Supplementary material for

Prevalence of bovine viral diarrhoea in Scottish beef suckler herds

F. Brülisauer[a,*], F. I. Lewis[a], A. G. Ganser[a], I. J. McKendrick[b], G. J. Gunn[a]

[a] *Epidemiology Research Unit, SAC (Scottish Agricultural College), King's Buildings, West Mains Road, Edinburgh EH9 3JG*
[b] *Biomathematics & Statistics Scotland, King's Buildings, West Mains Road, Edinburgh, EH9 3JZ*

[*] Corresponding author
Email addresses: franz.brulisauer@sac.ac.uk (F. Brülisauer)

1. Power Calculations

A sample size of 300 study farms implied a power of 80% to estimate the true prevalence of herds with PI animals, hence active BVDV infection, with a 95% confidence interval of +/- 6% or less. These calculations were based on the conservative assumption that 60% of the herds included a PI animal. A further assumption was that the probability of a calf having been infected by a PI animal and detected by antibody ELISA given that a PI animal was in the same management group depended on age, with probabilities of 0.95 and 0.97 for calves aged 7 and 12 months respectively.

2. Bayesian Finite Mixture Modelling

Our data comprises the observed number of animals which tested positive for BVDV antibodies (BVDV seropositive) out of a sample of typically ten animals on each of 301 farms. From this we infer the distribution of 'within herd BVDV seroprevalence in young stock' (from now on referred to as seroprevalence) taking into account the sampling variability within and between herds. Our inferencial approach comprises a hierarchical Bayesian mixture model with an unknown number of mixture components (where these denote statistically distinct seroprevalence cohorts in the population, to be inferred from the sample data). In addition we explicitly include error in the classification given by the BVDV antibody ELISA (that is we do not assume the test is a gold standard). We follow the mixture modelling approach set out by Diebolt and Robert (1994). The likelihood for the data assuming binomial sampling in each individual herd is

$$\prod_{i=1}^{N} q_i^{r_i}(1-q_i)^{n_i-r_i}, \tag{1}$$

$$\text{where} \quad q_i = S_e\Delta_i + (1-S_p)(1-\Delta_i) \tag{2}$$

and we adopt the no gold standard parameterisation set out by Joseph et al. (1995) where S_e and S_p denote the sensitivity and specificity of the BVDV antibody ELISA respectively, r_i and n_i denote the observed number of seropositive animals and the total number of animals sampled from herd i respectively, and q_i denotes the probability that a given animal in herd i tests BVDV seropositive. We define $\Delta = \Delta_1, \ldots, \Delta_N$ as independent observations (representing seroprevalence on each farm) from a mixture density with k (assumed known and finite) components. Let $p(\Delta \mid \pi, \mu, \theta)$ denote the probability density for seroprevalence in the population of beef suckler herds:

$$p(\Delta \mid \pi, \mu, \theta) = \pi_1 f(\Delta; \mu_1, \theta) + \cdots + \pi_k f(\Delta; \mu_k, \theta), \tag{3}$$

where $\pi = (\pi_1, \ldots, \pi_k)$ are the mixture proportions and $\mu = (\mu_1, \ldots, \mu_k)$ are component specific parameters (representing mean seroprevalence within each component of the mixture), θ is a vector parameter which is common to all components (representing S_e and S_p) and f is a density (either binomial or beta-binomial). When f is beta-binomial then μ comprises of component specific vector parameters representing mean seroprevalence and in addition a covariance related parameter. Following Stephens (2000) we assume that each observation Δ_i arose from an unknown component z_i of the mixture where $z_1, \ldots, z_N$, are

realisations of independent and identically distributed discrete random variables $Z_1, \ldots, Z_N$ with probability mass function

$$Pr(Z_i = j \mid \pi, \mu, \theta) = \pi_j \quad \text{for} \quad i = 1, \ldots, N \quad \text{and} \quad j = 1, \ldots, k. \tag{4}$$

Hence the Z_i random variables are used to allocate each individual farm to a component of the mixture distribution. Conditional on the Z's, $\Delta_1, \ldots, \Delta_N$ are then independent observations from the densities

$$p(\Delta_i \mid Z_i = j, \pi, \mu, \theta) \quad = \quad f(\Delta_i; \mu_j, \theta) \quad \text{for} \quad i = 1, \ldots, N. \tag{5}$$

