Course 52558: Problem Set 2 Solution

1. Computing with a non-conjugate model. Suppose \( y_1, \ldots, y_n \) are independent samples from a Cauchy distribution with unknown center \( \theta \) and known scale 1 so that
\[
p(y_i | \theta) \propto \frac{1}{1 + (y_i - \theta)^2}
\]
Assume, for simplicity, that the prior distribution for \( \theta \) is uniform on [0, 1].
Given the observations \((y_1, \ldots, y_5) = (-2, -1, 0, 1.5, 2.5)\):

**Solution:**
Use Cauchy.m (along with cauchyrnd.m and cauchyinv.m) in the course’s website to get figures for all sections of this question in matlab

(a) Compute (i.e. write a function for) the unnormalized density function \( p(\theta)p(y | \theta) \), on a grid of points \( \theta = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1 \), for some large integer \( m \). Using the grid approximation, compute and plot the normalized posterior density function, \( p(\theta | y) \), as a function of \( \theta \).

(b) Sample 1000 draws of \( \theta \) from the posterior density and plot a histogram of the draws.

(c) Use the 1000 samples of \( \theta \) to obtain 1000 samples from the predictive distribution of a future observation, \( y_6 \), and plot a histogram of the predictive draws.

2. Jeffrey’s prior. Suppose \( y | \theta \sim \text{Poisson}(\theta) \). Find Jeffrey’s prior density for \( \theta \), and then find \( \alpha \) and \( \beta \) for which the Gamma(\( \alpha, \beta \)) density is a close match to Jeffrey’s density. Use a plot to verify that the densities are indeed similar.

**Solution:**
The Poisson density function is \( p(y | \theta) = \theta^y e^{-\theta} / y! \) and so
\[
I(\theta) = E \left[ -\frac{\partial^2 \log p(y | \theta)}{\partial \theta^2} | \theta \right] = E \left[ \frac{y}{\theta^2} \right] = \frac{1}{\theta}
\]
and so the Jeffry’s prior is \( \theta^{-1/2} \). Mathematically, this is an improper Gamma density with \( \alpha = 0.5 \) and \( \beta = 0 \). Using matlab to draw the proper Gamma distribution, you can verify that for any value \( 0 < \beta \leq 1 \) we get a reasonable approximation.
3. Binomial and Multinomial models. Suppose data \((y_1, \ldots, y_J)\) follow a multinomial distribution with parameters \((\theta_1, \ldots, \theta_J)\). Also suppose that \(\theta = (\theta_1, \ldots, \theta_J)\) has a Dirichlet distribution.

\[
P(\theta) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_J) \theta_1^{\alpha_1-1} \ldots \theta_J^{\alpha_J-1}}{\prod_{j=1}^{J} \Gamma(\alpha_i)}
\]

(a) Derive the posterior \(p(\theta \mid y)\)

**Solution:**

\[
P(\theta \mid y) \propto \prod_{j=1}^{J} \theta_j^{y_j \alpha_j - 1} \prod_{j=1}^{J} \theta_j^{\alpha_j} = \prod_{j=1}^{J} \theta_j^{y_j + \alpha_j - 1}
\]

which is a Dirichlet\((\alpha_1 + y_1, \ldots, \alpha_J + y_J)\) distribution.

(b) Let \(\delta = \frac{\theta_1}{\theta_1 + \theta_2}\). Derive the marginal posterior for \(\delta\).

**Solution:**

This can be solved rather cumbersomely using transformation of variables and then integration. However, we can also use the known property of the Dirichlet distribution that the marginal of \((\theta_1, \theta_2, 1 - \theta_1 - \theta_2)\) is also Dirichlet with parameters \((\alpha_1, \alpha_2, \alpha_{rest} - 1)\) where \(\alpha_{rest}\) is the sum of all \(\alpha\)'s except the first two (this can be proceed via induction - sorry for not giving you this property and making your life needlessly difficult). Thus,

\[
P(\theta_1, \theta_2 \mid y) = \theta_1^{\alpha_1 + y_1 - 1} \theta_2^{\alpha_2 + y_2 - 1} (1 - \theta_1 - \theta_2)^{\alpha_{rest} + y_{rest} - 1}
\]

Now, do a change of variables to \((\alpha, \beta) = \left(\frac{\theta_1}{\theta_1 + \theta_2}, \theta_1 + \theta_2\right)\) so that \(\theta_1 = \alpha \beta\) and \(\theta_2 = (1 - \alpha) \beta\). The determinant of the Jacobian of this transformation is \(\beta\) and so

\[
P(\alpha, \beta \mid y) \propto \beta^{(\alpha_1 + y_1 - 1)(1 - \alpha) \beta^{\alpha_2 + y_2 - 1}(1 - \beta)^{\alpha_{rest} + y_{rest} - 1}}
\]

\[
= \alpha^{\alpha_1 + y_1 - 1}(1 - \alpha)^{\alpha_2 + y_2 - 1} \beta^{(1 - \beta) \alpha_{rest} + y_{rest} - 1}
\]

\[
\propto \text{Beta}(\alpha \mid y_1 + \alpha_1, y_2 + \alpha_2)
\]

\[
\times \text{Beta}(\beta \mid y_1 + y_2 + \alpha_1 + \alpha_2, y_{rest} + \alpha_{rest})
\]

