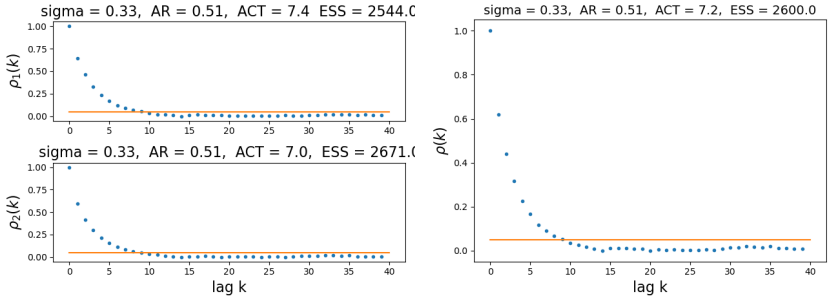

Figure 8: ACT: individual dimension(left), combined (right)

ACT for Hamiltonian

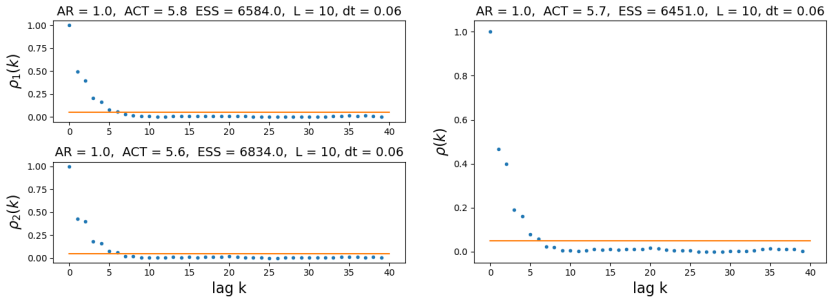

Figure 9: ACT: individual dimension(left), combined (right)

I can notice the following observations:

- As expected HMC has lowest ACT, while MALA has 2nd lowest.
- For all the methods, there is negligible difference b/w ACT -combined and for individual dimensions. This might be because the C_x and C_y of target distribution are almost identical.

1.2.3 ESS, AR, $\hat{R}$ comparison table

For calculation of AR,ACT,ESS I have used a MCMC chain of length 40,000 for all the three methods. The sigma for MH and MALA was chosen as 0.33, while the step length in HMC was chosen as 0.6 (L=10, epsilon = 0.6), with a burn in of 1500.

For calculation of convergence diagnostic $\hat{R}$, the sigma and step length parameters are unchanged, however we draw m=10 chains of length 10000 each.

	AR	**ACT**	**ESS**	$\hat{R}(\theta_1)$	$\hat{R}(\theta_2)$
MH	0.6	17.72	1279	1.0015924	1.0022
MALA	0.5038	7.19397	2600.51	1.0001512	1.00015
HMC	0.99633	5.71378	6451.07	1.0000497	1.00005

Table 1: Comparison of MH, MALA and HMC

I can notice the following observations:

- While HMC takes the longest to run, it provides the best convergence statistic $\hat{R}$, with closest values to 1.
- HMC also has the lowest ACT and consequently the best sample size or ESS as well.
- While vanilla MH and MALA have almost identical AR, however the ACT for MALA is lower and thus it provides better ESS.
- MALA shows better convergence than MH.
- All the 3 methods do provide sufficiently good convergence tho, as the $\hat{R}$ is close to 1 for all the 3 methods.

2 Question 2: Variational Inference

2.1 Task 1: Proofs

2.1.1 KL divergence is nonnegative and nonsymmetric

We know KL-divergence for two densities p,q is given by:

$$KL(p(x)||q(x)) = E_{p(x)}[log\frac{p(x)}{q(x)}]$$

Hence we can simplify this as:

$$\begin{aligned}
KL(p(x)||q(x)) &= E_{p(x)}[log\frac{p(x)}{q(x)}] \\
&\text{By definition of expectation} \\
&= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\
&\text{Let } I \text{ be the set of all } x \text{ for which } p(x) \text{ is non-zero.} \\
&= \sum_{x\in\mathcal{I}} p(x)\log\frac{p(x)}{q(x)} \\
&\text{For all } x>0, \text{ it holds that } \ln x \le x-1 \implies \log\frac{1}{x} \ge 1-x \\
\implies KL(p||q) &= \sum_{x\in\mathcal{I}} p(x)\log\frac{p(x)}{q(x)} >= \sum_{x\in\mathcal{I}} p(x)\log(1-\frac{q(x)}{p(x)}) \\
&= \sum_{x\in\mathcal{I}} p(x) - \sum_{x\in\mathcal{I}} q(x) \\
&\text{But } \sum_{x\in\mathcal{I}} p(x) = 1 \\
&\text{But since I is chosen basis p hence } \sum_{x\in\mathcal{I}} q(x) <= 1 \\
&\implies \sum_{x\in\mathcal{I}} p(x) - \sum_{x\in\mathcal{I}} q(x) \ge 0 \\
&\implies KL(p(x)||q(x)) \ge 0
\end{aligned} \tag{3}$$

