CS6190: Probabilistic Modeling **Due Th 2/11**

Homework 1: Exponential Families, Conjugate Priors

Instructions: Be sure to electronically submit your answers in either R Markdown (*.Rmd) or Sweave (*.Rnw) format. Include all of the output of your code, plots, and discussion of the results in your write up. You may work together and discuss the problems with your classmates, but write up your final answers entirely on your own.

Written Part

1. **Expectation of sufficient statistics:** Consider a random variable X from a continuous exponential family with *natural* parameter $\eta = (\eta_1, \ldots, \eta_n)$. Recall that this means the pdf is of the form:

$$p(x) = h(x) \exp\left(\eta \cdot T(x) - A(\eta)\right).$$

   (a) Show that $\mathrm{E}[T(X) \,|\, \eta] = \nabla A(\eta) = \left(\frac{\partial A}{\partial \eta_1}, \ldots, \frac{\partial A}{\partial \eta_d}\right)$.

   Hint: Start with the identity $\int p(x)dx = 1$, and take the derivative with respect to η.

   (b) Verify this formula works for the Gaussian distribution with unknown mean, μ, and known variance, σ^2.

   Hint: Start by thinking about what the natural parameter η and the function $A(\eta)$ are, then verify that the expectation of the Gaussian is the same as $\nabla A(\eta)$.

2. **Noninformative priors for the Poisson distribution:** Let $X \sim \mathrm{Pois}(\lambda)$. Recall the pmf for the Poisson is

$$P(X = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

   (a) Rewrite the above pmf in exponential family form. What is the natural parmater? What is the sufficient statistic?

   (b) Give at least two different options for noninformative priors for $p(\lambda)$.

   (c) What are the resulting posteriors for your two options? Are they proper (i.e., can they be normalized)?

3. **Non-conjugate priors:** Let X_i be from a Gaussian with known variance σ^2 and mean μ with *uniform prior*, i.e.,

$$\mu \sim \mathrm{Unif}(a, b)$$
$$X_i \sim N(\mu, \sigma^2)$$

   What is the posterior pdf, $p(\mu \,|\, x_1, \ldots, x_n; \sigma^2, a, b)$?

   Hint: There will be an integral that you won't be able to analytically solve (just leave it in integral form).

R Coding Part

In this exercise we will be using data from the OASIS brain database, a publicly-available resource:

`http://www.oasis-brain.org`

You will be analyzing the volume of the hippocampus, a brain structure that is critical to memory, and its role in Alzheimer's disease. The data consists of the hippocampal volume, derived from MRI, for 188 elderly subjects, including healthy control subjects, and those with mild to moderate dementia. First, download the data and the R code for loading it from the class website.

Model the right hippocampal volume (`RightHippoVol`) as a normal random variable Y_{ij} for $j \in \{1, 2, 3\}$ and $i = 1, \ldots, n_j$, where j represents the dementia categorization (1 = "Control", 2 = "Mild", 3 = "Dementia") and n_j is the sample size for the jth group. Each group will have its own mean and variance, that is,

$$Y_{ij} \sim N(\mu_j, \sigma_j^2).$$

4. Use a conjugate prior $(\mu_j, \sigma_j^2) \sim \text{N-IG}(0, 10^{-6}, 10^{-6}, 10^{-6})$, independently for each group j.

   (a) What is the joint posterior density $p(\mu_j, \sigma_j^2 | y_{ij})$? What is the marginal posterior $p(\sigma_j^2 | y_{ij})$? Write a function to sample from this distribution. Plot a histogram for 10^6 samples of σ_1^2 and draw the density function over it. Draw the sample variance $\hat{\sigma}_1^2$ as a vertical line.

   (b) What is the marginal posterior distribution $p(\mu_j | y_{ij})$? Plot the marginal posterior densities for each mean $j = 1, 2, 3$ in the same figure. Draw the sample means $\hat{\mu}_j$ as vertical lines.

   (c) Consider the random variable $d_{12} = \mu_1 - \mu_2$. What is the conditional density $p(d_{12} | \sigma_1^2, \sigma_2^2, y_{ij})$? Generate 10^6 samples from the posterior $p(d_{12} | y_{ij})$ and estimate the probability $P(d_{12} < 0 | y_{ij})$. **Note:** You should sample σ_1^2, σ_2^2 from $p(\sigma_j^2 | y_{ij})$, then given those values, sample from $p(d_{12} | \sigma_1^2, \sigma_2^2, y_{ij})$. Repeat the same analysis for $d_{13} = \mu_1 - \mu_3$ and $d_{23} = \mu_2 - \mu_3$. What do these probabilities tell you?

   (d) How does your analysis compare to a one-sided t-test of the difference between two means? Please describe in terms of the logic of the two approaches and in terms of the resulting probabilities. **Hint:** you can do a t-test in R using the function `t.test`.

5. Repeat the same steps in parts 4 (a)–(c), but this time use independent noninformative Jeffreys' priors for μ_j and σ_j, i.e., $p(\mu_j, \sigma_j) \propto \sigma_j^{-2}$. What can you say about the relationship between this prior and the conjugate prior you used in 4? Discuss how the results are different and why.

6. A study from Schuff et al. [1] using a different data set calculated that the average hippocampus volume for 127 elderly controls was 2,133 mm^3, with a standard deviation of 279 mm^3. Use this information as pseudo-observations to construct conjugate priors for the means, μ_j (again use independent priors for each group). Repeat the analysis of 4 (a)–(c). How do the results change? Do you think this is a good prior to use? Explain why or why not.

Optional: If you want to go further, try a Bayesian analysis of a linear regression model with age as the independent variable. Similar to how you compared the means, you can compare the slopes of the three groups.

[1] Schuff et al., *Brain*, 132(4), pp. 1067–1077, 2006.