HOMEWORK 4

Handed out: Wednesday, Oct. 4 2017 Due: Monday, Oct. 23 midnight

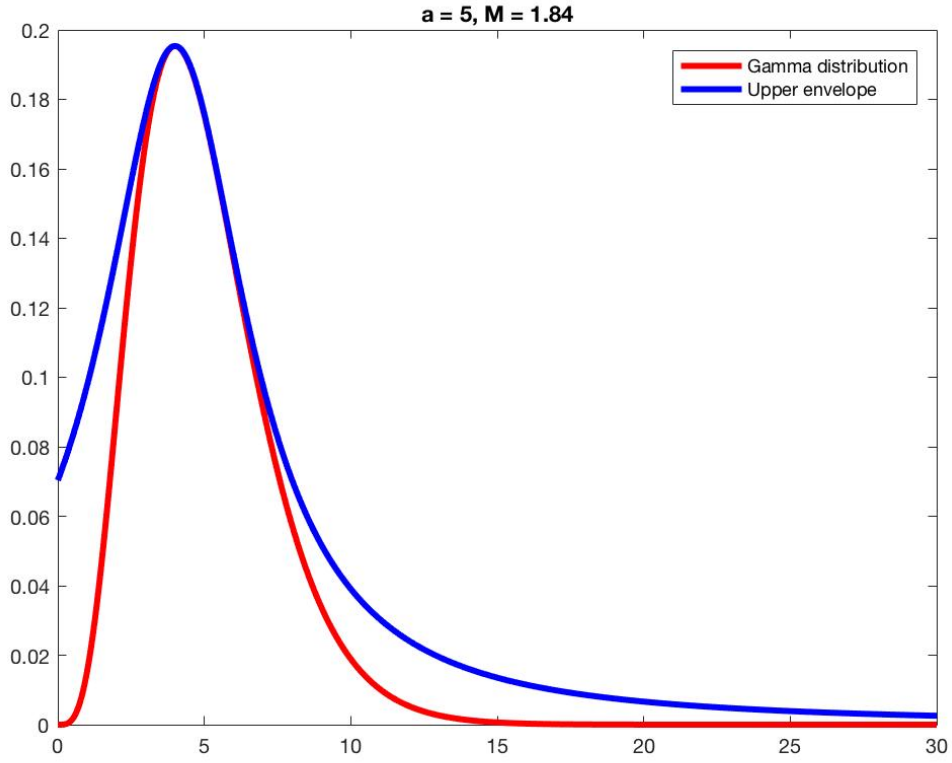
Solution 1. We wish to same from $\mathcal{G}(x|a,\lambda)$ whose density is $f(x) = \frac{1}{\Gamma(a)} x^{a-1}\lambda^a e^{-\lambda x}, a > 1$, with a proposal Cauchy $h(x|b,c) = \frac{\sqrt(c)/\pi}{1+c(x-b)^2}, b = a-1, c = \frac{1}{2a-1}$. We have to find the constant M such that $Mh(x)$ is nowhere less than $f(x)$. The optimal (smallest) one can be derived as follows:

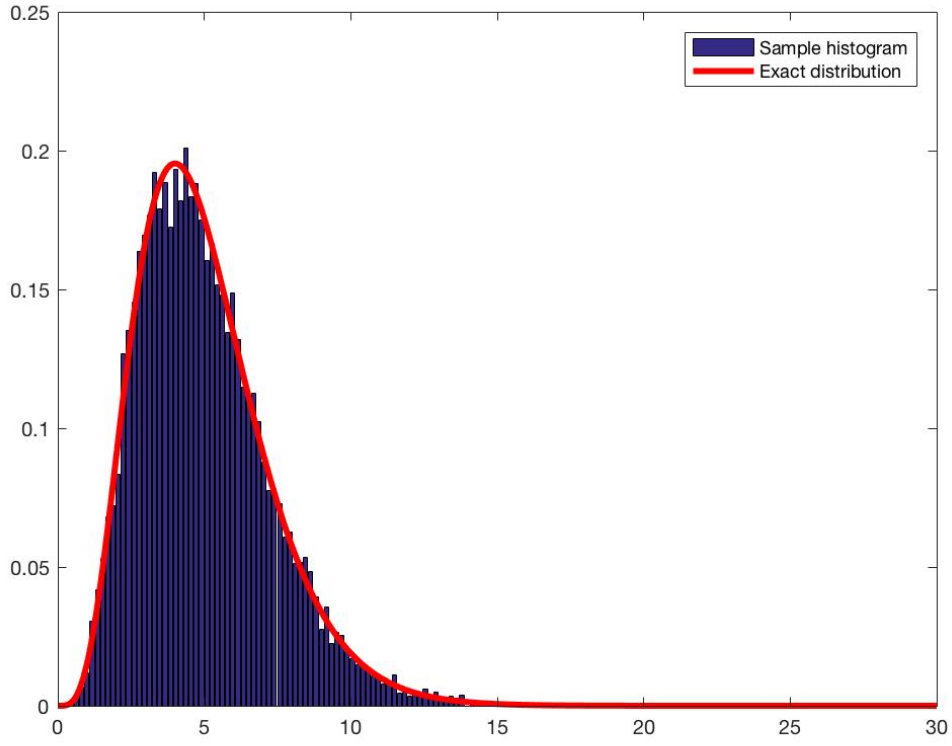
$$f(x) = \frac{1}{\Gamma(a)} x^{a-1}\lambda^a e^{-\lambda x} \leq \frac{1}{\Gamma(a)} e^{(-(a-1))}(a-1)^{a-1}\pi\sqrt(2a-1)h(x|b,c) \tag{1}$$

Therefore $M = \frac{1}{\Gamma(a)} e^{(-(a-1))}(a-1)^{a-1}\pi\sqrt(2a-1)$. The algorithm is detailed in Algorithm Q1 while the two figures appear later.