Hence KL divergence is always **non-negative**. Also we can show:

$$\begin{aligned}
KL(p(x)||q(x)) &= E_{p(x)}[log\frac{p(x)}{q(x)}] \\
&= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\
&\neq \sum_{x\in\mathcal{X}} q(x)\log\frac{q(x)}{p(x)} \\
&\neq KL(q(x)||p(x))
\end{aligned} \tag{4}$$

Hence KL divergence is generally **non-symmetric**, except for the case:

$$KL(p|q) = KL(q|p) = 0 \text{ iff } p = q, \forall x$$

2.1.2 ELBO is a lower bound of the log marginal likelihood/evidence

$$
\begin{aligned}
KL(p(x)||q(x)) &= E_{q(z)}[\log \frac{q(z)}{p(z|x)}] \\
&= E_{q(z)}[\log \frac{q(z)p(x)}{p(z,x)}] \\
&= E_{q(z)}[\log \frac{q(z)}{p(z,x)}] + E_{q(z)}[\log p(x)] \\
&= E_{q(z)}[\log \frac{q(z)}{p(z,x)}] + \log p(x) \quad \text{(p is constatnt wrt Expectation)} \\
&= E_{q(z)}[\log q(z)] - E_{q(z)}[p(z,x)] + \log p(x) \qquad (5) \\
&\text{Now by def. of ELBO} \\
KL(p(x)||q(x)) &= -ELBO(q) + \log p(x) \\
\implies ELBO(q) &= \log p(x) - KL(p(x)||q(x)) \\
&\text{But as proved earlier } KL \geq 0 \\
&\text{also } p(x) \in [0,1] \implies \log p(x) \leq 0 \\
&\text{Hence we can say:} \\
ELBO(q) &= \log p(x) - positive \\
\implies ELBO(q) &\leq \log p(x)
\end{aligned}
$$

Hence we prove that ELBO is lower bound for Evidence or $\log p(x)$

2.2 Task 2: Bayesian Regression

2.2.1 Regression of $log(price) \sim X_2, X_3$

The point estimates of Regression coefficients in this case are given below and serve as a comparision.

```
Learned parameters for x2,x3:
weight [[-0.0013261  -0.00021149]]
bias [3.786944]
```

The R^2 for the above modes is very low at 0.6 and thus linear model is not the best fit for the data. The regression is a 2d plane in the case of these point estimates:

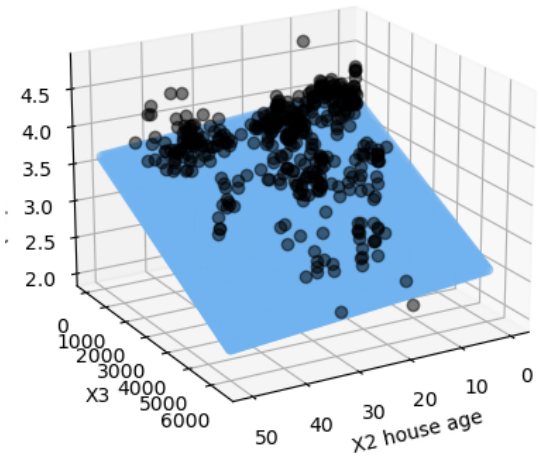

Figure 10: Regression plane for point estimate of parameters

The posterior density of regression parameters for SVI with diagonal normal for one seed is given below:

