MIDTERM EXAM

ISyE 8843: Bayesian Statistics
Brani Vidakovic
Friday, 10/15/2004.

Name ____

1. Nematodes. Some varieties of nematodes (roundworms that live in the soil and are frequently so small they are invisible to the naked eye) feed on the roots of lawn grasses and crops such as strawberries and tomatoes. This pest, which is particularly troublesome in warm climates, can be treated by the application of nematocides. However, because of size of the worms, it is very difficult to measure the effectiveness of these pesticides directly. To compare three brands of nematocides (A, B, and C), the yields of equal-size plots of one variety of tomatoes were collected. The data (yields in pounds per plot) are shown in the table.

Nematocide A	25.8	19.8	28.6	29.4	22.3	33.8	33.8	27.8	29.6					
Nematocide B	16.5	23.5	13.5	34.6	16.9	18.8	26.1	18.4	17.2	11.6	20.2			
Nematocide C	24.0	29.1	16.0	24.8	27.0	10.9	11.8	23.2	17.7	23.9	24.6	24.0	27.2	23.7

Assume the general one way ANOVA model,

$$y_{ij} = \mu_i + \epsilon_{ij},$$

where $i = 1, \ldots, k$, and $k = 3$ (3 treatments). Within each treatment, $j = 1, \ldots, n_i$, ($n_1 = 9, n_2 = 11$, and $n_3 = 14$). Total sample size is $n = n_1 + n_2 + n_3 = 34$.

Complete this ANOVA model as follows. All (hyper)parameters with index 0 are specified numbers.

$$\begin{aligned}
y_{ij} &= \mu_i + \epsilon_{ij} \\
\left[\epsilon_{ij}|\sigma_i^2\right] &\sim \mathcal{N}(0, \sigma_i^2),\ i = 1, \ldots, k;\ j = 1, \ldots, n_i. \\
\left[\mu_i|\psi, \tau^2\right] &\sim \mathcal{N}(\psi, \tau^2) \\
\left[\psi|\tau^2\right] &\sim \mathcal{N}(\psi_0, \frac{\tau^2}{\zeta_0}) \\
\left[\sigma_i^2|\sigma^2\right] &\sim \mathcal{IG}amma(\frac{a_0}{2}, \frac{a_0\sigma^2}{2}) \\
\left[\tau^2\right] &\sim \mathcal{IG}amma(\frac{c_0}{2}, \frac{d_0}{2}) \\
\left[\sigma^2\right] &\sim \mathcal{G}amma(\frac{f_0}{2}, \frac{g_0}{2}).
\end{aligned}$$

Assume $a_0 = 1, c_0 = 1, d_0 = 1, f_0 = 1, g_0 = 0.1, \psi_0 = 10$, and $\zeta_0 = 0.1$. Denote by $\boldsymbol{\theta}$ the vector of all parameters in the model,

$$\boldsymbol{\theta} = (\mu_1, \mu_2, \ldots, \mu_k, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2, \sigma^2, \psi, \tau^2),$$

and by $\boldsymbol{\theta}\backslash\{\theta_i\}$ all parameters in $\boldsymbol{\theta}$ except θ_i, where θ_i is one of the parameters in $\boldsymbol{\theta}$. Denote by $\boldsymbol{y}$ all y_{ij}'s. Let $\bar{y}_i$ and s_i^2 be ith class sample means and variances, $\bar{y}_i = \frac{1}{n_i}\sum_j y_{ij}, s_i^2 = \frac{1}{n_i - 1}\sum_j (y_{ij} - \bar{y}_i)^2$. Then the full conditionals are:

$$
\begin{aligned}
[\mu_i|\boldsymbol{\theta}\backslash\{\mu_i\},\boldsymbol{y}] &\sim \mathcal{N}\left(\frac{\frac{n_i\bar{y}_i}{\sigma_i^2}+\frac{\psi}{\tau^2}}{\frac{n_i}{\sigma_i^2}+\frac{1}{\tau^2}},\frac{1}{\frac{n_i}{\sigma_i^2}+\frac{1}{\tau^2}}\right),\ i=1,\ldots,k \\
[\psi|\boldsymbol{\theta}\backslash\{\psi\},\boldsymbol{y}] &\sim \mathcal{N}\left(\frac{\sum_i\mu_i+\zeta_0\psi_0}{k+\zeta_0},\frac{\tau^2}{k+\zeta_0}\right) \\
\left[\tau^2|\boldsymbol{\theta}\backslash\{\tau^2\},\boldsymbol{y}\right] &\sim \mathcal{IG}amma\left(\frac{c_0+k+1}{2},\frac{d_0+\sum_{i=1}^k(\mu_i-\psi)^2+\zeta_0(\psi-\psi_0)^2}{2}\right) \\
\left[\sigma_i^2|\boldsymbol{\theta}\backslash\{\sigma_i^2\},\boldsymbol{y}\right] &\sim \mathcal{IG}amma\left(\frac{a_0+n_i}{2},\frac{a_0\sigma^2+(n_i-1)s_i^2+n_i(\bar{y}_i-\mu_i)^2}{2}\right),\ i=1,\ldots,k \\
\left[\sigma^2|\boldsymbol{\theta}\backslash\{\sigma^2\},\boldsymbol{y}\right] &\sim \mathcal{G}amma\left(\frac{f_0+ka_0}{2},\frac{g_0+a_0\sum_{i=1}^k\sigma_i^{-2}}{2}\right).
\end{aligned}
$$

Program this example in MATLAB or R and run the MCMC with suggested starting values. After discarding 1000 runs as burn-in, simulate 5000 runs and plot histograms of all parameters (for which the full conditionals are given). Give the posterior sample means and variances for all the parameters.

2. Rainfall Data in Marquitia. The data in the table below consists of daily maximums of rainfall data measured at the airport Marquietia, near Caracas, Venezuela in the period 1951-1999. Assume that your data consists of years 1951-1998 and that the rainfall in 1999 is going to be predicted. The value of maximum rainfall in 1999 is not a typo.[1]

Year	1951	52	53	54	55	56	57	58	59	1960	61	62	
DayMax	154	49.6	46.7	58.3	70.5	90	70.1	105.7	37.4	40.8	34.7	58.9	
Year	1963	64	65	66	67	68	69	1970	71	72	73	74	
DayMax	72.2	30	71.6	100	33.7	49.9	56.1	142.3	28.6	54.8	74.1	60	
Year	1975	76	77	78	79	1980	81	82	83	84	85	86	
DayMax	50.9	38.6	53.4	132.5	50.7	40.8	84.3	38.8	27.4	67	118.7	23.2	
Year	1987	88	89	1990	91	92	93	94	95	96	97	98	1999
DayMax	55	67.9	87.3	89	98.7	47.1	71.6	83.6	44.3	41.2	35.9	44.3	410.4

Assume the Gumbel distribution for the maximum rainfall,

$$F(y|\mu,\sigma) = \exp\left\{-\exp\left\{-\frac{y-\mu}{\sigma}\right\}\right\},$$

with density

$$f(y|\mu,\sigma) = \frac{1}{\sigma} \times \exp\left\{-\frac{y-\mu}{\sigma}\right\} \times \exp\left\{-\exp\left\{-\frac{y-\mu}{\sigma}\right\}\right\}.$$

The parameter $\theta = (\mu, \sigma)$ is given independent componentwise priors. Assume that the prior on μ is $\mathcal{N}(0, 10^2)$ and on σ lognormal $\mathcal{LN}(0, 10^2)$ (See handout 0).

(a) Express the posterior $\pi(\mu, \sigma|y_1, \ldots, y_n)$ up to a normalizing constant.

(b) Develop Metropolis-Hastings algorithm that samples posterior values of $\theta = (\mu, \sigma)$ in a form of Markov chain. Experiment to properly tune the algorithm.

(c) Use the output to estimate the probability $P(y^* \geq 410|y)$ where y^* is a new observation. This can be done in several ways. Of course, condition on data y from 1951 till 1998.