Q1 Reject sampling

1: Generate a sample $x* \sim h$.
2: Draw $u \sim U(0,1)$.
3: **if** $Mu < f(x*)/h(x*)$ **then**
4: Accept the candidate.
5: **else**
6: Reject the sample.
7: **end if**


a = 5, M = 1.84
Gamma distribution
Upper envelope
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
5
10
15
20
25
30

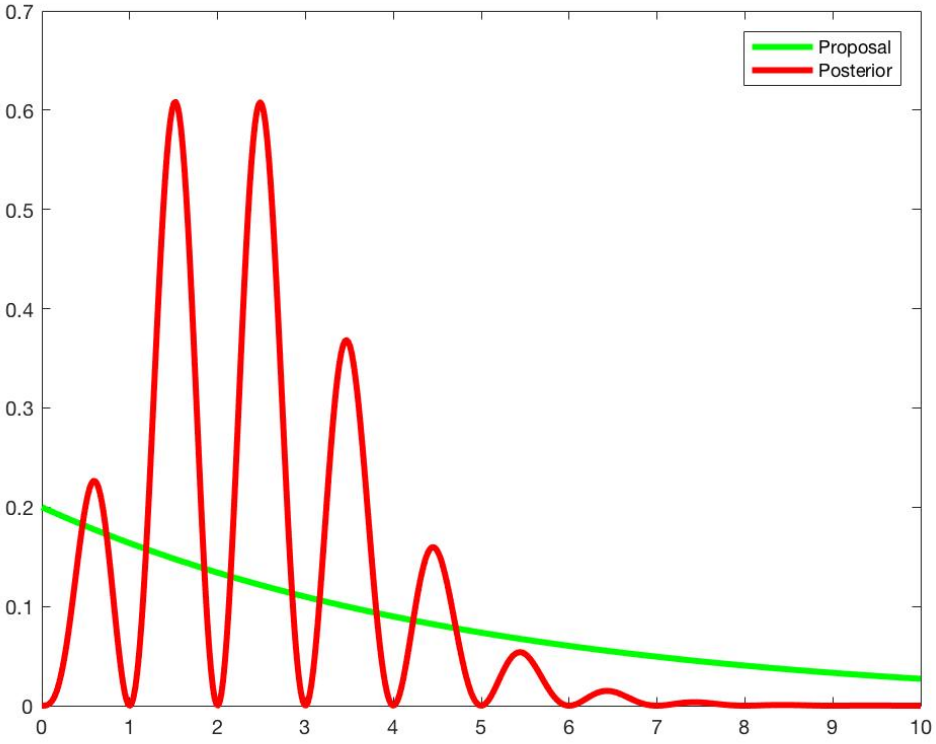

Sample histogram
Exact distribution
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30

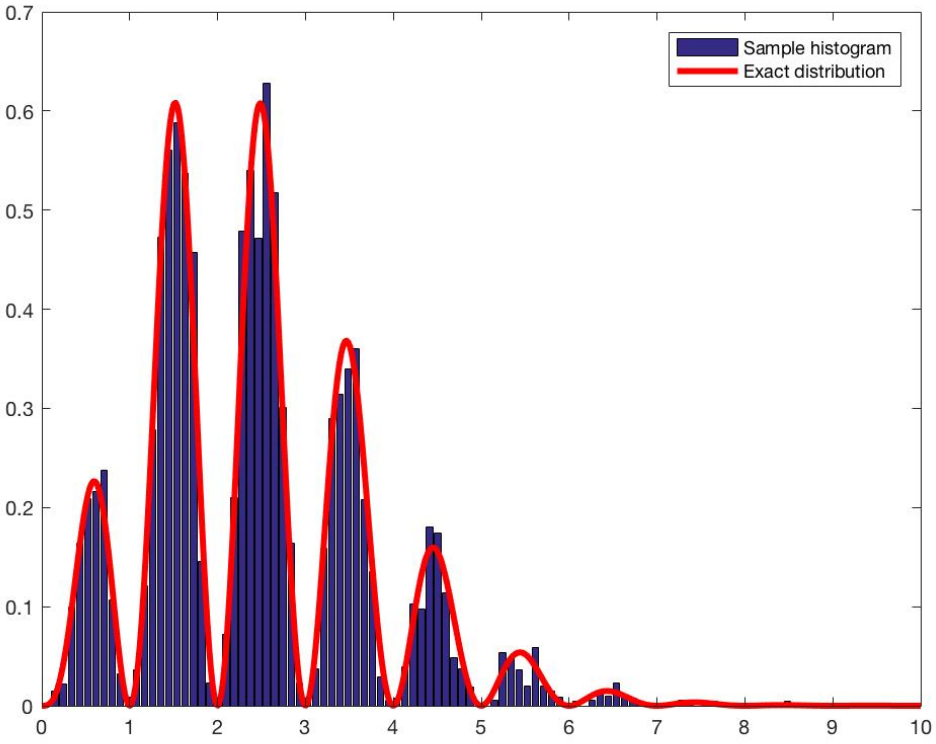
Solution 2. We wish to sample from $p(a|y) \propto p(y|a)p(a)$ with an independent Metropolis sampler. The algorithmic procedure is detailed in Algorithm Q2. The independent sampler is very similar to the reject sampler in Q1, but it has no requirement of knowing the optimal constant M, which can be frequently hard to obtain. The drawback is unlike reject sampling, the generated samples can be correlated.