From here, since the distribution decomposes for \(\alpha\) and \(\beta\), we have \(\delta \mid y \equiv \alpha \mid y \sim \text{Beta}(\alpha \mid y_1 + \alpha_1, y_2 + \alpha_2)\)

(c) Show that \(p(\delta \mid y)\) is identical to the posterior distribution for \(\delta\) obtained by treating \(y_1\) as an observation from the binomial distribution with probability \(\delta\) and sample size \(y_1 + y_2\), ignoring the data \(y_2, \ldots, y_J\).

**Solution:**

We have already seen that using a \(\text{Beta}(\alpha, \beta)\) prior distribution and a binomial observation \(y_1\) with sample size \(y_1 + y_2\) gives us the same posterior as the one we derived above.
This result justifies the application of the binomial distribution to multinomial problems when we are only interested in two of the categories. For example, see the next problem.

4. **Comparison of two multinomial observations.** On September 25, 1988, the evening of a Presidential campaign debate, ABC News conducted a survey of registered voters in the United states; 639 person were polled before the debate and 639 different persons were polled after. The results are displayed in the following table:

<table>
<thead>
<tr>
<th>Survey</th>
<th>Bush</th>
<th>Dukakis</th>
<th>No opinion/other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-debate</td>
<td>294</td>
<td>307</td>
<td>38</td>
<td>639</td>
</tr>
<tr>
<td>post-debate</td>
<td>288</td>
<td>322</td>
<td>19</td>
<td>639</td>
</tr>
</tbody>
</table>

Assume that both surveys are independent random samples from the population of registered voters. Model the data with two different multinomial distributions. For \( j = 1, 2 \), let \( \alpha_j \) be the proportion of voters who preferred Bush, out of those who had a preference for either Bush or Dukakis at the time of survey \( j \). Plot a histogram of the posterior density \( \alpha_2 - \alpha_1 \). What is the posterior probability that there was a shift toward Bush?

**Solution:**

Assuming independent prior on the multinomial parameters, the posterior distributions are then independent multinomial:

\[
(\theta_1, \theta_2, \theta_3) \mid y \sim \text{Dirichlet}(295, 308, 39)\\
(\phi_1, \phi_2, \phi_3) \mid y \sim \text{Dirichlet}(289, 333, 20)
\]

and \( \alpha_1 = \frac{\theta_1}{\theta_1 + \theta_2} \), \( \alpha_2 = \frac{\phi_1}{\phi_1 + \phi_2} \). Using the previous question we then have

\[
\alpha_1 \mid y \sim \text{Beta}(295, 308)\\
\alpha_2 \mid y \sim \text{Beta}(289, 333)
\]

We can now easily sample 1000 draws of \( \alpha_1 \) and \( \alpha_2 \) and construct a posterior distribution \( \alpha_2 - \alpha_1 \) from the samples. A histogram of this distribution is shown below. Using these samples we estimate that the probability of a shift toward Bush is around 20%.
5. **Binomial with unknown probability and sample size.** Consider data \(y_1, \ldots, y_n\) modeled as iid \(\text{Bin}(N, \theta)\), with both \(N\) and \(\theta\) unknown. Defining a convenient family of prior distributions on \((N, \theta)\) is difficult, partly because of the discreteness of \(N\).

Raftery (1988) considers a hierarchical approach based on assigning the parameter \(N\) a Poisson distribution with unknown mean \(\mu\). To define a prior distribution over \((N, \theta)\), Raftery defines \(\lambda = \mu \theta\) and specifies a prior distribution on \((\lambda, \theta)\). This is done because 'it would seem easier to formulate prior information about \(\lambda\), the unconditional expectation of the observations, than about \(\mu\), the mean of the unobserved quantity \(N\').

**Solution:**
Rather than write out a solution here, I decided to put on the website Raftery’s original paper in the hope that your curiosity will drive you to glance at it. With the tools we already learned, this paper is relatively straightforward to read.

(a) A suggested non-informative prior distribution if \(p(\lambda, \theta) \propto \lambda^{-1}\). What is a motivation for this non-informative distribution? Is the distribution improper? Transform to determine \(p(N, \theta)\).

(b) The approach is applied to counts of water-bucks obtained by remote photography on five separate days in Kruger Park in South Africa. The counts were 53, 57, 66, 67, 72. Perform the Bayesian analysis on these data and display a scatter-plot of posterior simulations of \((N, \theta)\). What is the posterior probability that \(N > 100\).

(c) Why not simply use a Poisson prior distribution with a fixed \(\mu\) as a prior distribution for \(N\)?