

KEY - Second Mid-Term - April 22, 2005

Read all the questions carefully and start answering the ones for which you know the answer. None of the questions require extensive calculations or derivations. Be precise. Show all your work; partial credit will be given. Each part of a problem is worth 10 points (except for parts 2.b, 2.c and 2.d which are worth 3.33 points), for a total of 100 points. Good luck!

Problem 1

In this problem, you are presented with various examples from Congdon (2001), somewhat modified. In each case, you are asked to write down the appropriate sampling model and possible prior and hyperprior distributions, and also to describe (in detail!) how you would go about providing researchers with the answers they seek. Read the problem descriptions carefully and justify all your choices.

(1.1) Baseball scores

Morris and Christiansen (1996) present data on runs per game Y_i for 14 American Baseball League teams in 1993. Detroit led that season with an average of 5.549 runs per game. The variances V_i of runs per game for each team were computed for that season from the $y_{ij} = 162$ games played during the season. The data available for analysis are therefore the 14 averages Y_i and the 14 standard deviations $V_i^{1/2}$. The Y_i range from 0.76 to 5.549.

(1.1.a) Write down an exchangeable model for the Y_i , assuming that the average runs per game for a team in 1993 is generated by a population distribution for the team with mean equal to the usual (over all seasons) average runs per game for that team. Further, assume that the usual team averages are exchangeable draws from a distribution with mean representing the average runs over all teams and all seasons. Propose hyperpriors for other parameters.

Answer: The outcomes are continuous average runs per game and therefore a normal sampling model is justifiable as a first choice. An exchangeable model is

$$\begin{aligned} Y_i &\sim N(\theta_i, V_i) \\ \theta_i &\sim N(\mu, \tau^2), \end{aligned}$$

where $i = 1, \dots, 14$ and the V_i are known. Priors for μ, τ^2 could be chosen to be uniform and inverted Gamma with very low precision, for example.

(1.1.b) In detail, describe how you would estimate the probability that each team is best regarding average runs per game.

Answer: Suppose that we use the Gibbs sampler to obtain M draws of $\theta_1, \dots, \theta_{14}$ as well as of μ and τ^2 . In each draw, we rank the θ_i from smallest to largest. The team that is 'best' in the draw is the one with the largest θ_i . We then count to see how many times (out of M) each team was best, meaning, had the highest θ_i .

We could also draw values \tilde{y}_{ij} , for $j = 1, \dots, N$ to simulate N games for the i th team and then rank the teams according to the mean of the \tilde{y}_{ij} over the N draws and M parameter values, but this is a lot more work and would not result in different rankings (except for numerical reasons).

(1.1.c) You now wish to forecast team performance in the first 10 games of the 1994 season. In detail, describe how you would approach this prediction problem.

Answer: This is a prediction problem. We sample 10 values from the posterior predictive distribution of each team. For team i , we proceed as follows:

- a- We have M draws of θ_i from $p(\theta_i|y)$ obtained in part 1.1.b.
- b- Get 10 draws $Y.pred_{ij} \sim N(\theta_i, 10V_i)$.

There is a bit of hand-waiving in the last step. The variances V_i that we assume known were computed from an entire season that includes 162 games. Thus, it is likely to underestimate the variance of games in a much shorter season of only 10 games.

(1.2) Conduct disorder

Johnson and Albert (1999) consider the number of times, out of four total observations on each of 172 school children, that a student exhibited conduct disorder. For each child, conduct was assessed at grades 6, 8, 10 and 12. Child-level covariates included sex (1 = females, 0 = males), aggressive behavior observed while in 3rd grade ($A_i = 1$ if child was aggressive, 0 otherwise), and social rejection as rated in 3rd grade ($R_i = 1$ if rejected,

0 otherwise). Researchers wish to know whether the probability of conduct disorder is associated to the covariates.

(1.2.a) Write down the appropriate sampling distribution, as well as the exchangeable population distribution for the model parameter(s). Include all prior distributions as well, and explain your choices.

Answer: In each of four occasions, a child was classified as either well behaved or misbehaved. Thus, observation was $Y_{ij} = 0$ if the i th child did not exhibit conduct disorder during the j th grade and was $Y_{ij} = 1$ otherwise. We let Y_i denote the sum of Y_{ij} for the i th child. An appropriate sampling distribution is

$$y_i \sim \text{Bin}(\theta_i, n),$$

with $n = 4$. Given covariates, children are exchangeable, so we can formulate a logit model for the probabilities of conduct disorder θ_i as follows:

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 S_i + \beta_2 A_i + \beta_3 R_i,$$

where $\text{logit}(\theta_i) = \log(\theta_i/(1 - \theta_i))$, S is a dummy for sex, A is a dummy for the aggression covariate and R is the dummy for the rejection covariate. The regression coefficients β_0, \dots, β_3 are assumed to have a flat prior distribution or perhaps a normal distribution with very low precision.