Q2 Independent Metropolis sampling

Generate a sample $a^{(1)} \sim q$. Set $i = 1$.
Compute its posterior $p^{(1)} = p(a^{(1)}|y)$ and proposal density $q(a^{(1)})$.
while $i < M$ **do**
  Generate a candidate sample $a* \sim p$.
  Compute its posterior and proposal density $p*, q*$.
  Draw $u \sim U[0, 1]$.
  if $u < (p*/p^{(i)})/(q^{(i)}/q*)$ **then**
    Accept candidate sample. Let $a^{(i+1)} = a*$, $p^{(i+1)} = p*, q^{(i+1)} = q*$.
  else
    Reject and keep the previous sample: $a^{(i+1)} = a^{(i)}, p^{(i+1)} = p^{(i)}, q^{(i+1)} = q^{(i)}$.
  end if
  $i = i + 1$.
end while

To generate the following plots, note that candidate samples should be generated from an exponential distribution with mean $\mu = 5$, not the rate parameter λ.


Proposal
Posterior
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10


Sample histogram
Exact distribution
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9 10

Solution 3. (a) The conditional distributions are correspondingly:

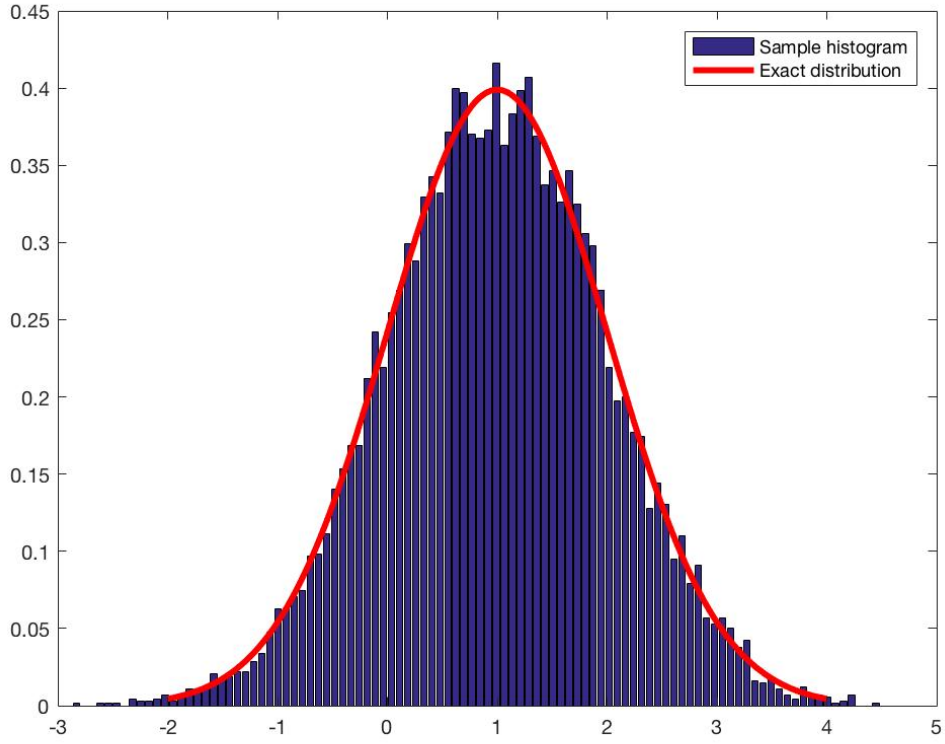
$$x_1|x_2 \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), (1-\rho^2)\sigma_1^2\right)$$
$$x_2|x_1 \sim \mathcal{N}\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(x_1 - \mu_1), (1-\rho^2)\sigma_2^2\right)$$

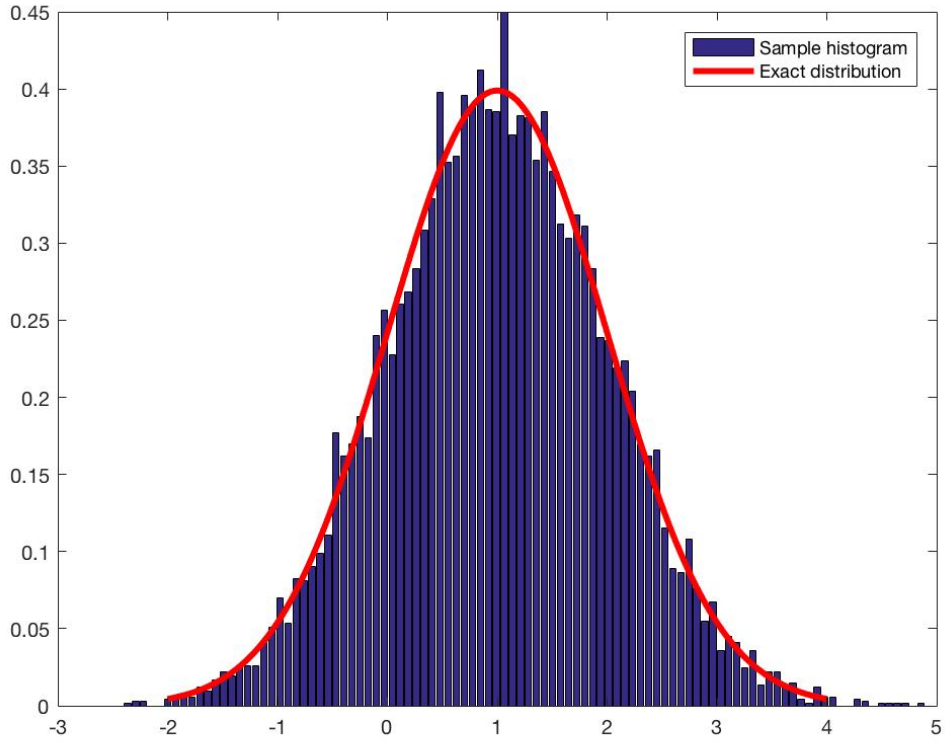
With such explicit knowledge on the forms of conditional distributions, we can conduct Gibbs sampling to generate samples of x. The algorithmic procedure for Gibbs sampling is detailed in Algorithm Q3a.