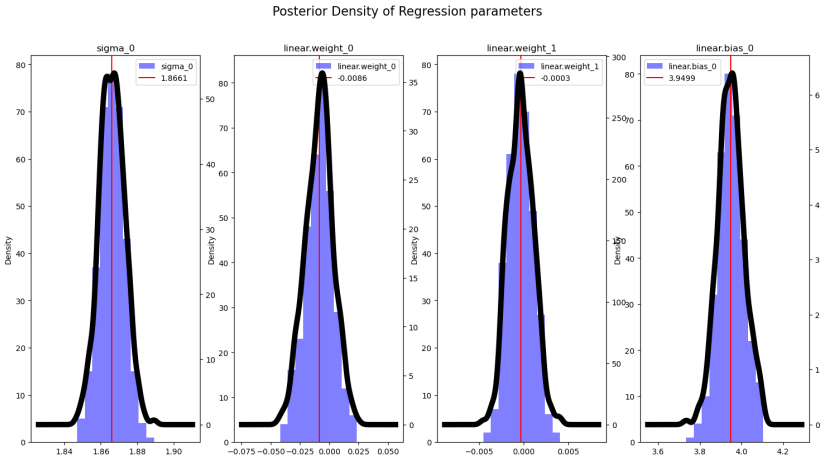

Figure 11: Posterior estimate of regression params (SVI-diagonal)

We see that the maximum likelihood estimate differs slightly from the point estimates.

The posterior density of regression parameters with SVI (Diagonal and Multi variate) and HMC

The priors on weights and data noise is given below:

$$w \sim \mathcal{N}(-0.001, 0.25) \quad bias \sim \mathcal{N}(4.0, 0.25) \quad \sigma^2 \sim \mathcal{N}(0.0, 1.0)$$

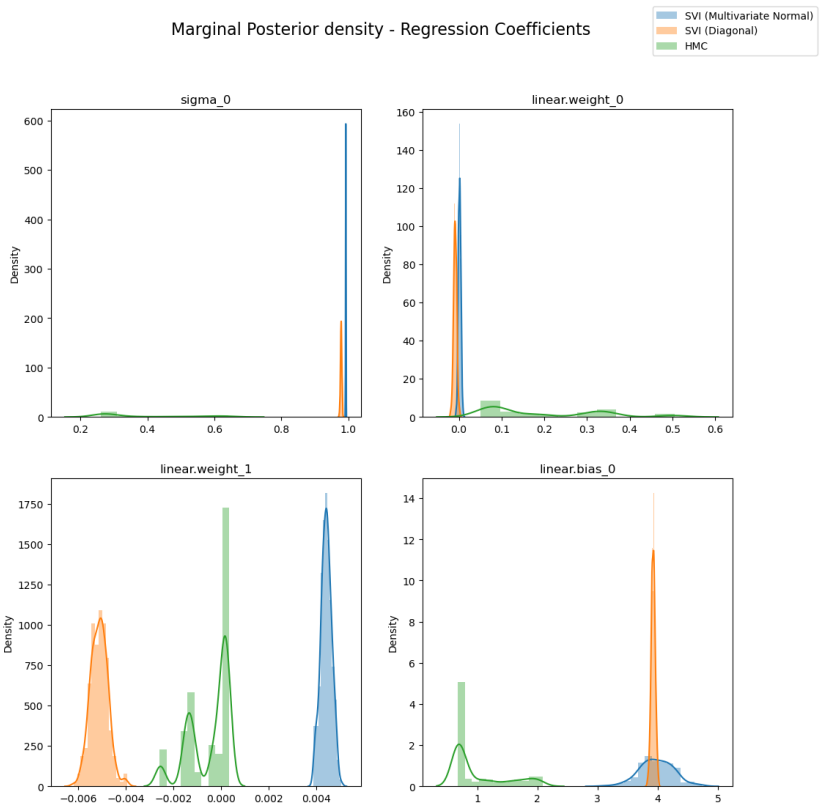

Figure 12: Posterior estimate of regression params (all methods)

As we can see that for this particular setting the density estimates of all the three methods have great divergence, which is probably due to the original observation that Linear model is not the best choice for this dataset.

