Supplementary material for Estimating bacterial pathogen concentrations in New Zealand bulk tank milk.

J.C. Marshall, T.K. Soboleva and N.P. French

May 27, 2015

1 Total Bacterial count model

1.1 Model formulation

We assume that total bacterial counts on farms come from a mixture of two Poisson distributions, one of which describes the low-level contamination inherent in milking systems, and the other the high-level fecal contamination introduced through one-off events such as the dropping of a milking cluster. We assume that each distribution is farm-dependent, and thus use a hierarchical model, where the mean of each distribution on each farm is a random effect (α_j, β_j) centered on some common means (μ_α, μ_β) across all farms for each of the low (α) and high (β) distributions, with precisions τ_α and τ_β respectively.

The low and high distributions are combined based on the expected rate w of high-level contamination events which we assume is constant across the farms. Thus the model takes the form

$$\begin{aligned} Y_{ij} &\sim \mathsf{Poisson}(\lambda_{ij}) \\ \log(\lambda_{ij}) &= (1 - z_{ij})\alpha_j + z_{ij}\beta_j \\ z_{ij} &\sim \mathsf{Bernoulli}(w) \\ \alpha_j &\sim \mathsf{Normal}(\mu_\alpha, \tau_\alpha) \\ \beta_j &\sim \mathsf{Normal}(\mu_\beta, \tau_\beta). \end{aligned}$$

1.2 Priors

We fit the model within a Bayesian context, by specifying uninformative priors on the hyper-parameters as follows:

$$\begin{aligned} w &\sim \mathsf{Bernoulli}(1,1) \\ \mu_\alpha, \mu_\beta &\sim \mathsf{Normal}(0, 0.001) \\ \tau_\alpha, \tau_\beta &\sim \mathsf{Gamma}(0.001, 0.001) \end{aligned}$$

1.3 MCMC algorithm

The MCMC algorithm for sampling from the posterior is as follows.

1. Gibbs sample per-count latent variables X_{ij} for whether count i is from the high or low distribution for farm j, given the parameters α_j, β_j, w using the likelihoods of the data under high and low.

$$\begin{aligned} p^\beta_{ij} &= w e^{\beta_j Y_{ij} - e^{\beta_j}} \\ p^\alpha_{ij} &= (1-w) e^{\alpha_j Y_{ij} - e^{\alpha_j}} \\ x_{ij} &\sim \mathsf{Bernoulli}(\frac{p^\beta_{ij}}{p^\alpha_{ij} + p^\beta_{ij}}) \end{aligned}$$

2. Sample α_j for each farm j given the latent variables $X_i j$ and parameters μ_α, τ_α using Metropolis-Hastings with a Normal proposal distribution.

3. Sample β_j for each farm j given the latent variables $X_i j$ and parameters μ_β, τ_β using Metropolis-Hastings with a Normal proposal distribution.

4. Propose a swap between α_j and β_j, and latent variables $X_i j$ given parameters w, μ_α, μ_β, τ_α, τ_β.

5. Gibbs sample $w \sim \mathsf{Beta}(1 + \sum_{i,j} X_{ij}, 1 + \sum_j N_j - \sum_{i,j} X_{ij})$ where N_j is the number of samples on farm j.

6. Gibbs sample the overall means μ_α and μ_β and precisions τ_α, τ_β:

$$\begin{aligned} \mu_\alpha &\sim \mathsf{Normal}(\frac{\tau_\alpha \sum \alpha_j}{\xi_{\mathsf{prec}} + N\tau_\alpha}, \xi_{\mathsf{prec}} + N\tau_\alpha) \\ \mu_\beta &\sim \mathsf{Normal}(\frac{\tau_\beta \sum \beta_j}{\xi_{\mathsf{prec}} + N\tau_\beta}, \xi_{\mathsf{prec}} + N\tau_\beta) \\ \tau_\alpha &\sim \mathsf{Gamma}(\xi_{\mathsf{shape}} + N/2, \xi_{\mathsf{rate}} + \sum (\alpha_j - \mu_\alpha)^2/2) \\ \tau_\beta &\sim \mathsf{Gamma}(\xi_{\mathsf{shape}} + N/2, \xi_{\mathsf{rate}} + \sum (\beta_j - \mu_\beta)^2/2) \end{aligned}$$

   where N is the number of farms, $\xi_{\mathsf{prec}} = 0.001$ is the prior precision on μ_α and μ_β, and $\xi_{\mathsf{shape}} = 0.001$, $\xi_{\mathsf{rate}} = 0.001$ are prior shape and rates for τ_α and τ_β.

The model was fit by running 5 chains of 10,000 iterations with disperse initial values sampled from the priors. The posterior distributions for α_j, β_j and w are used in the following section as uncertainty distributions.

2 Simulating pathogen counts

We assume that faecal contamination of bulk tank milk occurs via one of two processes.

1. The small amount of continuous contamination inherent in milking systems, likely originating from all animals on the farm.
2. From discrete contamination events that produce higher levels of contamination than is normal, more likely originating from a single cow.

We assume further that the proportion of high level events is given by the proportion of TBC counts from the high count distribution (w), and that we can model faecal contamination of milk (in g/l) as being proportional to TBC by noting that there are approximately $10^{10} - 10^{11}$ bacteria per gram of faeces. The amount of pathogen contamination in each of the processes above is then dependent on the proportion of positive farms p_f, the proportion of positive animals on farms p_a, and the shedding distribution f (in cfu/g) of positive animals.

In the case of process 1, we assume faecal contamination is sourced from a pooled sample from all animals on the farm, so that for a sufficiently large farm, we may utilise the central limit theorem to derive a normal distribution f_p for the number of cfu per gram of faecal contamination as follows. If we assume the animal counts have a log normal distribution conditional on being positive, the unconditional distribution for animal counts will be the product of a log-normal distribution with $\mathsf{Bernoulli}(p_a)$. The distribution of the pooled sample will then be the distribution of the mean of a product of log-normal and Bernoulli distributions. By the central limit theorem, this will be normal, with mean given by the mean of the product distribution, and variance equal to the variance of the product distribution divided by the farm size. We thus have

$$f_p \sim \mathsf{Normal}\left(p_a m_a, \frac{p_a v_a + p_a(1-p_a)m_a}{n}\right),$$

where $m_a = 10^{\mu_f + \frac{\sigma_a^2}{2}}$ and $v_a = \left(10^{\sigma_a^2} - 1\right) 10^{2\mu_a + \sigma_a^2}$ are the mean and variance of the log normal distribution f_a on the natural scale, p_a is the proportion of animals positive, and n is the number of cows on the farm.

In the case of process 2, we assume faecal contamination is sourced from a single animal on the farm, thus will contain pathogen only if the cow is positive.

We simulate the pathogen counts in milk (in cfu/ml) using the following process:

1. Sample $w, \alpha_j, \beta_j, p_f$ and p_a from their uncertainty distributions.

2. Sample the event type from $\mathsf{Bernoulli}(w)$.
3. If a low level event
   (a) Sample x from $\mathsf{Bernoulli}(p_f)$.
   (b) Sample y from f_p.
   (c) Sample t from a mixture of Poissons with means e^{α_j}.
4. If a high level event
   (a) Sample x from $\mathsf{Bernoulli}(p_f p_a)$.
   (b) Sample y from f.
   (c) Sample t from a mixture of Poissons with means e^{β_j}.
5. Sample b from the bacterial distribution $d \sim \mathsf{LogNormal}(10.5, 0.3)$.
6. $\frac{xyt}{b}$ is then a sample from the pathogen count per ml of milk.