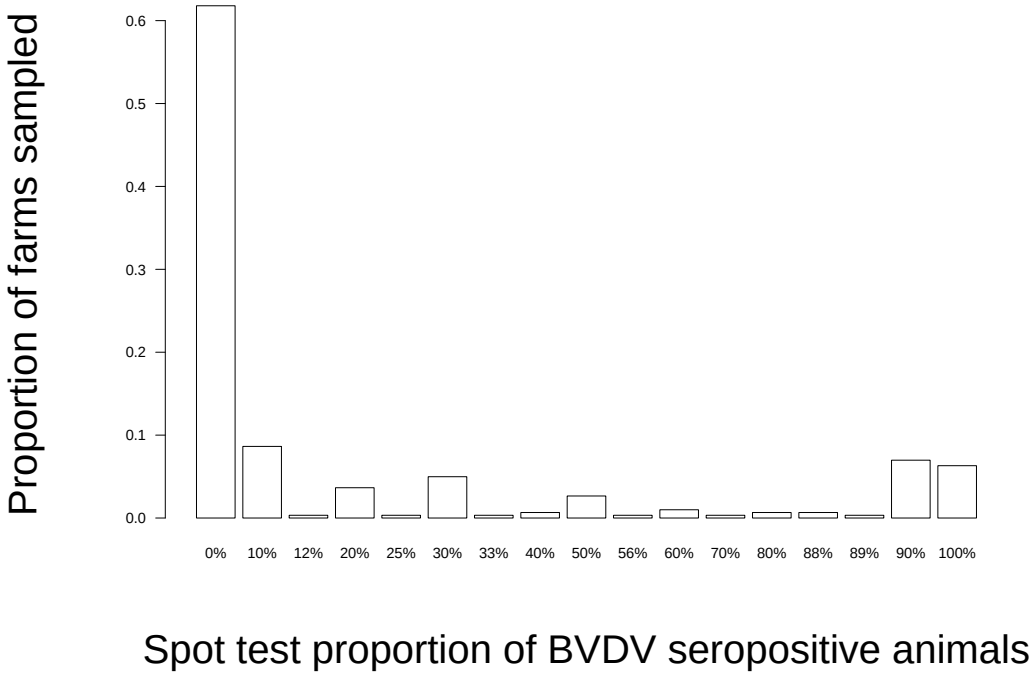
We include $f(\Delta_i; \mu = 0, \theta)$ as a possible component in each of the models to allow for the possibility that each herd could have recently been free from BVDV exposure with non-zero probability.

Our model formulation assumes that k, the number of components in the mixture distribution is known. Our goal is to determine the maximum number of components supported by the observed sample data, and we achieve this by simply enumerating over models with increasing values of k, assessing the goodness of fit in each case using Bayes factors (or more specifically log marginal likelihoods). We then use the model which maximises the goodness of fit. In each case we estimate the posterior distribution $p(Z, \pi, \mu, \theta \mid \Delta)$ using an implementation of the slice sampler due to Neal (Neal, 2003) written in C using the GNU scientific library (Galassi et al., 2006). In the absence of prior knowledge relating to estimates of parameters such as S_e and S_p (in specific relation to Scottish beef suckler herds) we adopt non-informative priors for all parameters:, namely $\beta(1, 1)$ priors for each of S_e and S_p, Dirichlet$(1, \ldots, 1)$ for π and $\beta(1, 1)$ for each μ_j for $j = 1, \ldots, k$ and finally a prior of $P(Z_i = j) = 1/k$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

3. Observed BVDV seroprevalence

In Figure 1 in the main text we showed the empirical distribution of seroprevalence for the 274 farms in the study where exactly 10 young stock were sampled. Figure 1 in the supplementary material shows this distribution but over all 301 herds, where the breakdown of the number of animals sampled on each farm is: 274 farms with 10 sampled; 12 farms with 9 sampled; 6 farms with 8 sampled; 5 farms with 7 sampled; 2 farms with 11 sampled; 1 farm with 18 sampled; 1 farm with 20 sampled.

Figure 1: Observed frequency distribution of seroprevalence for 301 beef suckler herds.



4. Model Selection

Determining the optimal number of components, k, in our mixture model is a key part of our statistical analysis as this is integral to the mechanism used to account for within and between herd variability. We identify the largest value of k supported by the sample data by simply successively fitting models with increasing values of k, and comparing the goodness of fit of each model. We also compare models for $k = 2$ and $k = 3$ in which one of the seroprevalence cohorts is the exposure free state to reflect the reality that an individual herd may be unexposed to the BVDV. That is we allow the true estimated seroprevalence in the herd to take the value zero with some positive probability, rather than assume the seroprevalence in each component has a continuous posterior probability density.

An additional complication which requires consideration in our analysis is the possibility that the data may exhibit greater variance than that which could be accounted for by use of a binomial likelihood. This is typically termed overdispersion and could reasonably be present in our observed data due to the presence of within herd correlation in the immune status of the animals sampled. The term "clustering" is also often used in the veterinary epidemiology literature (e.g. McDermott and Schukken (1994)) to describe such extra between farm/herd variability. Statistical methods to correct for overdispersion are discussed by Lindsey (1999). A standard approach, and the one we adopt, is to compare the goodness of fit of our standard binomial model with the goodness of fit using an extension of this model which explicitly incorporates overdispersion. If the latter model fits the data better than the former, then we use the overdispersed model for inference and vice versa. A number of extended binomial models exist (see Lindsey

(1999)) and we use the most common of these, the beta-binomial model, which is discussed in detail by Prentice (1986) including its various different parameterisations and how these relate to the standard binomial model.

We discuss the results of our overdispersion (beta-binomial) model after presenting full results from our analyses using the standard binomial model. In short there is strong evidence that the standard binomial model is better supported by the data than the overdispersion model. However as parameter estimation and interpretation in the latter model is relatively complex we defer discussion of this until later.

