

# APTS Statistical Modelling: Exercises

The work provided here is intended to take students up to half a week to complete. Students should talk to their supervisors to find out whether or not their department requires this work as part of any formal accreditation process (APTS itself has no resources to assess or certify students). It is anticipated that departments will decide on the appropriate level of assessment locally, and may choose to drop some (or indeed all) of the parts, accordingly.

0. If you have not already done so, complete the three APTS week practical sessions.
1. (a) Suppose that  $y = (y_1, \dots, y_n)$  are modelled as observations of i.i.d. random variables from  $f(y_i|\theta)$  and that the prior information about  $\theta = (\theta_1, \dots, \theta_p)$  is represented by the prior probability distribution  $\pi(\theta)$  (which does not depend on  $n$ ). Writing the *marginal likelihood*,  $p(y)$ , as

$$p(y) = \int f(y | \theta)\pi(\theta) d\theta = \int \exp\{-h(\theta)\} d\theta,$$

apply the Laplace approximation to the integral, and show that if  $O(1)$  terms are neglected, then

$$-2\log p(y) \doteq \text{BIC} = -2\log f(y | \hat{\theta}) + p \log n, \quad n \rightarrow \infty.$$

Hence, what can you say about Bayesian posterior model probabilities as  $n \rightarrow \infty$ ?

- (b) Show that AIC for a normal linear model with  $n$  responses,  $p$  covariates and unknown  $\sigma^2$  may be written as  $n \log \hat{\sigma}^2 + 2p$  where  $\hat{\sigma}^2 = RSS/n$  is the maximum likelihood estimate of  $\sigma^2$ . If  $\hat{\sigma}_0^2$  is the unbiased estimate under some fixed ‘correct’ model with  $q$  covariates, show that AIC is equivalent to using  $n \log \{1 + (\hat{\sigma}^2 - \hat{\sigma}_0^2)/\hat{\sigma}_0^2\} + 2p$  as a model comparison criterion, and that this is approximately equal to  $n(\hat{\sigma}^2/\hat{\sigma}_0^2 - 1) + 2p$ . Deduce that model selection using  $C_p$  approximates that using AIC.
  - (c) In the same context as (b), show that  $C_p = (q - p)(F - 1) + p$  where  $F$  is the  $F$ -statistic for comparison of the models with  $p$  and  $q > p$  covariates, and deduce that if the model with  $p$  covariates is correct then  $E(C_p) \doteq q$  but that otherwise  $E(C_p) > q$ .
2. The data frame `bacteria` are discussed in Chapter 10 of *Modern Applied Statistics with S (4th edition)* by Venables and Ripley (Springer, 2002). They are available in R by loading the library `MASS`. The response `y` indicates presence or absence of a particular bacteria when assessed on 50 individuals (`ID`) at each of up to 6 time points (`week`). Each individual has received one of three treatments (`trt`: `placebo`/`drug`/`drug+`).

Model the dependence of `y` on `trt` and `week` using binary GLMs and GLMMs (to account for intra-subject dependence in the response), fitted by maximum likelihood and associated approximations. Functions which you might wish to investigate for doing this include `glmPQL` (from the `MASS` library), `glmMML` (from the library of the same name) and `lmer` (from the `lme4` library). Use the library documentation provided to learn about the required arguments of these functions. Compare the inferences obtained by different fitting methods (quadrature, Laplace, PQL).

3. For the `bacteria` data, Venables and Ripley (2002, p297) propose the binary GLMM with

$$Y_{ij} \sim \text{Bernoulli}(\mu_{ij}), \quad g(\mu_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + b_{0i}, \quad b_{0i} \sim N(0, \sigma_b^2)$$

where  $X_1, X_2, X_3$  are the three binary explanatory variables  $I(\text{trt} = \text{drug})$ ,  $I(\text{trt} = \text{drug}+)$ , and  $I(\text{week} > 2)$  and  $g$  is the logit link function.

If instead  $g$  is the probit link ( $\Phi^{-1}$ ), then a Bayesian analysis of this model, using a Gibbs sampler, is straightforward, utilising the following latent variable formulation: The GLMM above (with  $g = \Phi^{-1}$ ) is equivalent to

$$Y_{ij} = I(Z_{ij} > 0), \quad Z_{ij} \sim N(\mu_{ij}, 1), \quad \mu_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + b_{0i}, \quad b_{0i} \sim N(0, \sigma_b^2)$$

where the  $Z_{ij}$  are latent continuous-valued variables, one for each observed  $Y_{ij}$ .

- (a) Establish this equivalence, that is, for the latent variable model, show that  $P(Y_{ij} = 1 | b_{0i}) = \Phi(\mu_{ij})$ .

As the  $Z_{ij}$  are unobserved, they can also be generated in any Gibbs sampler scheme. It is immediately obvious that, given  $Z$ , the conditional distributions for  $\beta, b, \sigma_b^2$  are exactly as for a LMM (with known error variance  $\sigma^2 = 1$ ). Hence, a Gibbs sampler for this GLMM can be obtained by a straightforward modification of our LMM Gibbs sampler from Practical 3. We simply need to generate the  $Z_{ij}$  at each step.

- (b) Show that the conditional distribution for  $Z_{ij} | Y, \beta, b, \sigma_b^2$  is  $N(\mu_{ij}, 1)$ , restricted to the range  $(0, \infty)$  when  $Y_{ij} = 1$ , or the range  $(-\infty, 0]$  when  $Y_{ij} = 0$ .
- (c) Modify the R function you used for an LMM Gibbs sampler in Practical 3, to perform a Bayesian analysis of the model above. Use the initial diffuse priors  $\beta_i \stackrel{\text{ind}}{\sim} N(0, 25)$  and  $\sigma_b^{-2} \sim \text{Gamma}(0.01, 0.01)$ . It is reasonable to suppose *a priori* that the probability of bacteria presence decreases over time. Perform an alternative analysis with the more informative prior distribution  $\beta_3 \sim N(-2, 4)$ . How are your results affected?
- (d) Compare your results with the logit model results obtained using maximum likelihood in Question 2. [Note that, if  $g_1$  and  $g_2$  are logit and probit links respectively, then linear approximation gives  $g_1 \approx 4g_2/(2\pi)^{1/2}$ .]