PROBLEM SET # 5

EC/ACM/CS 112: Bayesian Statistics

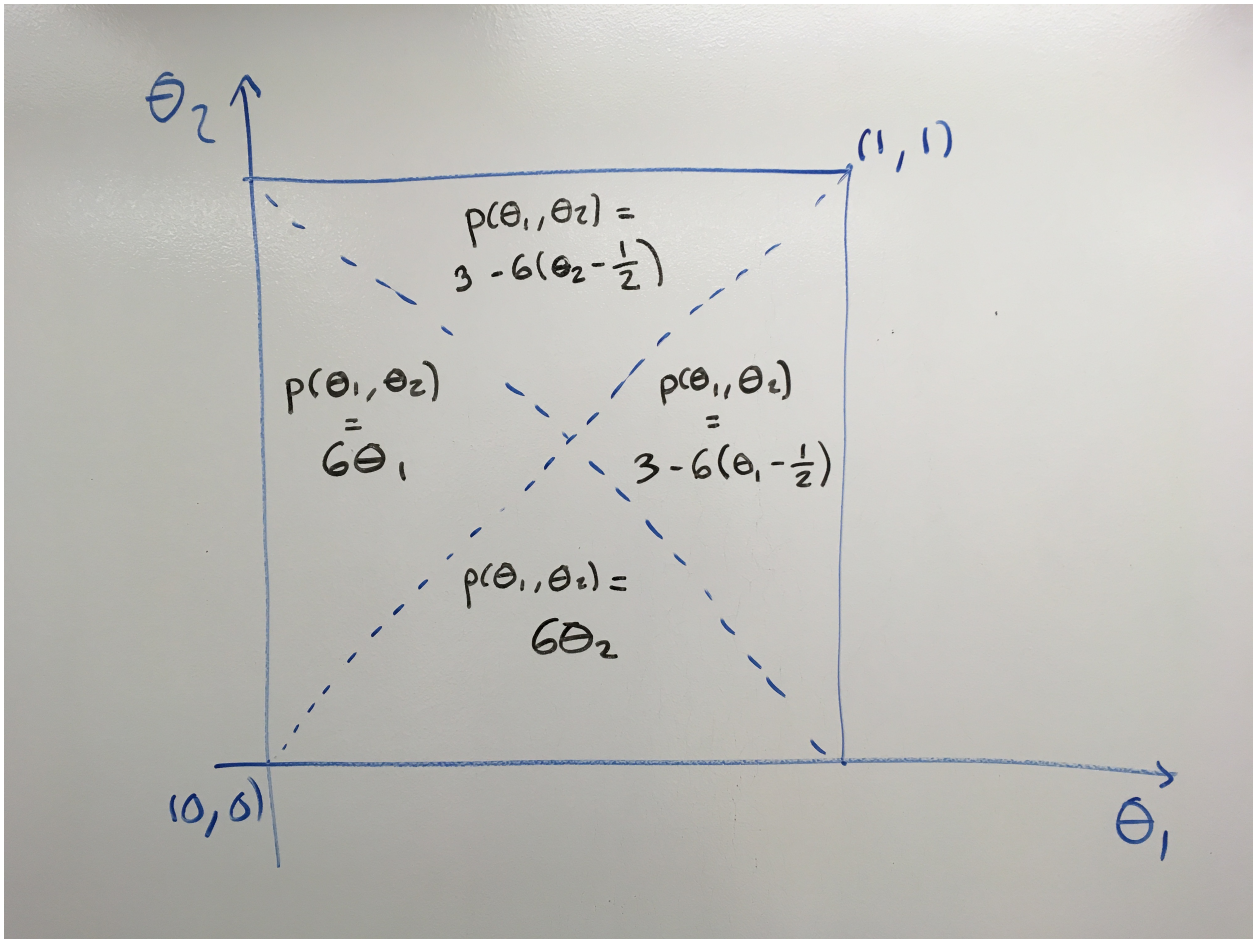
Due date	Tuesday, February 9, 10:30 am
Submission instructions	>> Create a pdf of the R Notebook with your solutions (details below) >> Submit in Canvas
Additional files included in the problem set package	>> dataset for problem set >> solutions template (.Rmd file)

QUESTION 1. BUILD YOUR OWN METROPOLIS-HASTINGS SAMPLER

LEARNING GOAL

- Build your own Metropolis-Hasting sampler in order to deepen your understanding of how it works before applying it to an actual dataset

Step 1. Program a series of functions that will allow you to use the Metropolis-Hastings algorithm to sample from the following target distribution:



Note that the vector of parameters that we want to sample from is 2-dimensional, and that the target distribution only has non-zero probability in the unit square depicted above.

