Stat 515, Spring 2015: Take home final

Due Wednesday, April 29, 2015 at midnight

1. Consider a regression of a variable Y on X where the regression model is as follows, $Y_i \sim EMG(\beta_0 + \beta_1 X, \sigma_i, \lambda)$, where the exponentially modified Gaussian random variable, EMG(μ, σ, λ), has pdf $f(x; \mu, \sigma, \lambda) = \frac{\lambda}{2} \exp(\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x))\text{erfc}\left(\frac{\mu+\lambda\sigma^2-x}{\sqrt{2}\sigma}\right)$, and erfc is the complementary error function defined as

$$\text{erfc}(x) = \frac{2}{\pi}\int_x^\infty e^{-t^2} dt.$$

The EMG distribution is obtained when a normal density is convolved with an exponential density. That is, for the regression above, the error contains a normal error (with standard deviation σ) added to an exponential error (with rate λ, that is, expected value $1/\lambda$). `R` code for the EMG density function is here: `http://sites.stat.psu.edu/~mharan/515/hwdir/emg.R`

   (a) Assume that $\beta_0 = 5, \lambda = 0.4$, and $\sigma_i = 1$ for all i. Let the prior for β_1 be N$(0, 10)$ (parameterization: N(mean, sd)). Write a Metropolis-Hastings algorithm to approximate the posterior distribution, $\pi(\beta_1 \mid \mathbf{Y}, X)$ for Dataset#1 on Angel. Clearly and succinctly describe the algorithm you used, with enough detail so anyone reading it should be able to write code based on your description. You should also provide important details such as starting values (e.g. arbitrary values, values based on several preliminary MCMC runs, a random draw from some initial distribution you chose etc.) *Note: here, as in other parts, you will lose points if your answer is either unclear or incomplete.*

   (b) Report your estimate of the posterior expectation of β_1. This is a point estimate for β_1. Also report the MCMC standard error associated with this estimate.

   (c) Report a 95% credible interval for β_1 based on your samples. Credible intervals are the Bayesian analogue of frequentist confidence intervals. The interpretation is that if (L, B) is the credible interval, $P(\beta_1 \in (L, B) \mid \mathbf{Y}, X) = 0.95$. A simple 95% credible interval may be obtained by reporting the 2.5th and 97.5th sample percentiles from your Markov chain. In R, if you have stored your Markov chain in the vector mySamples, you can use the command `quantile(mySamples, c(0.025, 0.975))`.

   (d) Plot an estimate of the posterior pdf of β_1 from a smoothed density plot of the samples. In R, you can use the command `plot(density(mySamples))`.

   (e) Describe how you determined that your approximations above were accurate, along with any supporting information as discussed in class, e.g. plots of autocorrelations, MCMC standard errors etc. *All your plots must be clearly labelled and referenced in your text.*

2. Now assume that only $\sigma_i = 1$ is known. Write a Metropolis-Hastings algorithm to approximate the posterior distribution, $\pi(\beta_0, \beta_1, \lambda \mid \mathbf{Y}, X)$ for Dataset #2 on Angel. Assume priors $\beta_0 \sim N(0, 10), \beta_1 \sim N(0, 10)$ (parameterization: N(mean, sd)), $\lambda \sim$Gamma$(0.01, 100)$ (w/ Gamma parameterization such that prior expected value of λ is 1 and variance is 100).

(a) Clearly and succinctly describe the algorithm you used, with enough detail so anyone reading it should be able to write code based on your description.

(b) Provide, preferably in a well organized table, for $\beta_0, \beta_1, \lambda$, the posterior mean w/ estimate of MCMC standard error in parentheses, posterior 95% credible intervals.

(c) Provide an approximation of the correlation between β_0, β_1.

(d) Provide approximate density plots for the marginal distributions of $\beta_0, \beta_1, \lambda$.

(e) Describe how you determined that your algorithm was producing reliable approximations. Provide relevant plots and justifications.

3. Repeat Problem #2 above except do it for Dataset#3 on Angel.

(a) Provide, preferably in a well organized table, for $\beta_0, \beta_1, \lambda$, the posterior mean w/ estimate of MCMC standard error in parentheses, posterior 95% credible intervals.

(b) Provide approximate density plots for the marginal distributions of $\beta_0, \beta_1, \lambda$.

(c) Explain (if and) how you modified your MCMC algorithm in order to make it work better for this problem.

Please read the following instructions carefully.
Important (*you will lose points if you do not follow these instructions*):

1. Your writeup must be **no longer than 5 pages** (it may be shorter, but not longer) including all plots and discussions. *Anything after the 5th page will be ignored.*

2. You may not discuss the exam with anyone except the instructor, that is, you cannot even talk to the T.A. Please feel free to ask me as many questions about this as you like. I may choose not to answer some of the questions, but I would rather you checked with me if you are unsure. If I choose to answer the question, I will email my response to the entire class.

3. You are required to submit your R code to `Angel` **by the due date/time**. It is an electronic system so it will not allow you to submit it even if you are a few seconds late. Use the following naming convention for the file you upload – first initial, name, .R extension. For instance if your name is John Newman you would attach a file called jnewman.R. If you have more than one program file, simply give them extensions. For e.g.: jnewman1.R jnewman2.R etc. If you use some other programming language, say Matlab or C, use the appropriate extension for that language, e.g. jnewman.c or jnewman.m

4. You are required to submit your pdf code to `Angel` **by the due date/time**. Follow the same naming convention, for e.g.: jnewman.pdf . Please do not submit your writeup in any other format.

5. Do not include your code in the writeup.

6. Make sure you add comments to your code; this is generally a good practice and can also be helpful in explaining the organization of your program to me.

7. Use a sensible editor for your program; otherwise, your program will be very difficult for me to read (and you will have more trouble with your programming).

8. *It should be obvious how I would run your code, if I were to chose to do so.* If I find the code hard to run, for e.g. if your code calls files in other directories (that I don't have!) you will lose points. Test your code carefully.

9. Avoid having too many plots or unreadable plots, and make sure they are all well labeled. Note: you can produce multiple plots on the same page by using the `par` command in `R`.