Q3a Gibbs sampling

1: Set up an initial sample $x^{(1)}$. Set $i = 1$.
2: **while** $i < M$ **do**
3: Generate the 1st dimension of the new sample $x_1^{(i+1)}$ from $x_1|x_2^{(i)}$.
4: Generate the 2nd dimension of the new sample $x_2^{(i+1)}$ from $x_2|x_1^{(i+1)}$.
$i = i + 1$.
5: **end while**

Histograms for x_1 and x_2 both match marginal densities pretty well.


Sample histogram
Exact distribution
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3
-2
-1
0
1
2
3
4
5


Sample histogram
Exact distribution
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3
-2
-1
0
1
2
3
4
5

(b) Let's consider the block-wise Metropolis Hastings as detailed in Algorithm Q3b:

Q3b Block-wise MH

Set up an initial sample $x^{(1)}$. Set $i = 1$.

Compute its target density $p^{(1)} = \mathcal{N}\left(x^{(1)}|\mu, \Sigma\right)$.

while $i < M$ **do**

  Generate a candidate sample $x* \sim \mathcal{N}\left(x^{(i)}, I\right)$.

  Compute its target density $p* = \mathcal{N}\left(x*|\mu, \Sigma\right)$.

  Draw $u \sim U[0, 1]$

  if $u < (p*/p^{(i)})$ **then**

    Accept candidate sample: $x^{(i+1)} = x*$, $p^{(i+1)} = p*$.

  else

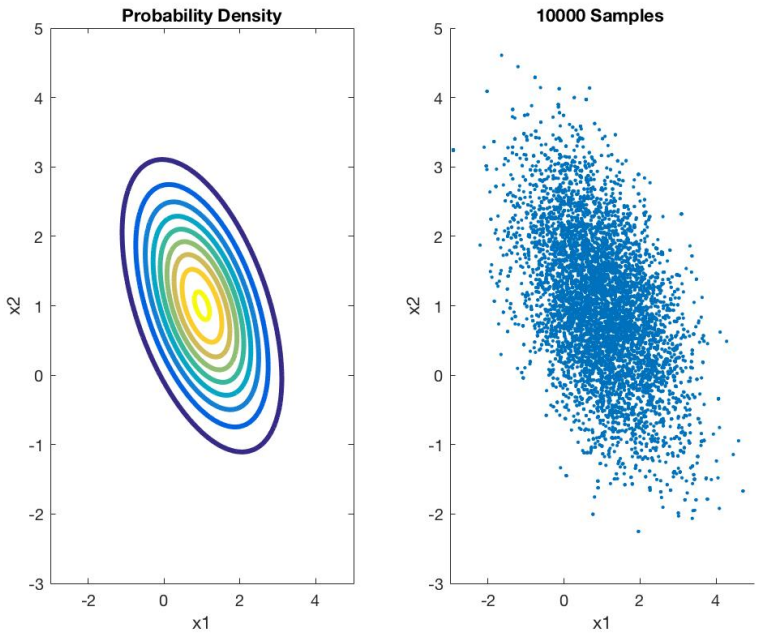
    Reject and keep the previous sample: $x^{(i+1)} = x^{(i)}, p^{(i+1)} = p^{(i)}$.

  end if

  $i = i + 1$.

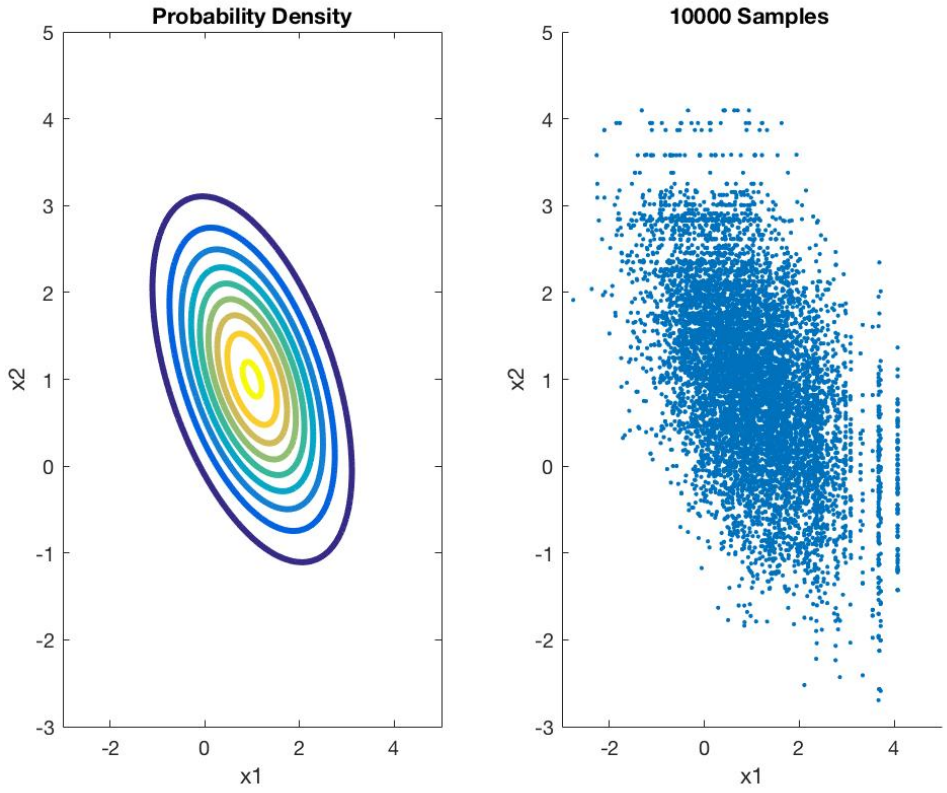
end while

Figure 1: Block-wise MH: Target density vs. Samples



(c) Let's consider the component-wise Metropolis Hastings. When it is possible to draw from each full-conditional distribution associated with the target, this is just a Gibbs sampler. When impossible to fully exploit the conditional distribution, we can use a component-wise proposal density (univariate Normal q) and update samples component-wise.