As is standard in Bayesian model comparison we use Bayes factors to compare models, or more specifically since we assume each model has the same prior probability of being the "optimal model", we compare the log marginal likelihood of each model. To ensure robust estimation of the log marginal likelihood we follow Congdon (2001) and divide the output into batches, calculate the harmonic mean in each batch and then mean these values. Up to eight batches were generated along with the use of the median rather than mean to average over batches. Such variations had negligible impact on the resulting marginal likelihood values giving confidence in the robustness of our estimates.

The goodness of fit for models with $k = 1, 2, 3$ is shown in Table 1. For the three values of k so far examined $k = 3$ has a substantially superior goodness of fit. We also find that there is very little difference in goodness of fit between models with a exposure free cohort and otherwise (however we later provide justification of why the inclusion of the exposure free cohort provides a more robust model when $k = 3$). See Congdon (2001) Table 10.1 for approximate

Table 1: Goodness of fit comparisons between mixture models with 1, 2 and 3 components

Mixture Model	log(marginal likelihood)
1 component	-1496.927
2 components	-700.116
2 components (inc. exposurefree)	-700.059
3 components	-552.551
3 components (inc. exposure free)	-552.138

guidelines on the magnitude of differences required between Bayes factors to be notable, ranging from weak support (denoting the smallest difference between log marginal likelihoods) through to very strong support (denoting a difference in log marginal likelihoods of at least 5). As with all Bayesian analyses it is important to verify that the output from the Markov chain Monte Carlo sampler is sensible (i.e. that the sampler appears to have reached the stationary distribution and that the subsequent mixing is suitably stable). Next, we briefly illustrate diagnostics for our models with $k = 1, 2, 3$; an exercise which is of particular relevance when we later consider models with $k = 4$, as diagnostics for the latter are highly unstable and suggest overfitting.

4.1. Model diagnostics

Figures 2 and 3, and 4 and 5, respectively show typical Markov chain Monte Carlo output for models with two and three components, both with and without the inclusion of a exposure free component. The initial burn-in period in each run was both obvious and very short (several thousands iterations) and the output

shown is from 10,000 to 1,000,000 iterations. The mixing appears generally quite satisfactory, and the same was true for the other model parameters which are not show. The possible exception is the mixing of the seroprevalence estimates for cohort 3 (the component with the highest seroprevalence) in the model with $k = 3$ and no exposure free cohort, the output in green in Figure 5. Running the chains for this model very much longer produces identical output which suggest that the problem is not due, for example, to excessively long burn-in and a failure to reach stationarity. The latter mixing is somewhat slow and unstable, and in addition the seroprevalence estimates for cohort 1 (the component with the lowest seroprevalence) are very low (mean and medians of 0.007 and 0.006 respectively), and much less that the minimum possible observed seroprevalence (of 0.05, 1 BVDV seropositive animal out of a spot test of 20). This suggests that the model with $k = 3$ including a exposure free component is the more robust of the two $k = 3$ models, as the model without the explicit exposure free cohort appears to be trying to approximate a exposure free cohort but is restricted to a continuous posterior prevalence distribution (with an obvious boundary at zero). This may in some part explain the rather unsatisfactory mixing behaviour observed for cohort 3. Hence the model with $k = 3$ with a exposure free cohort is our preferred model of those so far presented.

5. Models with 4 or more components

Figure 6 shows similar output to Figures 2 through 5 but now for the model with $k = 4$ components (including a exposure free component). Clearly there are serious issues with the mixing of this model. In direct comparison to the

Figure 2: Markov chain Monte Carlo output for mixture model with two components, one of which is the exposure free cohort. Shown is the marginal trace of estimates for seroprevalence in the non-exposure free cohort.

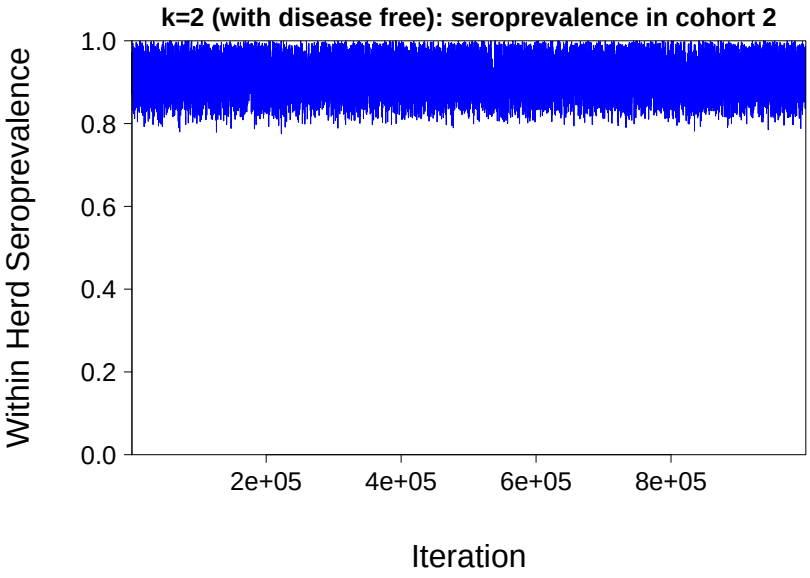


Figure 3: Markov chain Monte Carlo output for mixture model with two components. Shown are the marginal traces of estimates for seroprevalence in each of the two cohorts in the model.