I do see ELBO loss decreasing with increasing number of iterations:

```
Optimising ELBO loss
[iteration 0001] loss: 17175.9852
[iteration 0301] loss: 7.0548
[iteration 0601] loss: 61.6511
[iteration 0901] loss: 63.8351
[iteration 1201] loss: 43.4407
[iteration 1501] loss: 1.6850
[iteration 1801] loss: 4.7171
[iteration 2101] loss: 2.0058
[iteration 2401] loss: 9.2491
[iteration 2701] loss: 2.4573
[iteration 3001] loss: 28.0266
[iteration 3301] loss: 2.2259
[iteration 3601] loss: 2.0369
[iteration 3901] loss: 1.7892
[iteration 4201] loss: 13.5306
[iteration 4501] loss: 3.4967
[iteration 4801] loss: 3.6425
[iteration 5101] loss: 5.2875
```

The predictive distribution of output

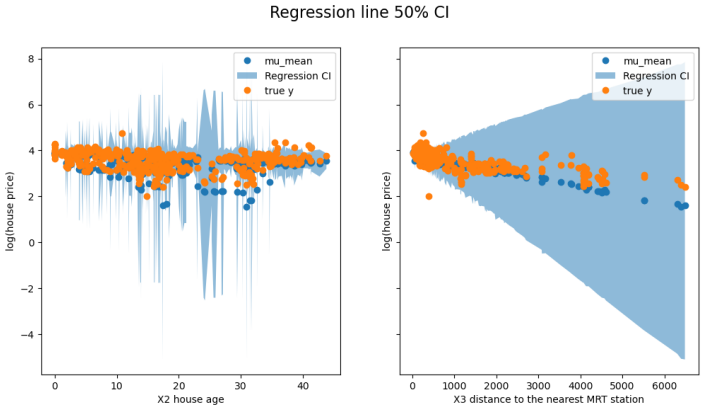

Figure 13: Predictive distribution of mean output

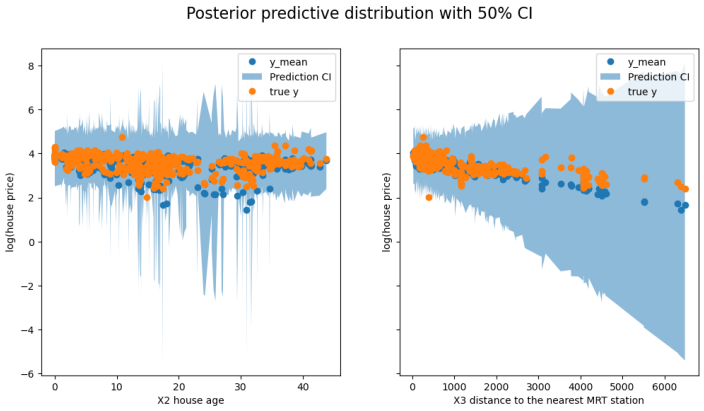

Figure 14: Predictive distribution of obs

As in the case of tutorial though, I observe that SVI Multivariate provides more dispersed posterior estimates, while SVI-diagonal has independent features. This behaviour is apparent in the image below:

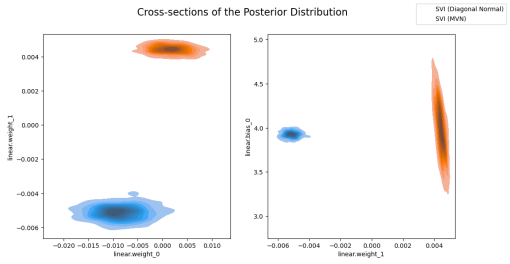

Figure 15: SVI-diagonal vs SVI Multivariate

2.2.2 Regression of $log(price) \sim X_1, X_2, \ldots X_6$

The point estimates of Regression estimates in this case are given below:

```
Learned parameters:
weight [[ 6.8186955e-03 -6.3773077e-03 -1.8886816e-04  3.3255335e
    -02
  -5.7910925e-01  3.9249312e-02]]
bias [-0.2893918]
```

Note these point estimates are arrived at using Adam optimiser and are different from exact OLS solution arrived using Normal equation.

The R^2 for the above modes is also very low at 0.68 and the X1 column (date) does not make much sense in it's current form. The fir for different variables is given below:

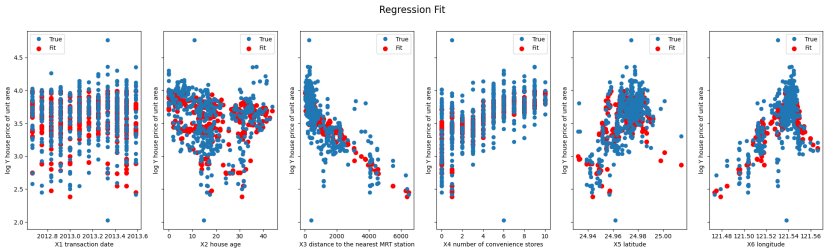

Figure 16: Regression fit for point estimate of parameters

The posterior density of regression parameters for SVI with diagonal normal for one seed is given below:

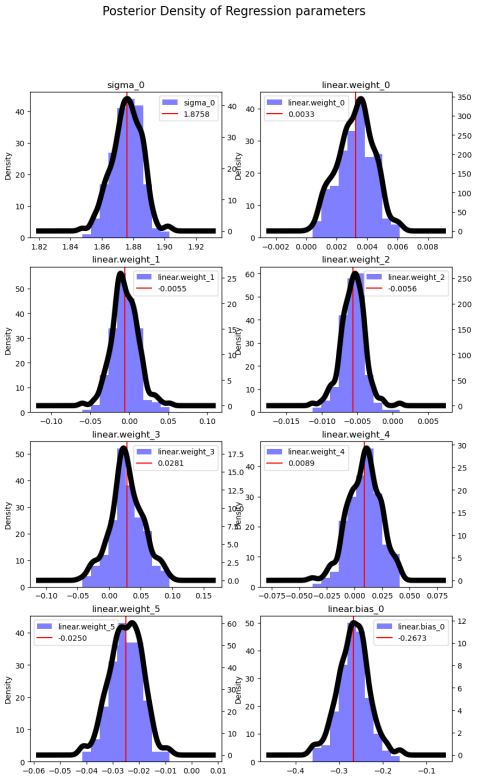

Figure 17: Posterior estimate of regression params (SVI-diagonal)

Again the maximum likelihood estimate differs slightly from the point estimates.

With same priors as before on weights and data noise, we get the following Posterior for Regression params:

$$w \sim \mathcal{N}(-0.001, 0.25) \quad bias \sim \mathcal{N}(4.0, 0.25) \quad \sigma^2 \sim \mathcal{N}(0.0, 1.0)$$