The first function that you should program, called **TargetDtbn**, should take as input a vector of parameters $\theta = (\theta_1, \theta_2)$ and return the value of the target distribution $p(\theta)$. Recall that the target distribution is the one that we want to sample from.

The second function that you should program, called **SampleProposalDtbn**, should take as input the current vector of parameters θ_t and return a candidate vector of parameters θ^* by sampling from a time-independent proposal distribution $J(\theta^*|\theta_t)$ = Normal(θ_t, Σ), with $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$.

The third function that you should program, called **DensityProposalDtbn**, should take as input two vectors of parameters, θ and θ', and return the conditional density $J(\theta|\theta')$.

The fourth function that you should program, called **MHsampling**, should take as input the following elements:

- The TargetDtbn function
- The SampleProposalDtbn function
- The DensityProposalDtbn function
- The number of desired samples
- A vector θ_0 that initializes the sampling process (recall that it must satisfy $p(\theta_0) > 0$).

The function returns a matrix with the samples, one for each row.

TIPS:

- The R library **mvtnorm** is useful in programming the proposal distribution functions.
- You might want to add a progress feedback message to your MHsampling function to help you keep track of the progress of the sampling algorithm. The simplest way of doing this is to print periodically a message like *"step t out of T completed"* .

Step 2. Draw 250,000 samples of the target distribution, using $\sigma^2 = 0.25$ in your proposal distribution. Visualize the results using a scatter plot of the samples. Note that each sample is a point in the plot. Label the axes properly and customize the plot to get a good visualization.

TIPS:

- Use the option **col=rgb(red=0.0, green=0.0, blue=1.0, alpha=0.0075)** to make the points semi-transparent, which helps with the visualization.

Step 3. Compute the marginal distribution of θ_1 analytically and compare it with the one resulting from your sampler. To accomplish this:

- Make a histogram of the θ_1 samples generated by your sampler.
- Add a line with the marginal posterior function $p(\theta_1)$ that you computed analytically.

TIPS:

- Use the **prob=TRUE** option in the histogram plot.

- Use the **breaks=100** option in the histogram plot to improve the quality of the visualization.

Step 4. Use the samples to compute the following properties of the target distribution:

- Mean of θ_1
- Mean of θ_2
- Variance of θ_1
- Variance of θ_2
- Correlation between θ_1 and θ_2.

QUESTION 2. USING METROPOLIS-HASTINGS ON A FAMILIAR DATASET

LEARNING GOAL

- Apply the MH-Algorithm to estimating the posterior in a real dataset.
- Build your intuition and comfort with the MH-algorithm by applying it to a familiar dataset and statistical model