For an extensive Bayesian treatment, see

- Coles S., Pericchi L. R., Sisson S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, **273**, 35–50.
- Coles S., Pericchi L. R. (2003). Anticipating catastrophes through extreme value modelling. *Appl. Statist.*, **52**, 405–416.
- Data sets (annual maximum daily rainfall and daily rainfall levels recorded at Maiquetia) can be found at `http://www.blackwellpublishing.com/rss/Readmefiles/Coles.htm` .

[1] On 15 and 16 December, 1999, exceptional rainfall struck the northern coast of Venezuela. This strip of land between the El Avila mountain range and the sea has been considerably urbanized in the past thirty years, in particular on the alluvial cones produced by the torrents that rush down the slopes of the mountains. The equivalent of a year's average rainfall for the region - 900 mm - fell in two days on ground that was already saturated by the heavy rains of the preceding months. The torrents, more or less perpendicular to the coastline, swelled and washed away everything in their path, then dumped tons of sediment, blocks of stone, and vegetation debris from the slopes on the urbanisations located downstream. The damage was great both on the human level (some 10,000 deaths) and as regards property and economic impact.

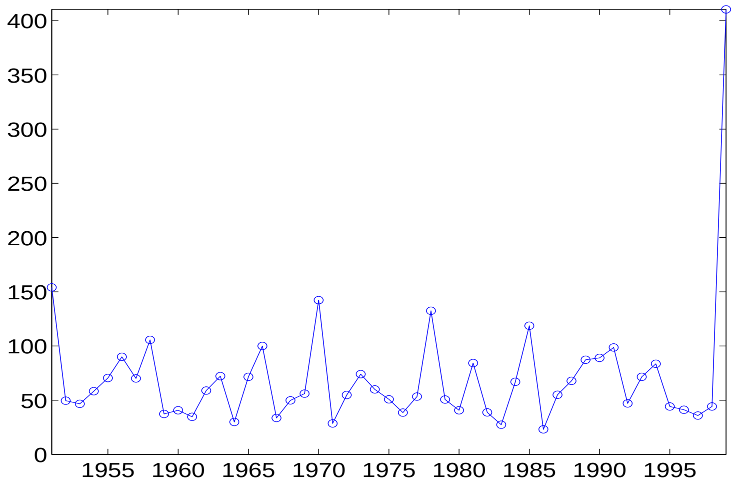

Figure 1: Marquietia Rainfall Data: Yearly maximum of daily rainfall.

3. A Baby Bayes NonParametric via BUGS. The underground train at Hartsfield-Jackson airport is arriving at the starting station [baggage claim] every four minutes. The number of people Y entering the train is random variable with Poisson distribution,

$$[Y|\lambda] \sim \mathcal{P}oi(\lambda).$$

The prior on λ is any discrete distribution supported on integers [1, 17],

$$[\lambda|P] \sim \mathcal{D}iscr\left((1, 2, \dots, 17), P = (p_1, p_2, \dots, p_{17})\right),$$

where $\sum_i p_i = 1$. The hyperprior on probabilities P for is Dirichlet,

$$[P] \sim \mathcal{D}ir(\alpha_1, \alpha_2, \dots, \alpha_{17}).$$

This is a baby Bayesian nonparametric problem. We are interested in posterior inference on λ.

Check for ODC file [model, data and inits folded] titled `midt3.odc` here where the exam is.

Import the simulations to MATLAB or R and plot histograms of posterior distributions for λ, $p_2, p_7, p_9, p_{10}, p_{13}$, and p_{17}. Provide all posterior statistics. You may consult Handout 14.

```
#model
model
{
for (i in 1:N)
    {
     y[i] ˜ dpois(lambda)
    }
    lambda ˜ dcat(P[])
    P[1:bins] ˜ ddirch(hyper[])
}
#data
list(
bins=17,
hyper=c(1,1,1,2,2,3,3,4,4,5,6,5,4,3,2,1,1),
 y=c(9,7,7,8,8,11,8,7,5,7,13,5,7,14,4,6,18,9,8,10),  N=20
)

#inits
list(
lambda=12,
P=c(0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0)
)
```