Figure 2: Component-wise MH: Target density vs. Samples



Q3c Component-wise MH

Set up an initial sample $x^{(1)}$. Set $i = 1$.

Compute its target density $p^{(1)} = \mathcal{N}\big(x^{(1)}|\mu, \Sigma\big)$.

while $i < M$ **do**

For 1st dimension:

Generate a candidate $x_1* \sim \mathcal{N}(0, 1)$ and compute the proposal density q_1*,

and the target density with such update: $p_1* = \mathcal{N}\big([x_1*, x_2^{(i)}]|\mu, \Sigma\big)$.

Draw $u \sim U[0, 1]$.

if $u < (p_1 * /p^{(i)})/(q_1^{(i)}/q_1*)$ **then**

Accept the candidate: $x^{(i+1)} = [x_1*, x_2^{(i)}]$, $p^{(i+1)} = p_1*$, $q_1^{(i+1)} = q_1*$.

else

Reject and keep the previous sample: $x^{(i+1)} = x^{(i)}$, $p^{(i+1)} = p^{(i)}$, $q_1^{(i+1)} = q_1^{(i)}$.

end if

For 2nd dimension:

Generate a candidate $x_2* \sim \mathcal{N}(0, 1)$ and compute the proposal density q_2*,

and the target density with such update: $p_2* = \mathcal{N}\big([x_1^{(i+1)}, x_2*]|\mu, \Sigma\big)$.

Draw $u \sim U[0, 1]$.

if $u < (p_2 * /p^{(i+1)})/(q_2^{(i)}/q_2*)$ **then**

Accept the candidate: $x_2^{(i+1)} = x_2*$, $p^{(i+1)} = p_2*$, $q_2^{(i+1)} = q_2*$.

else

Reject and keep the previous sample: $x_2^{(i+1)} = x_2^{(i)}$, $p^{(i+1)} = p^{(i)}$, $q_2^{(i+1)} = q_2^{(i)}$.

end if

$i = i + 1$.

end while

(d) Let's consider the Hamiltonian Monte Carlo. It introduces an *momentum* variable q, and uses Hamiltonian dynamics to generate samples. With proper configuration, it can converge more quickly to the target distribution than conventional MCMC.

Define the *potential energy* $U(x) = -\log p(x)$, the *kinetic energy* $K(q) = -\log p(q)$ (where $p(x)$ is the target density and $p(q)$ is the proposal density for q, usually a normal distribution), and the summation *Hamiltonian* $H(x,q) = U(x) + K(q)$. If we can generate samples $\propto \exp(-H(x,p)) = p(x)p(q)$, the resulting x samples will be distributed according to the target one.

We can then generate new candidate samples based on the Hamilton′s equation of motion:

$$\begin{aligned}\frac{\partial x_i}{\partial t} &= \frac{\partial H}{\partial p_i} = \frac{\partial K(q)}{\partial p_i} \\ \frac{\partial q_i}{\partial t} &= -\frac{\partial H}{\partial x_i} = -\frac{\partial U(x)}{\partial x_i}\end{aligned}$$

For numerical implementation, Hamilton′s equations must be approximated by discretizing time, using some small step size, ϵ. Starting with the state at time 0, we iteratively compute the states at $\epsilon, 2\epsilon, 3\epsilon$, till L steps of ϵ. One step of the leapfrog method works as follows to obtain states at $t+\epsilon$ from t for $i-$th dimension:

$$\begin{aligned}q_i(t+\epsilon/2) &= q_i(t) - (\epsilon/2)\frac{\partial U}{\partial x_i}x(t) \\ x_i(t+\epsilon) &= x_i(t) + \epsilon\frac{\partial K}{\partial q_i}q(t+\epsilon/2) \\ q_i(t+\epsilon) &= q_i(t+\epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial q_i}q(t+\epsilon/2)\end{aligned}$$

In other words, it starts with half step update for the momentum variable, and then do a full step for x using the updated momentum, and then do the other half step for momentum. The resulting $x(L\epsilon)$ and $q(L\epsilon)$ is our candidate sample.

Q3d Hamiltonian MC

Set up an initial sample $x^{(1)}$. Set $i = 1$.

Compute its target density $p^{(1)} = \mathcal{N}\big(x^{(1)}|\mu, \Sigma\big)$.

while $i < M$ **do**

  Generate an initial momentum variable $q^{(i)} \sim \mathcal{N}(0, \Sigma_q)$,

  and compute its Hamiltonian $H(x^{(i)}, q^{(i)})$.

  Generate a candidate sample $x*$ by conducting L leapfrog schemes with step-size ϵ.