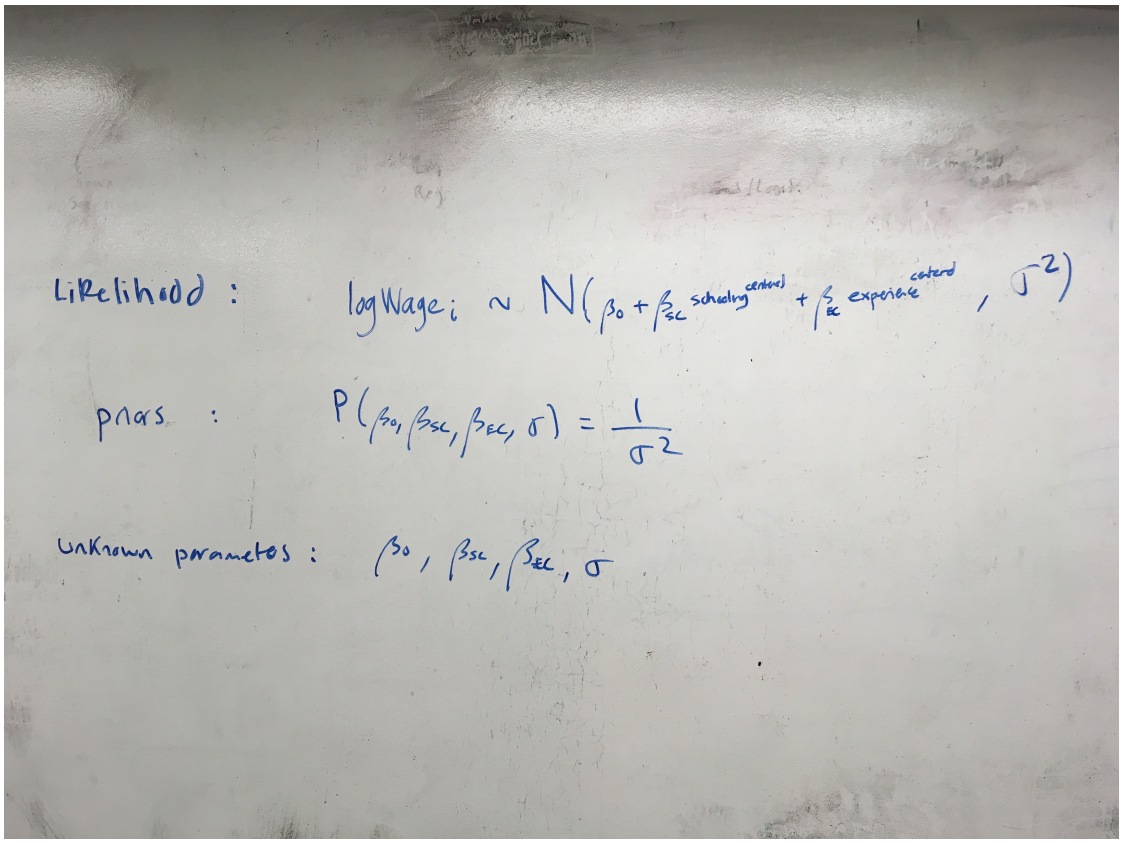
DATASET. The dataset for this problem is "Wages1.csv", which we have been using in class to study linear regression.

It contains measures of the following variables for 3,294 individuals:

- X = individual identifier
- wage = hourly wage per-hour (in $)
- schooling = years of schooling
- experience = years of full time work experience
- sex = "male" or "female"

The dataset has no missing observations.

STATISTICAL MODEL. We want to estimate the following linear regression model, which we have already estimated in previous lectures using the grid method.



Step 1. Modify your code from part 1 to be able to fit this model on the wage data using the MH-Algorithm.

Note the following:

- Your target function is now $p(\beta_0, \beta_{SC}, \beta_{EC}, \sigma|data)$.
- You can work with an improper posterior given by prior * likelihood.
- Every step, the MH-Algorithm generates samples over the four unknown parameters: $\beta_0, \beta_{SC}, \beta_{EC}$ and σ

You can use any proposal distribution that you want, but be careful with the following issues:

- It should return a strictly positive σ in every step, since the likelihood is not well-defined otherwise.
- You need to experiment with the variability of the proposal distribution to make sure that the MH-algorithm samples properly from the posterior.

TIPS:

- Use a proposal distribution that samples independently for each parameter.
- Work with logSums when computing likelihoods and acceptance ratios to avoid numerical floating problems due to the limited numerical precision of the software.

Step 2. Use the code from step 1 to generate 510,000 of the posterior. The first 10,000 are used as "burn-in samples" designed to help the chain converge to the steady state, and you should throw them away.

Plot a histogram of the marginal posterior distribution of each of the parameters. Use 100 bins to facilitate visualization. Add a red vertical line at the mean of the samples, as well as a blue vertical line at the ordinary-least-squares estimates $\hat{\beta}_0, \hat{\beta}_{SC}, \hat{\beta}_{EC}$.

TIP:

- If everything is working correctly, these two lines should lie on top of each other, and the histograms for each of the parameters should look similar to ones generated by normal distributions.

Step 3. Use the posterior samples to compute the standard deviation of the marginal posteriors $P(\beta_0|data)$, $P(\beta_{SC}|data)$, and $P(\beta_{EC}|data)$. Compute also the posterior correlation between β_{SC} and β_{EC}.