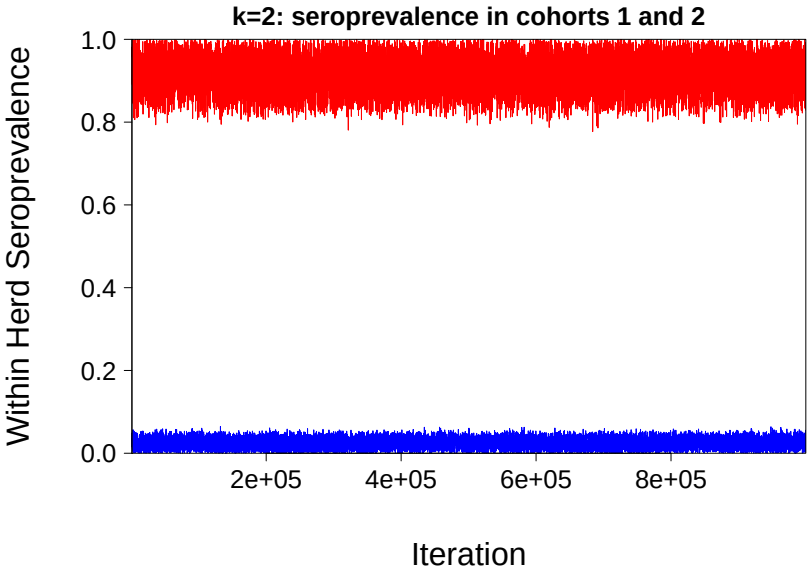


Figure 4: Markov chain Monte Carlo output for mixture model with three components, one of which is the exposure free cohort. Shown are the marginal trace of estimates for seroprevalence in the two non-exposure free cohorts.

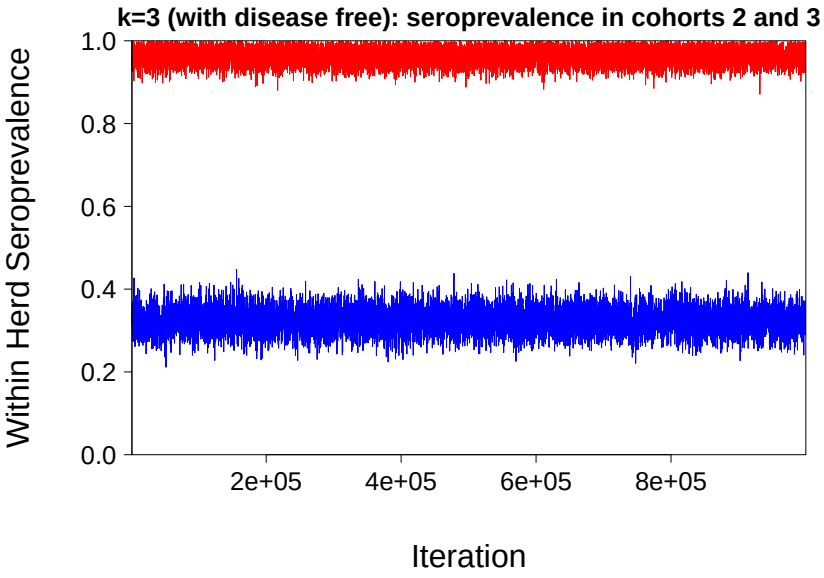
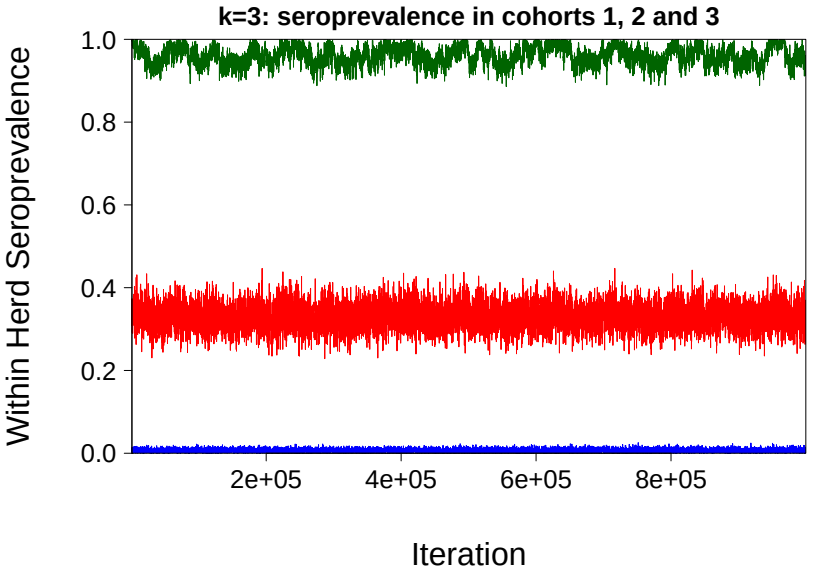


Figure 5: Markov chain Monte Carlo output for mixture model with three components. Shown are the marginal trace of estimates for seroprevalence in the three cohorts.



models with fewer than four components, where the seroprevalence estimates for each component were clearly distinct from each other, in the $k = 4$ model the various prevalence estimates are merging to cover the entire range of possible prevalence values from zero to unity. This behaviour suggests overfitting; for example the trace for seroprevalence in cohort 2 (in blue) appears to regularly visit the neighbourhood of zero but this model already explicitly has a exposure free component.