  Compute its Hamiltonian $H(x*, q*)$. Draw $u \sim U[0, 1]$.

  if $u < \exp\big(-H(x*, p*) + H(x^{(i)}, q^{(i)})\big)$ **then**

    Accept candidate sample:$x^{(i+1)} = x*$.

  else

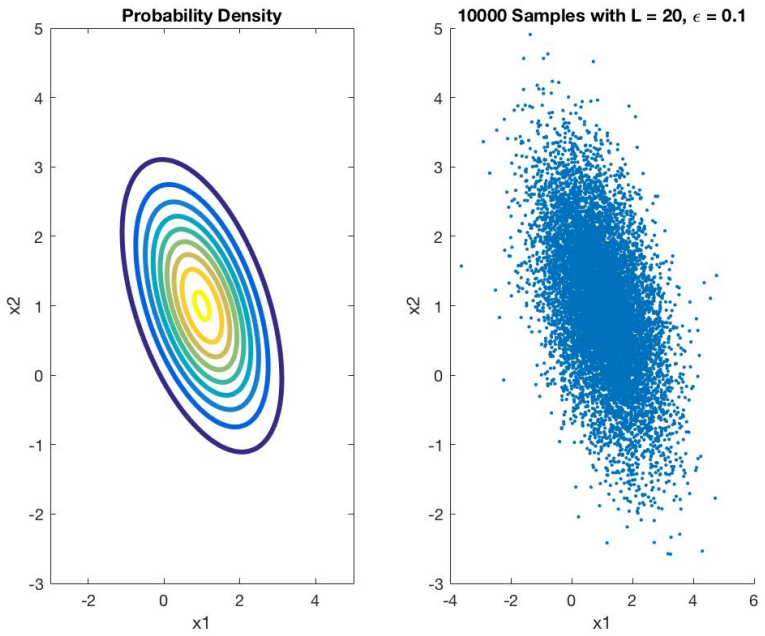
    Reject and keep the previous sample: $x^{(i+1)} = x^{(i)}$.

  end if

  $i = i + 1$.

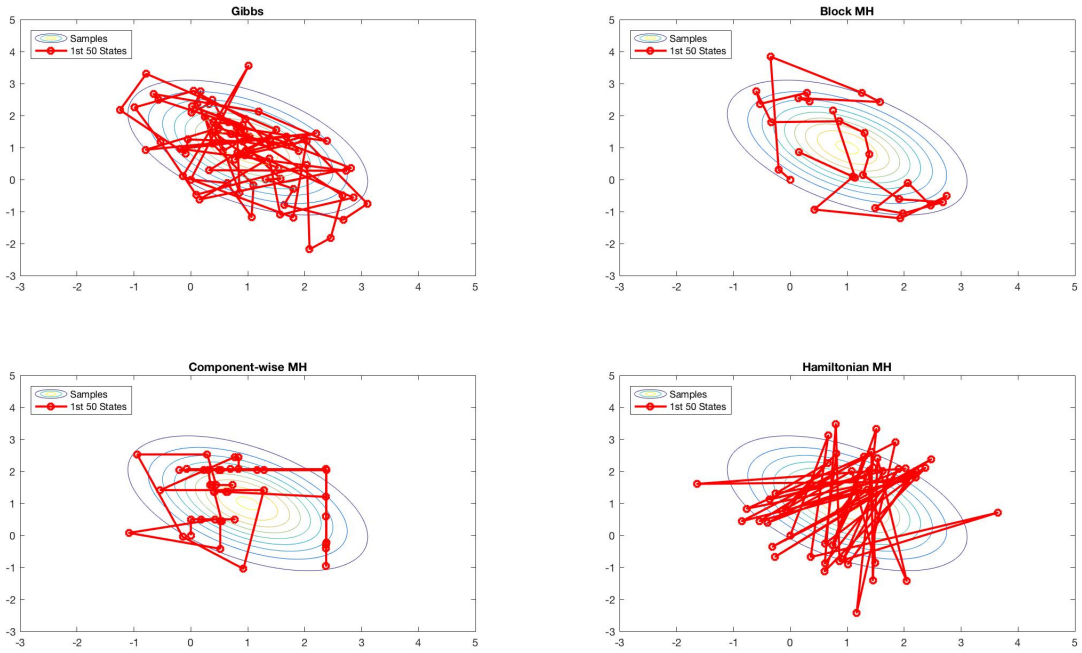
end while

Figure 3: Hamiltonian MC: Target density vs. Samples



For the convergence comparison, one can monitor the progress of sample mean/covariance matrix $\hat{\mu}, \hat{\Sigma}$ converging to the actual one μ, Σ, or one can show the trace plots of generated samples. Here we attach the trace plots of the first 50 samples, generated by all 4 algorithms tested above. It's clear that both Gibbs and HMC can explore the whole density more rapidly, since the former can exploit the full conditional density, while the latter utilizes gradient information. Block-wise and component-wise MH seem to perform inferior.

Figure 4: Trace plots of first 50 samples



Solution 4. The likelihood when the error is normal distributed can be written as (neglecting some constant terms)

$$l_n(\beta, \sigma^2 | y, X) = \left(\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2 \right]. \tag{2}$$

Observe the expression and it is clear that the MLE estimate of β is the one that minimizes the term $\sum_{i,j}(y_{ij} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$, which can be obtained by taking it derivative with respect to β, and set them to 0, we get:

$$\frac{\partial}{\partial\beta_0}(\sum_{i,j}(y_{ij}-\beta_0-\beta_1x_i-\beta_2x_i^2)^2)\rightarrow\sum_{i,j}y_{ij}=\sum_i n_i(\beta_0+\beta_1x_i+\beta_2x_i^2)$$
$$\frac{\partial}{\partial\beta_1}(\sum_{i,j}(y_{ij}-\beta_0-\beta_1x_i-\beta_2x_i^2)^2)\rightarrow\sum_{i,j}y_{ij}x_i=\sum_i n_i(\beta_0+\beta_1x_i+\beta_2x_i^2)x_i$$
$$\frac{\partial}{\partial\beta_2}(\sum_{i,j}(y_{ij}-\beta_0-\beta_1x_i-\beta_2x_i^2)^2)\rightarrow\sum_{i,j}y_{ij}x_i^2=\sum_i n_i(\beta_0+\beta_1x_i+\beta_2x_i^2)x_i^2$$