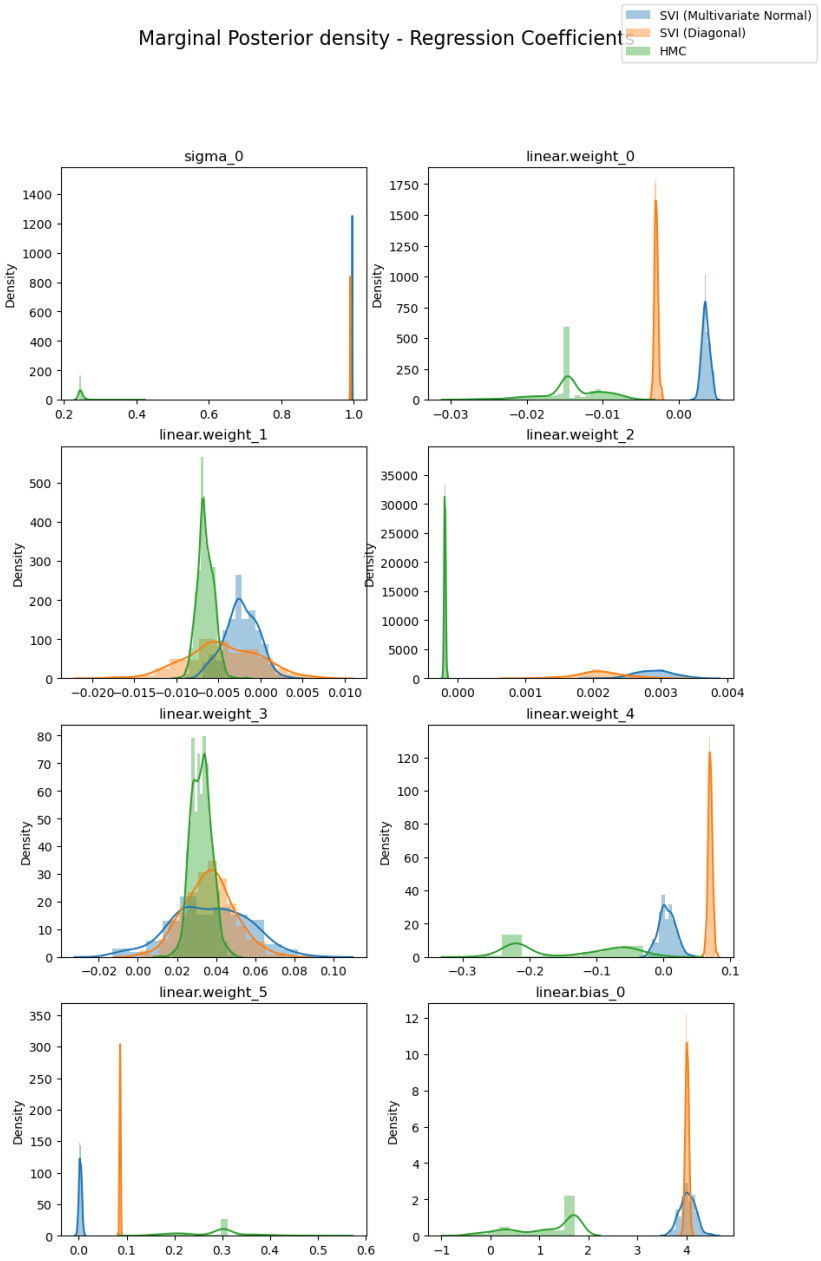

Figure 18: Posterior estimate of regression params (all methods)

The predictive distribution of output

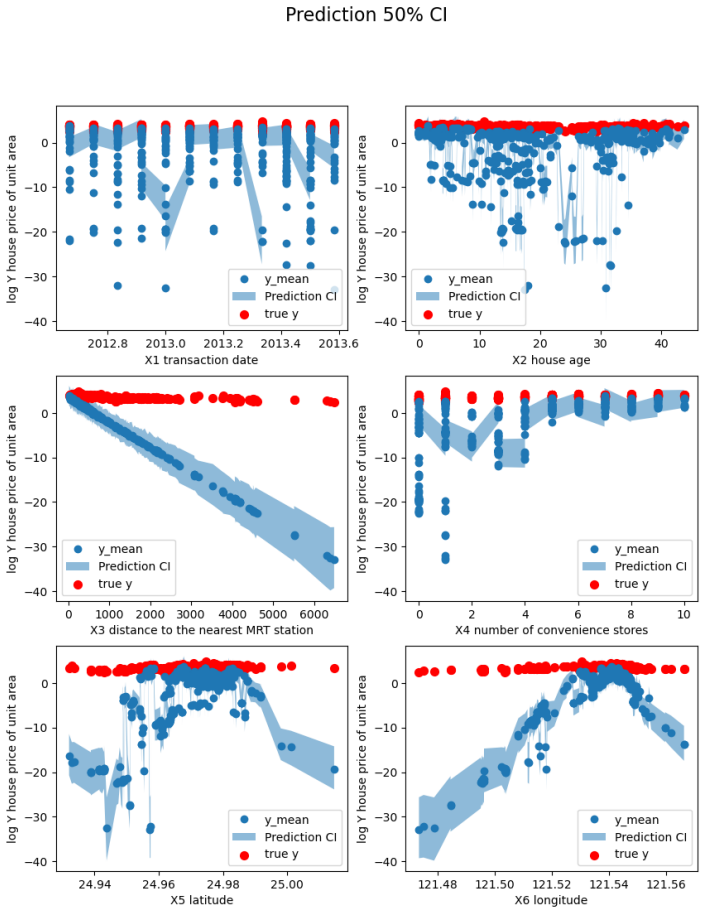

Figure 19: Predictive distribution of mean output

The above prediction densities do not seem very accurate and we need to adjust the parameters of priors for better suited posterior estimates. Having said that the prediction CI seems relatively more accurate wherever data is dense.

Similarly to the earlier case here too SVI Multivariate provides more dispersed posterior estimates, while SVI-diagonal has independent features. This behaviour is apparent in the image below:

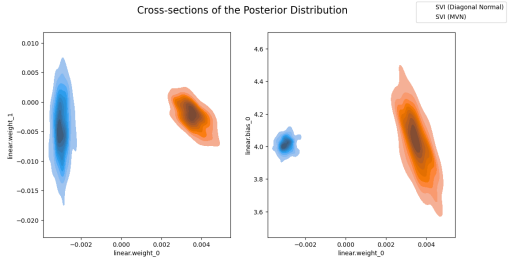

Figure 20: SVI-diagonal vs SVI Multivariate

3 References:

1. https://statproofbook.github.io/P/gibbs-ineq.html
2. https://github.com/pyro-ppl/pyro/blob/dev/tutorial/source/bayesian_regression.ipynb
3. https://github.com/pyro-ppl/pyro/blob/dev/tutorial/source/bayesian_regression_ii.ipynb
4. http://faculty.washington.edu/yenchic/19A_stat535/Lec9_HMC.pdf