In an attempt to improve the mixing we introduce constraints which bound away the various seroprevalence estimates from each other (in particular away from zero and aliasing with the exposure free cohort). Such constraints are commonly used to avoid label switching problems in Bayesian mixture models (see Stephens (2000)). We use a range of bounds, b, from 0.001 up to 0.01 where for each value of b the constraint is implemented by adjusting the current set of seroprevalence estimates generated by the sampler to ensure that they differ by at least b from each other (if this is not already the case). The impact of these constraints on model estimation is shown in Figures 7 and 8.

Figure 7 is similar to Figure 6 but now the prevalence estimates in each component at each iteration are adjusted (if necessary) to remain at least 0.05 apart, as can be most clearly seen by the movement away from prevalence estimates close to zero. The exposure free cohort is also present this model. It is clear that the same behaviour exists as in Figure 6 in that the prevalence estimates for cohort 2 (in blue) are still sampled regularly very close to the constrained lower bound of 0.05. Figure 8 is similar to Figure 7 but uses the greater bound of $b = 0.1$. The output from the sampler is again complex with potentially even

Figure 6: Markov chain Monte Carlo output for mixture model with four components, one of which is the exposure free cohort. Shown are the marginal trace of estimates for seroprevalence in the three non-exposure free cohorts.

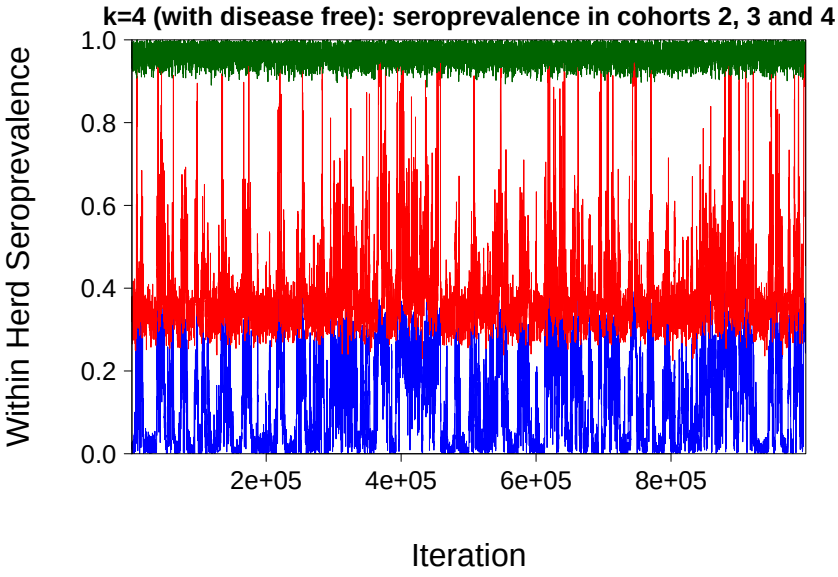
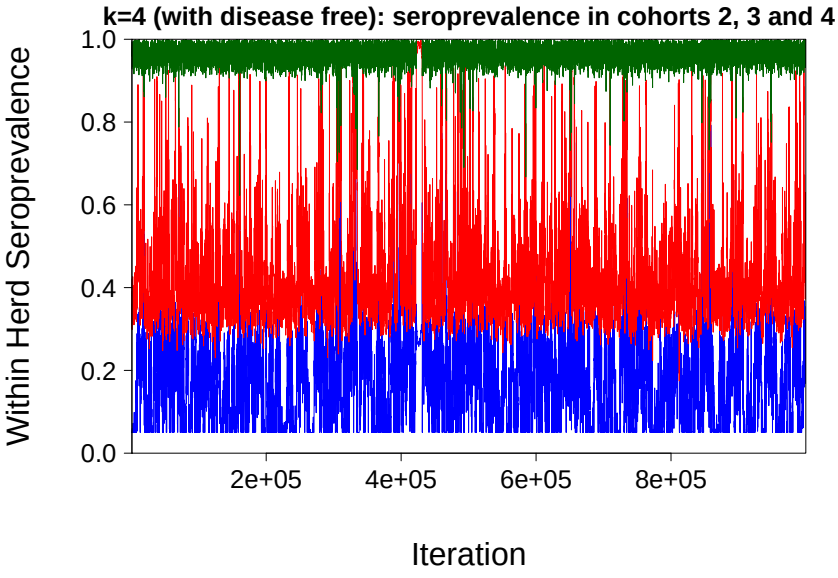
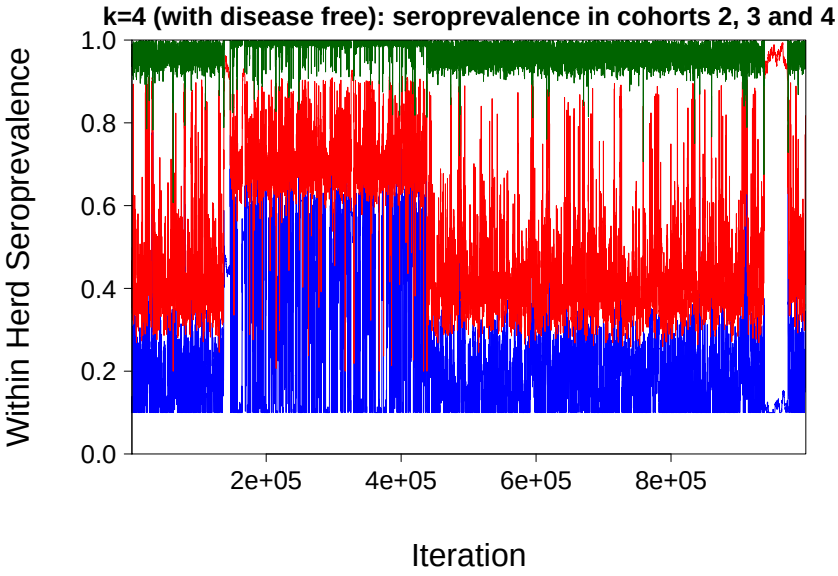