For convenience, we augment all the data y_{ij} into a N by 1 vector Y, such that $Y = [y_{1,1},\ldots,y_{1,n_1},y_{2,1}\ldots,y_{k,1},\ldots,y_{k,n_k}]$. We also create the corresponding input vector X by repeating the respective x_i by n_i times, i.e., $X = [x_1I_{(n_1\times1)},\ldots,x_kI_{(n_k\times1)}]$. The solution of β satisfying those equations has a closed form:

$$\hat{\beta}=\left([I,X,X^2]^T[I,X,X^2]\right)^{-1}[I,X,X^2]^TY$$

Plug in the estimated $\hat{\beta}$ back into the likelihood expression and minimize w.r.t. σ^2, we can obtained the MLE parameter's expression as

$$\hat{\sigma}^2=\frac{1}{N}\sum_{ij}(y_{ij}-\hat{\beta}_0-\hat{\beta}_1x_i-\hat{\beta}_2x_i^2)$$

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

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

The proposal density for β is then

$$\beta\sim\mathcal{N}\left(\hat{\beta},\mathbb{V}(\beta|X,\sigma^2)\right) \tag{3}$$

For the proposal density of parameter σ^2, notice that the following relationship holds (the well known Cochran's theorem):

$$\frac{N\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-3} = \mathcal{G}(\frac{N-3}{2}, 2) \rightarrow \frac{1}{\sigma^2} \sim \mathcal{G}(\frac{N-3}{2}, \frac{2}{N\hat{\sigma}^2})$$

Intuitively, it means the summation of squared independent normal random variables follows a $Chi - square$ distribution, and the degree of freedom is $N-3$ considering we have 3 regression terms I, X, X^2. Followed by some re-parameterization, we can see the precision parameter $\frac{1}{\sigma^2}$ follows the gamma distribution, which we'll use as the proposal distribution:

$$\frac{1}{\sigma^2} \sim \mathcal{G}(\frac{N-3}{2}, \frac{2}{N\hat{\sigma}^2}) \tag{4}$$

The final proposal density for β, σ^2 is then

$$p(\beta, \sigma^2) = \mathcal{N}\Big(\beta|\hat{\beta}, \mathbb{V}(\beta|X, \sigma^2)\Big)\mathcal{IG}(\sigma^2|\frac{N-3}{2}, \frac{N\hat{\sigma}^2}{2}) \tag{5}$$

$$= \mathcal{N}\left([2.47, 0.91, 0.1], \begin{bmatrix} 206.37 & -27.22 & 0.821 \\ -27.22 & 3.89 & -0.124 \\ 0.821 & -0.124 & 0.0041 \end{bmatrix}\right)\mathcal{IG}(23.5, 5405) \tag{6}$$

We choose the independent Metropolis sampler to generate samples of β, σ^2. We monitor convergence by observing the sample mean of β, σ^2 over generated samples. It's obvious that the mean all parameters converge to a stable value.

Figure 5: Convergence

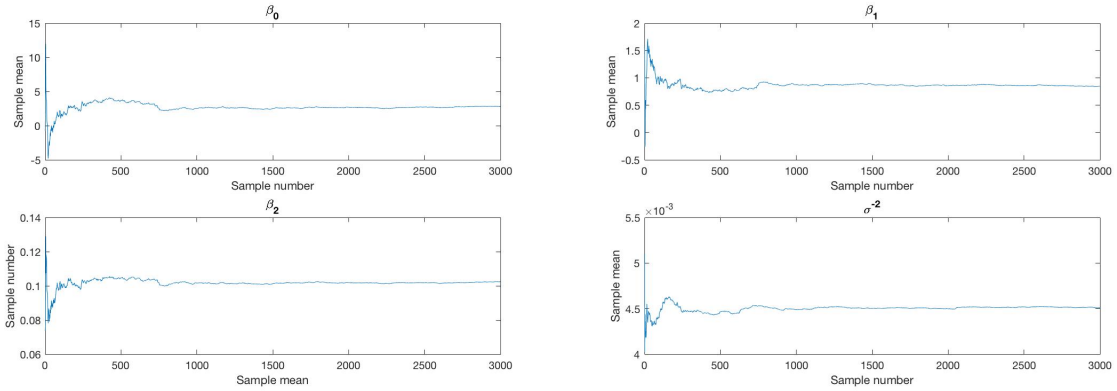
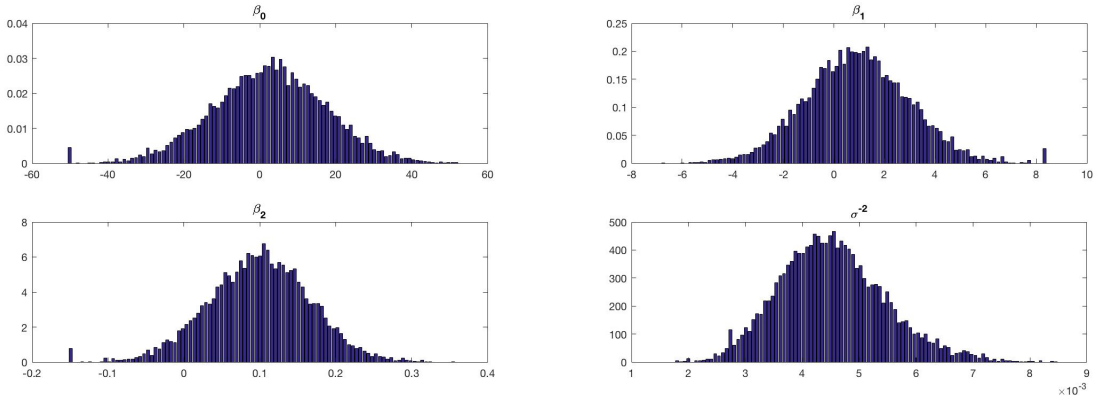


Figure 6: Histogram



Q4a Steps to generate M samples

1: Set the initial sample $\beta^{(1)}, \sigma^{2(1)}$ to be the MLE parameters $\hat{\beta}, \hat{\sigma}^2$.