(1.2.b) Researchers suspect that the model may not appropriately account for overdispersion in the observed data. Reformulate your model to incorporate a parameter to represent overdispersion and explain what this new parameter represents.

Answer: The linear model for $\text{logit}(\theta)$ can be augmented by adding a child-level random term e_i , which itself gets a population distribution

$$e_i \sim N(0, \sigma^2).$$

The child-level random effect accounts for the fact that repeated observations for the same child may be correlated. A prior for the variance component σ^2 must also be incorporated into the model. A safe choice is a diffuse inverted gamma prior.

(1.2.c) You wish to know which of the two models fits the data best. Explain how you would go about comparing the overall fit of the models. Choose a statistic and describe the steps you would follow to compute the statistic. Explain how, on the basis of your statistic, you would opt for one model or the other.

Answer: We might compute a global model fit measure such as the χ^2 statistic or the deviance information criterion DIC.

The χ^2 discrepancy is computed as

$$T(y, \theta) = \sum \frac{(y_i - E(y_i|\theta))^2}{\text{var}(y_i|\theta)}.$$

First, we compute the distribution of the discrepancy measure using the observed data and the draws of the parameters. We then replicate the data and recompute the distribution of the discrepancy measure over the replicates and over the parameter draws. We choose the model which has the most central Bayes p-value, i.e., the one for which the distributions of T in the observed and replicated datasets agree better.

An alternative is the Deviance Information Criterion or DIC

$$DIC = \hat{D}_{\text{avg}}^{\text{pred}} = 2\hat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y),$$

where $D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$ and $\hat{\theta}(y)$ is a point estimator of θ such as the posterior mean. The 'best' model is the one with smaller DIC.

(1.3) SIDS deaths in North Carolina

Cressie and Read (1989) present an analysis of SIDS (Sudden Infant Death Syndrome) deaths S_i in 100 counties in North Carolina. The number of total births in each county B_i is also known. One possible covariate believed to influence SIDS occurrences is the proportion of births to non-white mothers in each county (x_i). Researchers wish to know whether the covariate is in fact associated to the SIDS rate and also wish to obtain county-level estimates of SIDS rates.

(1.3.a) Write down an appropriate model for these data. Include the sampling distribution and all priors and hyperpriors. In English, interpret what each parameter represents.

Answer: The number of SIDS deaths can be modeled as Poisson, and the total number of births is used as an exposure. Then

$$y_i \sim \text{Poisson}(\lambda_i B_i),$$

where λ_i is the predicted SIDS rate in the i th county. We introduce the covariate in the second level:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i + \delta_i,$$

where the term $\delta_i \sim N(0, \sigma_\delta^2)$ is a random effect that accounts for potential overdispersion of log rates across counties. The regression coefficients (β_0, β_1) are unknown and are assigned flat priors (or normal priors with very large variance). An inverted Gamma prior can be used for σ_δ^2 .

(1.3.b) There is some indication of a spatial correlation in SIDS deaths in North Carolina. Extend your model by adding a spatial structure that would allow you to determine whether county-level SIDS death rates cluster or whether they are spatially uncorrelated. Describe the new parameters in your model and explain how you would interpret your results.

Answer: Because the number of SIDS deaths is an aggregate count in each county, we use the methods described for areal data to incorporate spatial correlation. Here, areas are counties. We first need to define the neighbors of each area (county) and one way to do so is to just consider adjacent counties to be the neighborhood. If county j is adjacent to county i , then $w_{ij} = 1$. Otherwise, $w_{ij} = 0$.

We add a county-level random effect to the second model tier:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i + \delta_i + \phi_i,$$

where ϕ_i capture clumping. We use a conditionally autoregressive prior on ϕ of the form

$$\phi_i \sim N(\bar{\phi}_i, 1/(\tau_e m_i)),$$

where

$$\bar{\phi}_i = \sum_{i \neq j} w_{ij} (\phi_i - \phi_j),$$

and m_i is the number of neighbors of area i . The relative sizes of the variances of δ and ϕ indicate how much of the variability in the SIDS death numbers across the state are due to global dispersion or to regional clustering. In this particular example, we might expect that inclusion of the covariate will already account for much of the regional clustering since the proportion of babies born to black mothers is likely to be geographically clustered (e.g., more of them in poor areas of the state).