Figure 7: Markov chain Monte Carlo output for mixture model with four components, one of which is the exposure free cohort, and using a between prevalence bound of $b = 0.05$. Shown are the marginal trace of estimates for seroprevalence in the three non-exposure free cohorts.



a new stationary "solution" being visited for some duration of the run. This may be explained by over-complexity in our model, in that fitting a model with four mixture components to data which does not support four distinct cohorts, will cause the process to drift through the parameter space unable to stabilise on a single (multivariate) stationary distribution. This appears to be exactly what is occurring with our four component model. Hence in summary our chosen model given the observed data is that with $k = 3$ components including an explicit exposure free cohort.

Figure 8: Markov chain Monte Carlo output for mixture model with four components, one of which is the exposure free cohort, and using a between prevalence bound of $b = 0.1$. Shown are the marginal trace of estimates for seroprevalence in the three non-exposure free cohorts.



6. Modelling Overdispersion

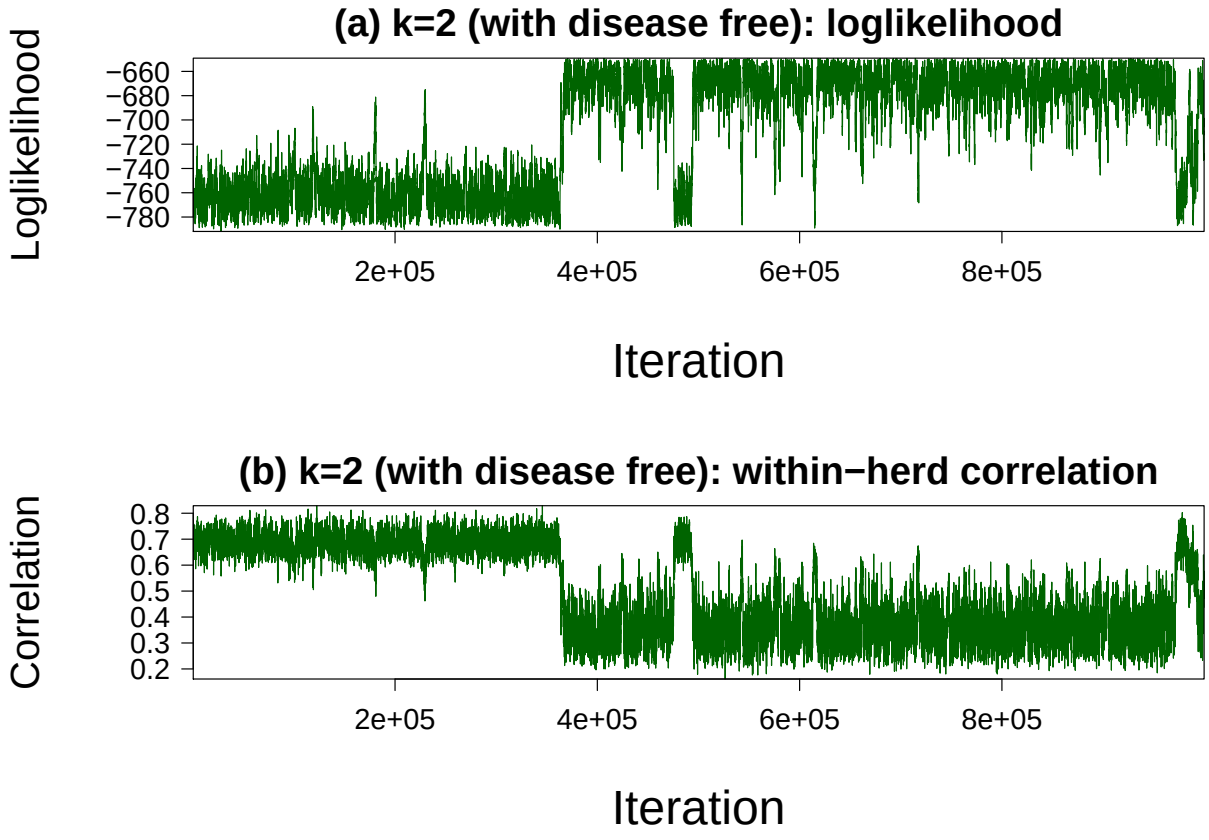
We use the beta-binomial extension of the binomial distribution to examine whether the data are better supported by a model with overdispersion. A possible parameterisation of this model is shown in eqn (6),