2: Compute its likelihood $l_n^{(1)} = l(\beta^{(1)}, \sigma^{2(1)}|y, X)$ as in Eq. (2),

and its proposal density $p^{(1)} = p(\beta^{(1)}, \sigma^{2(1)})$ as in Eq. (5).

3: **while** $i < M$ **do**

4: Generate a candidate sample with $\beta*, \sigma^2* \sim p$.

Compute its likelihood $l_n* = l_n(\beta*, \sigma^2 * |y, X)$ as in Eq. (2),

and its proposal density $p^{(1)} = p(\beta*, \sigma^2*)$ as in Eq. (5).

Generate $u \sim U[0, 1]$

5: **if** $u < (l_n * /l_n^{(i)})/(p^{(i)}/p*)$ **then**

6: Accept candidate sample:

$\beta^{(i+1)} = \beta*, \sigma^{2(i+1)} = \sigma^2*, l_n^{(i+1)} = l_n*, p^{(i+1)} = p*$.

7: **else**

8: Reject and keep the previous sample:

$\beta^{(i+1)} = \beta^{(i)}, \sigma^{2(i+1)} = \sigma^{2(i)}, l_n^{(i+1)} = l_n^{(i)}, p^{(i+1)} = p^{(i)}$.

9: **end if**

$i = i + 1$.

10: **end while**

For the case when residual errors are *student-T* distributed, first recall that the student-T distribution is indeed a scaled infinite mixture of normal distribution. In other words, the residual $\epsilon_{ij} \sim \mathcal{T}(v = 4, 0, \sigma^2)$ is equivalent to that $\epsilon_{ij} \sim \mathcal{N}(0, w\sigma^2)$, where w also follows an inverse gamma $\mathcal{IG}(\frac{v}{2} = 2, \frac{v}{2} = 2)$. Conditioned on any mixture component w, the process would be identical to the one with normal distributed error.

Since the MLE estimate for β doesn't depend on the variance of error, we can conclude that the MLE estimate for β under *student-T* distribution remains the same as the previous case. Plugging it back into the equation, the MLE parameter of σ^2 also remains unchanged. We can again rely on its covariance matrix approximation to determine the proposal density for β:

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

In this case, because of the robustness for *student-T* residuals, the covariance $\mathbb{V}_t(\beta|X, \sigma^2)$ is bigger than the previous one $\mathbb{V}(\beta|X, \sigma^2)$.

To deal with σ^2, we again resort to the previous proposal density for σ^2 as written in Eq. (4). The overall proposal density is

$$p(\beta, \sigma^2) = \mathcal{N}\Bigg(\beta|\hat{\beta}, \mathbb{V}_t(\beta|X, \sigma^2)\Bigg)\mathcal{IG}(\sigma^2|\frac{N-3}{2}, \frac{N\hat{\sigma}^2}{2}) \tag{7}$$

$$= \mathcal{N}\Bigg([2.47, 0.91, 0.1], \begin{bmatrix} 412.74 & -54.44 & 1.642 \\ -54.44 & 7.78 & -0.248 \\ 1.642 & -0.248 & 0.0082 \end{bmatrix}\Bigg)\mathcal{IG}(23.5, 5405) \tag{8}$$

The desired joint distribution of β, σ^2 can be written as:

$$l_t(\beta, \sigma^2, w|y, X) = \sigma^{-N} \prod_{i,j} \left(1 + \frac{1}{v}\frac{(y_{ij} - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2}{\sigma^2}\right)^{-(v+1)/2} \tag{9}$$

Q4b Steps to generate M samples

Set the initial sample $\beta^{(1)}, \sigma^{2(1)}$ to be the MLE parameters $\hat{\beta}, \hat{\sigma}^2$,

and sample $w^{(1)} \sim \mathcal{IG}(2,2)$.

Compute its likelihood $l^{(1)} = l(\beta^{(1)}, w^{(1)}\sigma^{2(1)}|y, X)$ as in Eq. (9),

and its proposal density $p^{(1)} = p(\beta^{(1)}, \sigma^{2(1)}, w^{(1)})$ as in Eq. (7).

while $i < M$ **do**

  Generate a candidate sample with $\beta*, \sigma^2*, w* \sim p$.

  Compute its likelihood $l* = l(\beta*, w * \sigma^2 * |y, X)$ as in Eq. (9),

  and its proposal density $p* = p(\beta*, \sigma^2*, w*)$ as in Eq. (7).

  Generate $u \sim U[0,1]$

  if $u < (l * /l^{(i)})/(p^{(i)}/p*)$ **then**

    Accept candidate sample:

    $\beta^{(i+1)} = \beta*, \sigma^{2(i+1)} = \sigma^2*, w^{(i+1)} = w*, \; l^{(i+1)} = l*, p^{(i+1)} = p*.$

  else

    Reject and keep the previous sample:

    $\beta^{(i+1)} = \beta^{(i)}, \sigma^{2(i+1)} = \sigma^{2(i)}, w^{(i+1)} = w^{(i)}, \; l^{(i+1)} = l^{(i)}, p^{(i+1)} = p^{(i)}.$

  end if

  $i = i + 1$.

end while

The histogram is also attached:

Figure 7: Histogram