Hw3 for 6448 (due March 16) (Debdeep Pati)

1. To understand the effect of a potential carcinogen, a study was undertaken where $n = 23$ rats were treated with the potential carcinogen, and the time to tumor occurrence (in months) was recorded. Of particular biological relevance was whether a tumor developed within 6 months, and hence the data was collected over a period of 6 months. Let z_i denote the time to tumor development (in months) for the ith rat, with $z_i \in \{1, ..., k+1\}$ and $k = 6$. For all the rats which didn't develop a tumor within the first 6 months, we set $z_i = 7$. The simplest version of a *continuation ratio* model for the time to tumor occurrence can be expressed as

$$\mathbb{P}(z_i = j) = p(1-p)^{j-1}, \quad j = 1, \ldots, k$$

independently for $i = 1, \ldots, n$. The parameter p is commonly referred to as the discrete hazards, with the interpretation that $p = \mathbb{P}(z_i = j \mid z_i \geq j)$, that is, p is the probability of the occurrence of the tumor during month j given that it has not occurred within the first $(j-1)$ months.

   (a) Assuming a Beta$(1/2, 1/2)$ prior on p, calculate the posterior distribution of p.

   (b) Out of the 23 rats in the study, 11 developed tumor within the first month, 5 during the second month, 2 during the third month, 2 during the fourth month, 1 during the fifth month and 1 during the sixth month. One rat did not develop tumor within the first 6 months. Use this information to calculate the posterior mean of p under the Beta$(1/2, 1/2)$ prior.

2. Let f and g be two probability density functions on $\mathbb{R}$ with $f(\theta)/g(\theta) \neq 0$ for all $\theta \in \mathbb{R}$. The Kullback-Leibler (KL) divergence between f and g, denoted $\mathrm{KL}(f||g)$, is defined as

$$\mathrm{KL}(f||g) = \int_{\theta \in \mathbb{R}} f(\theta) \log \left[\frac{f(\theta)}{g(\theta)}\right] d\theta$$

Like the total variation distance, $\mathrm{KL}(f||g)$ is a "measure of distance" between densities f and g, though KL is not a distance metric. If f and g respectively have $\mathrm{N}(\mu_1, \tau_1^2)$ and $\mathrm{N}(\mu_2, \tau_2^2)$ distributions, we often write $\mathrm{KL}\big[\mathrm{N}(\mu_1, \tau_1^2), \mathrm{N}(\mu_2, \tau_2^2)\big]$ instead of $\mathrm{KL}(f||g)$ Suppose $x \mid \theta \sim N(\theta, 1/n)$ and θ is assigned a $\mathrm{N}(0,1)$ prior. Let θ_n and σ_n^2 denote the posterior mean and variance of θ, so that the posterior distribution of $\theta \mid x$ is a $\mathrm{N}(\theta_n, \sigma_n^2)$ distribution. Suppose the true data generating parameter is θ_0; let $\mathbb{E}_0$ denote an expectation under a $\mathrm{N}(\theta_0, 1/n)$ distribution. Let

$$T_n = \mathrm{KL}\big[\mathrm{N}(\theta_n, \sigma_n^2)||\mathrm{N}(x, 1/n)\big].$$

Find T_n. Show that $\mathbb{E}_0(T_n) \to 0$ as $n \to \infty$. Interpret the result.

3. A coin with probability p of turning heads is independently flipped 10 times. Assume a $\mathrm{U}(0,1)$ prior on p. We are told that 7 out of the 10 flips landed in tails. The results for the remaining three flips are not disclosed. Based on this data, calculate the posterior distribution and the posterior mean of p; can you identify what distribution the posterior is?

4. Consider the linear regression model

$$Y = X\beta + \epsilon,$$

where Y and ϵ are n dimensional column vectors, X is an $n \times p$ matrix, and β is a k dimensional column vector of parameters. The error ϵ is assumed to have a $\mathrm{N}(0, \sigma^2 I_n)$ distribution, where I_n is the $n \times n$ identity matrix. Assuming $p \leq n$, and X to have full column rank, consider the g-prior for a fixed $g > 0$,

$$\beta \sim \mathrm{N}(\beta_0, g\sigma^2(X^{\mathrm{T}}X)^{-1}), \quad \pi(\sigma^2) \propto \frac{1}{\sigma^2}$$

   (a) Find the distribution of $(\beta, \sigma^2) \mid Y, X$ in closed form.

   (b) Find $\mathbb{E}[\beta \mid Y, X]$ and $\mathbb{E}[\sigma^2 \mid Y, X]$ in closed form.

   (c) Find a highest posterior credible interval for $\beta_j, j = 1, \ldots, p$.