$$\Pr(R = r; n) = \frac{\binom{n}{r} \prod_{k=0}^{r-1}(q + \gamma k) \prod_{k=0}^{n-r-1}(1 - q + \gamma k)}{\prod_{k=0}^{n-1}(1 + \gamma k),} \quad (6)$$

where q is the usual binary response probability within an experimental unit, $\gamma(1+\gamma)^{-1}$ the binary variate correlation parameter (see Prentice (1986) for more details) and we adopt the usual convention that $\prod_{k=0}^{x} c_k = 1$ for any $x < 0$. Our finite mixture model is analogous to the binomial approach detailed earlier except now our model contains mixtures of beta-binomial distributions, where again we have $q_i = S_e\Delta_i + (1 - S_p)(1 - \Delta_i)$.

Our beta-binomial mixture model is fitted to the observed data using exactly the same slice sampler approach as for the binomial model except that we now have additionally to estimate the within herd correlation parameter, or rather the parameter γ which is a function of the correlation parameter but results in much more stable sampling. As with all other model parameters we use an uninformative prior for γ, specifically uniform on the range 0, 100, and from MCMC output it was clear that this choice of prior did not impose any practical constraints on posterior realisations of γ. We fitted the range of models defined in Table 1 to the data using mixtures of beta-binomial densities. The parameter estimation for these models was more complex as the MCMC mixing appears to suggest the presence of possibly bi-modal posterior distributions where each mode corresponds to very different values of the correlation parameter. Figure 9 shows the

Figure 9: Markov chain Monte Carlo output for beta-binomial mixture model with two components (one exposure free). (a) estimates of the log likelihood at each iteration and (b) estimates of within herd correlation parameter.



log likelihood for the beta-binomial model with two components (with one set to be the exposure free component) and corresponding values for the correlation parameter. The burn-in period for this chain was short and clearly identifiable and has been discarded. For the 1×10^6 iterations shown it is clear that the log likelihood jumps between two modes, one where the log likelihood averages around -760 and the correlation parameter around 0.7, and a second mode with a higher mean log likelihood of approximately -670 and a much lower correlation parameter with a mean close to 0.3. Running much longer chains for this model results in a similar pattern of unpredictable jumps between these two seemingly stable posterior modes. We estimate the log marginal likelihood for the mode with the higher log likelihood as approximately -677.3 and for the other mode as approximately -758.2. The beta-binomial model with three components (with one exposure free component) has similar multi-modal behaviour to that of the two component model. However the behaviour of the model with three components suggests over-fitting with complex mixing (see Figure 10). In particular the correlation parameters (one for each of the two non-exposure free components) appear to "trade off" against each other as can clearly be seen in Figure 11. The mode with higher log likelihood corresponds to one correlation parameter being very close to unity while the other is close to 0.3 and vice versa. However it is clear that the typical log likelihood is sufficiently poor compared to that for the three component standard binomial model (Table 1), that the relatively complex nature of the beta-binomial model does not improve the fit relative to this, our optimal model.

Figure 10: Markov chain Monte Carlo output for beta-binomial mixture model with three components, one of which is the exposure free cohort. Shown are the marginal trace of estimates for within herd BVDV seroprevalence in the two non-exposure free cohorts.

