Kamaru-Deen Lawal

Homework 3

HW 3

(4.3) Let $\theta = LD50$ and $\upsilon = \beta$ (since we will use a change of coordinates). We know that the posterior mean and median should overlap. We also know that the posterior standard variance should be the inverse of the observed information, evaluated at the posterior mode. Note that $p(\theta|y) = \int p(\theta, \upsilon|y)\,d\upsilon$

$$= \int p(\alpha, \beta | y)\, |v|\, d\upsilon.$$

where $\alpha = -\theta v$ and $\beta = \upsilon$, $|v|$ is the jacobian associated with the change of coordinates we mentioned. From here we fix θ to be $-.11$, which is the approximate value of the posterior mode. From here we can just compute the last integral numerically. The region of integration is infinite, but decreases rapidly enough where we can approximate the value. Repeat this procedure for values near the posterior mode. Which leads to a resultant value of $-.115$. We can fix a small value of h, such as $h = .002$ to compute $\frac{d^2}{d\theta^2} \log p(\theta|y)$, evaluated at θ equal to the posterior mode, by the expression $[\log p(-0.115 + h|y) - 2\log p(-.0115|y) + \log p(-.115 - h|y)]/h^2$

Negating the preceding quantity, and putting it to the $-.5$ power gives us our estimate for the posterior standard deviation.

4.4 As $n \to \infty$, the posterior variance approaches 0. That is, the posterior distribution becomes concentrated near a single point. Any 1-to-1 continuous transformation on the real numbers is locally linear in the nbhd of that point.

5.3 (a) Based on the 1000 posterior simulations, we obtain

School	Pr(best)	A	B	C	D	E	F	G	H
A	.25	X	.64	.67	.66	.73	.70	.83	.61
B	.10	.36	X	.55	.53	.62	.61	.37	.49
C	.10	.33	.45	X	.46	.58	.53	.36	.45
D	.09	.34	.47	.54	X	.61	.58	.37	.47
E	.05	.27	.38	.42	.39	X	.48	.28	.38
F	.08	.31	.39	.47	.42	.82	X	.31	.40
G	.21	.47	.63	.64	.63	~~.63~~ .72	.69	X	.60

(b) In the model w/ τ set to ∞, the school effects θ_j are independent in their posterior distribution w/ $\theta_j | y \sim N(y_j, \sigma_j^2)$. Follows that $P(\theta_i > \theta_j | y) = \Phi\left(\frac{(y_i - y_j)}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right)$

The prob that θ_i is the largest of the school effects can be expressed as a single integral is given ~~below~~ here. $P(\theta_i \text{ is the largest}) = \int_{-\infty}^{\infty} \prod_{j \neq i} \Phi\left(\frac{\theta_i - y_j}{\sigma_j}\right) \cdot \phi(\theta_i | y_i, \sigma_i) d\theta_i$

We can evaluate this integral numerically. The results can be seen on the table on the next page.

5.3) (cont.)

(b).

School	Pr(best)	A	B	C	D	E	F	G	H
A	.556	X	.87	.92	.88	.95	.93	.72	.76
B	.034	.13	X	.71	.53	.73	.68	.24	.42
C	.028	.08	.29	X	.31	.46	.43	.14	.27
D	.034	.12	.47	.69	X	.70	.65	.23	.40
E	.004	.05	.27	.54	.30	X	.47	.09	.26
F	.013	.07	.32	.57	.35	.53	X	.13	.29
G	.170	.28	.76	.86	.77	.91	.87	X	.61
H	.162	.24	.58	.73	.60	.74	.71	.39	X

(c) The model w/ τ set to ∞ has more extreme probabilities. In the 1st column the prob that School A is the best increases from .25 to .56. This is also the case in the pairwise comparisons. For example, the prob that School A's prog is > school E's under the full hierarchical model is .73, while it is .95 under the $\tau = \infty$ model. The more conservative answer under the full hierarchical model highlights the evidence in the data that the coaching programs appear fairly similar in effectiveness. Preferred school in a pair can change depending on posterior distribution of τ. This occurs, because the standard errors differ.

(d) If $\tau = 0$, then all of the school effects are the same. Thus no school is better or worse than any other.

5.4 (a) Yes, they are exchangeable. The joint distribution is.

$$p(\theta_1, \ldots, \theta_{2J}) = \binom{2J}{J}^{-1} \sum_{p} \left(\prod_{j=1}^{J} N(\theta_{p(j)} \mid 1, 1) \prod_{j=J+1}^{2J} N(\theta_{p(j)} \mid -1, 1) \right)$$

where the sum is over all permutations p of $(1, \ldots, 2J)$. The density (7) is ~~obviously~~ invariant to permutations of the indexes $(1, \ldots, 2J)$.

(b) Pick any i, j. The covariance of θ_i, θ_j is negative. You can see this because if θ_i is large, then it is $\sim N(1,1)$ which means it is likely that $\theta_j \sim N(-1,1)$. (Since half of the parameters are assigned to each distribution). Therefore $\theta_j < 0$ with high probability. Same argument can be used when swapping θ_i and θ_j. So $p(\theta_1, \ldots, \theta_{2J})$ can not be written as a mixture of i.i.d components. We can formalize this argument by defining $\phi_1, \ldots, \phi_{2J}$, where half of the ϕ_j's are 1 and half are -1, and then setting $\theta_j \mid \phi_j \sim N(0,1)$. From here it is not difficult to show that $\text{cov}(\phi_i, \phi_j) < 0$, and then that $\text{cov}(\theta_i, \theta_j) < 0$

(c) As $J \to \infty$, the negative correlation between θ_i and θ_j approaches 0, and the joint distribution approaches i.i.d. Phrased more explicitly, as $J \to \infty$, the distinction between independently assigning each θ_j to one of two groups and picking exactly half of the θ_j's for each group vanishes.

5.12 $$E\left[\theta_j \mid \tau, y\right] = E\left[E\left[\theta_j \mid \mu, \tau, y\right] \mid \tau, y\right]$$

$$= E\left[\frac{\frac{1}{\sigma_j^2} y_j + \frac{1}{\tau_2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \;\middle|\; \tau, y\right]$$

$$= \boxed{\frac{\frac{1}{\sigma^2} y_j + \frac{1}{\tau_2}\hat{\mu}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}}.$$

$$\operatorname{var}\left[\theta_j \mid \tau, y\right] = E\left[\operatorname{var}(\theta_j \mid \mu, \tau, y) \mid \tau, y\right] + \operatorname{var}\left[E\left[\theta_j \mid \mu, \tau, y\right] \mid \tau, y\right].$$

$$= \boxed{\frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}\right)^2 V_\mu},$$

V_μ and $\hat{\mu}$ are defined as $\operatorname{var}(\mu \mid \tau, y)$ and $E[\mu \mid \tau, y]$ respectively.