Problem 2

Data on language scores was collected in 131 elementary schools (S) for 2287 students in two different grades. In each school a single class is observed. We are interested on the impact on language scores (LS) of student level factors such as IQ, and students social economic status (SES). The gender (G) of the students was also recorded (1 for girls, 0 for boys) as well as the class mean IQ (Z).

WinBUGS was used to fit the following model¹:

```

model {
  for (i in 1:2287) {LS[i] ~ dnorm(mu[i], T[i])

  log(T[i]) <- c[1]+c[2]*Gender[i]

  mu[i] <- b[S[i],1]+ b[S[i],2]*(IQ[i]-mean(IQ[]))+
          d[1]*(ses[i]-mean(ses[]))+ d[2]*G[i]
          }

  for (j in 1:131){b[j,1:2] ~ dmnorm(mu.b[j,1:2],Q[1:2,1:2])

  for (k in 1:2){mu.b[j,k] <- gam[1,k] + gam[2,k]*(Z[j]-mean(Z[]))}

  Q[1:2,1:2] ~ dwish(T.b[,,],2)
  V.b[1:2,1:2]<- inverse(Q[,,])
  for (i in 1:2){ for (j in 1:2){T.b[i,j] <- equals(i,j);
          C[i,j] <- V.b[i,j])
          R[i,j] <- C[i,j]/sqrt(C[i,i]*C[j,j])
          }
          }

  V.gend[1] <- 1/exp(c[1])
  V.gend[2] <- 1/exp(c[1]+c[2])
  V.diff <- step(V.gend[1]-V.gend[2])

  for (k in 1:2) {c[k] ~ dnorm(0,1)
          g[k] ~ dnorm(0,1)
          gam[1,k] ~ dnorm(0,0.0000001)
          gam[2,k] ~ dnorm(0,0.001)}}

```

a- Write down the likelihood, priors, and hyperpriors.

Answer: For $i = 1, \dots, n_j$, $j = 1, \dots, 131$, and $k = 1, 2$

¹the function `equals(i,j)` takes the value 1 if $i = j$, 0 otherwise.

$$y_{ij} \sim N(\mu_{ij}, \phi_{ij})$$

$$\mu_{ij} = \beta_{1j} + \beta_{2j}(\text{IQ}_{ij} - \bar{\text{IQ}}) + \delta_1(\text{SES}_{ij} - \bar{\text{SES}}) + \delta_2 G_{ij}$$

$$\beta_{pj} \sim N_2(\nu_j, V_\beta) \quad \text{for } p = 1, 2$$

$$\delta_p \sim N(0, 1) \quad \text{for } p = 1, 2$$

$$V_\beta \sim \text{Inv-Wishart}_2(\mathbf{I}) \quad \text{with } \mathbf{I} \text{ the } 2 \times 2 \text{ identity matrix}$$

$$\nu_{kj} = \gamma_{1k} + \gamma_{2k}(Z_{1j} - \bar{Z}_1)$$

$$\log(\phi_{ij}) = c_1 + c_2 G_{ij}$$

$$c_k \sim N(0, 1)$$

$$\gamma_{1k} \sim N(0, 100000000)$$

$$\gamma_{2k} \sim N(0, 1000)$$

- b- What does $R[1,2]$ measure in the above program? How will you interpret the fact that the posterior mean of $R[1,2]$ was equal to -0.55?

Answer: $R[1,2]$ measures the correlation between intercepts and IQ slopes. Because the mean of the correlation is -0.55, we infer that individual IQ had a higher impact on LS in classes with lower than average attainment.

- c- Suppose that we have run three chains and we obtained the following results based on 5000 iterations after a 500 burn-in period

	Mean	Std.	2.5%	97.5%
c_1	-3.650	0.040	-3.740	-3.570
c_2	0.064	0.064	-0.061	0.191

Interpret the obtained posterior distribution of c_2 in the context of the problem.

Answer: The 95% credible set for c_2 suggests that girls have higher precision (and hence lower variance) in their language scores. But the credible set includes zero which may throw doubt on a clear difference in variance between girls and boys.

- d- Which quantity should we keep track if we wish to test whether the variance of the boys exceeds that of girls?

Answer: We should keep track of $V.\text{diff}$.