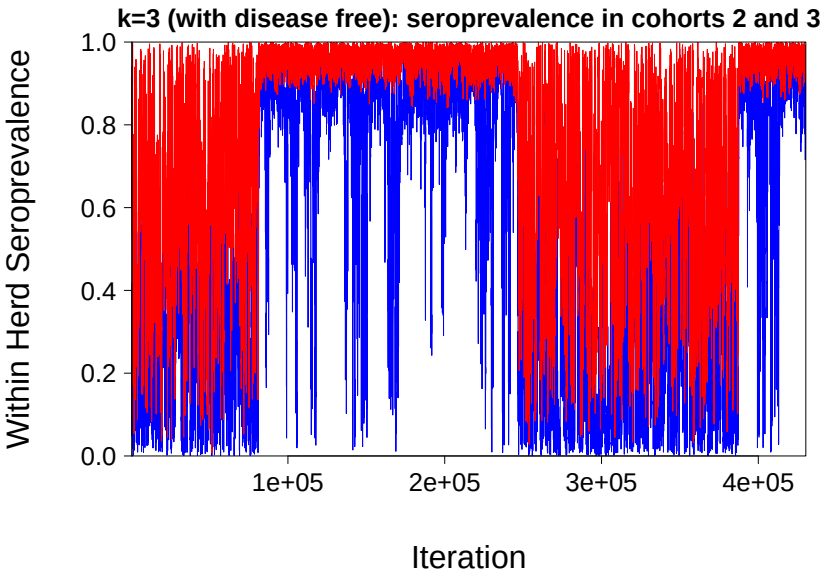
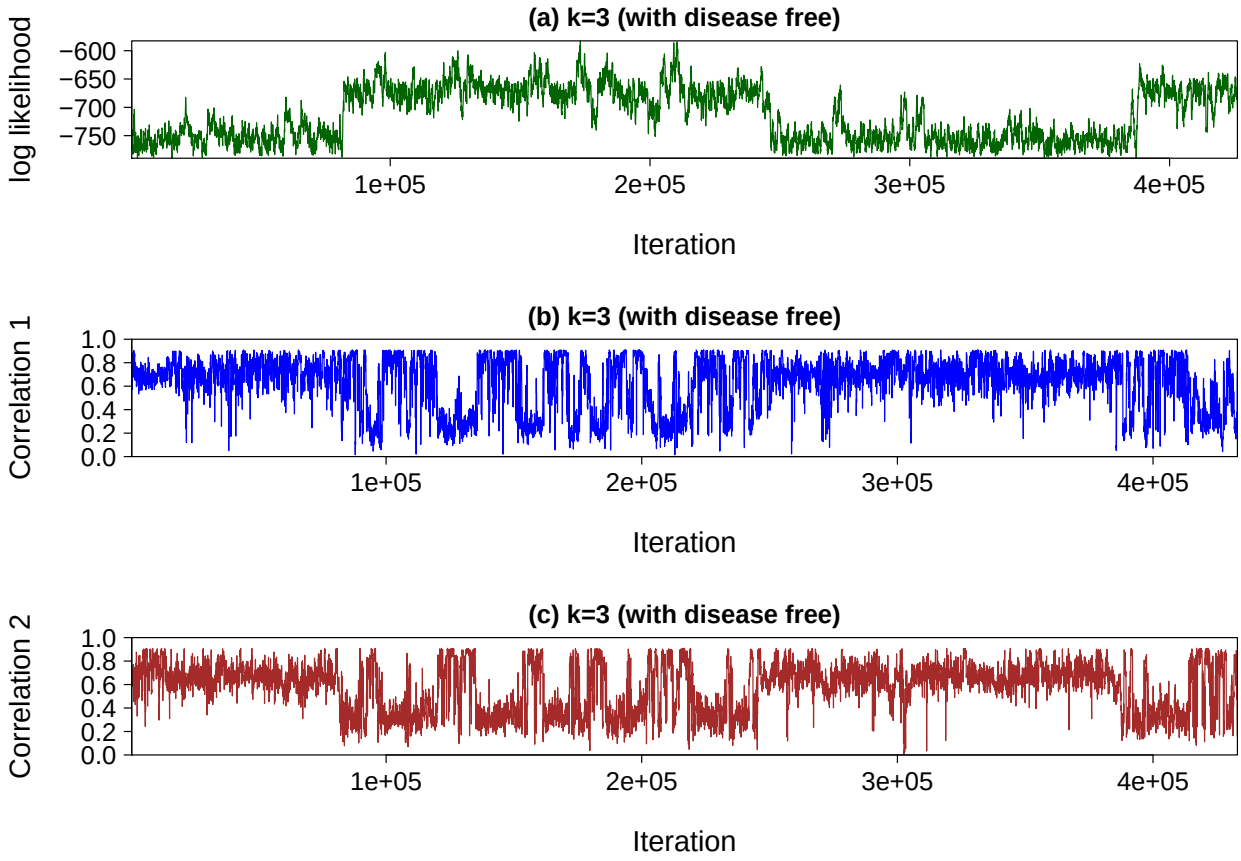


Figure 11: Markov chain Monte Carlo output for beta-binomial mixture model with three components (one exposure free). (a) estimates of the log likelihood at each iteration; (b) estimates of first correlation parameter and (c) estimates of second correlation parameter.



References

Congdon, P., 2001. Bayesian Statistical Modelling. Wiley.

Diebolt, J., Robert, C. P., 1994. Estimation of finite mixture distributions through bayesian sampling. Journal Of The Royal Statistical Society Series B-Methodological 56 (2), 363–375.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M., F., R., 2006. GNU Scientific Library Reference Manual - Revised Second Edition (v1.8). Network Theory Ltd.

Joseph, L., Gyorkos, T. W., Coupal, L., Feb. 1995. Bayesian-estimation of disease prevalence and the parameters of diagnostic-tests in the absence of a gold standard. American Journal Of Epidemiology 141 (3), 263–272.

Lindsey, J. K., 1999. On the use of corrections for overdispersion. Journal of the Royal Statistical Society Series C-Applied Statistics 48, 553–561.

McDermott, J. J., Schukken, Y. H., Feb. 1994. A review of methods used to adjust for cluster effects in explanatory epidemiologic studies of animal populations. Preventive Veterinary Medicine 18 (3), 155–173.

Neal, R. M., Jun. 2003. Slice sampling. Annals of Statistics 31 (3), 705–741.

Prentice, R. L., 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. Journal of the American Statistical Association 81, 321–327.

Stephens, M., 2000. Dealing with label switching in mixture models. Journal Of The Royal Statistical Society Series B-Statistical Methodology 62, 795–809.