Comparing the accuracy of maximum likelihood and MCMC estimates using a Zero-Inflated Gamma Poisson model

Denwood, M.J.[1], Toft, N.[2], Love, S.[1], Stear, M.J.[1], Reid, S.W.J.[1] and Innocent, G.T.[1]

This research was produced as part of the DEFRA-funded VTRI project 0101

[1] Comparative Epidemiology and Informatics, Institute for Comparitive Medicine, Department of Animal Production and Public Health, University of Glasgow Veterinary School, Bearsden Road, Glasgow G61 1QH, UK.
[2] Associate Professor, Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen, Grønnegårdsvej 8 1870 Frederiksberg C

The zero-inflated gamma poisson (negative binomial) model has applications in many fields including veterinary parasitology (*Nødtvedt et al, 2002*). The more commonly used maximum likelihood (ML) and comparatively modern Markov chain Monte Carlo (MCMC) techniques can be used to apply such a model. Therefore, comparison of these two methods when applied to simulated data is of interest, to determine which produces results that most accurately reflect the simulation parameters.

A total of 1,000 datasets each were produced for groups of 10 counts, 100 counts and 1,000 counts (a total of 3,000 datasets). Data were generated using R with values for mean count and overdispersion from a log normal distribution, and zero inflation from a normal distribution. Datasets containing all '0' counts were discarded before analysing the remaining 2956 datasets using WinBUGS, for MCMC, and R, for ML. The MCMC model successfully analysed and returned useable information on all parameters for 86% of the datasets, and the ML model returned useable results from 97%, 92% and 83% of the datasets for mean count, zero inflation and overdispersion respectively. As would be expected, more useable information was consistently returned by both models when larger sample sizes were used.

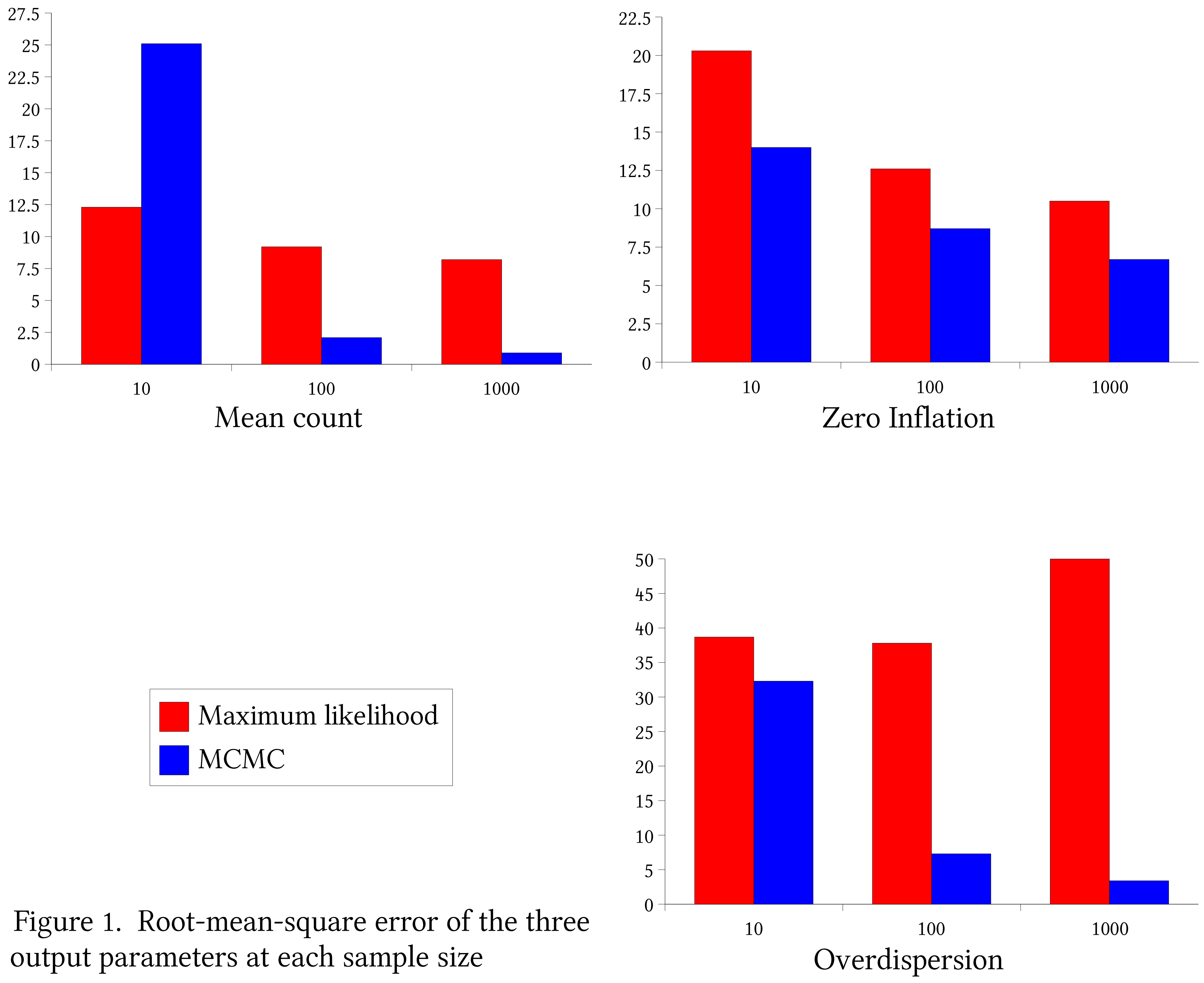

Figure 1. Root-mean-square error of the three output parameters at each sample size

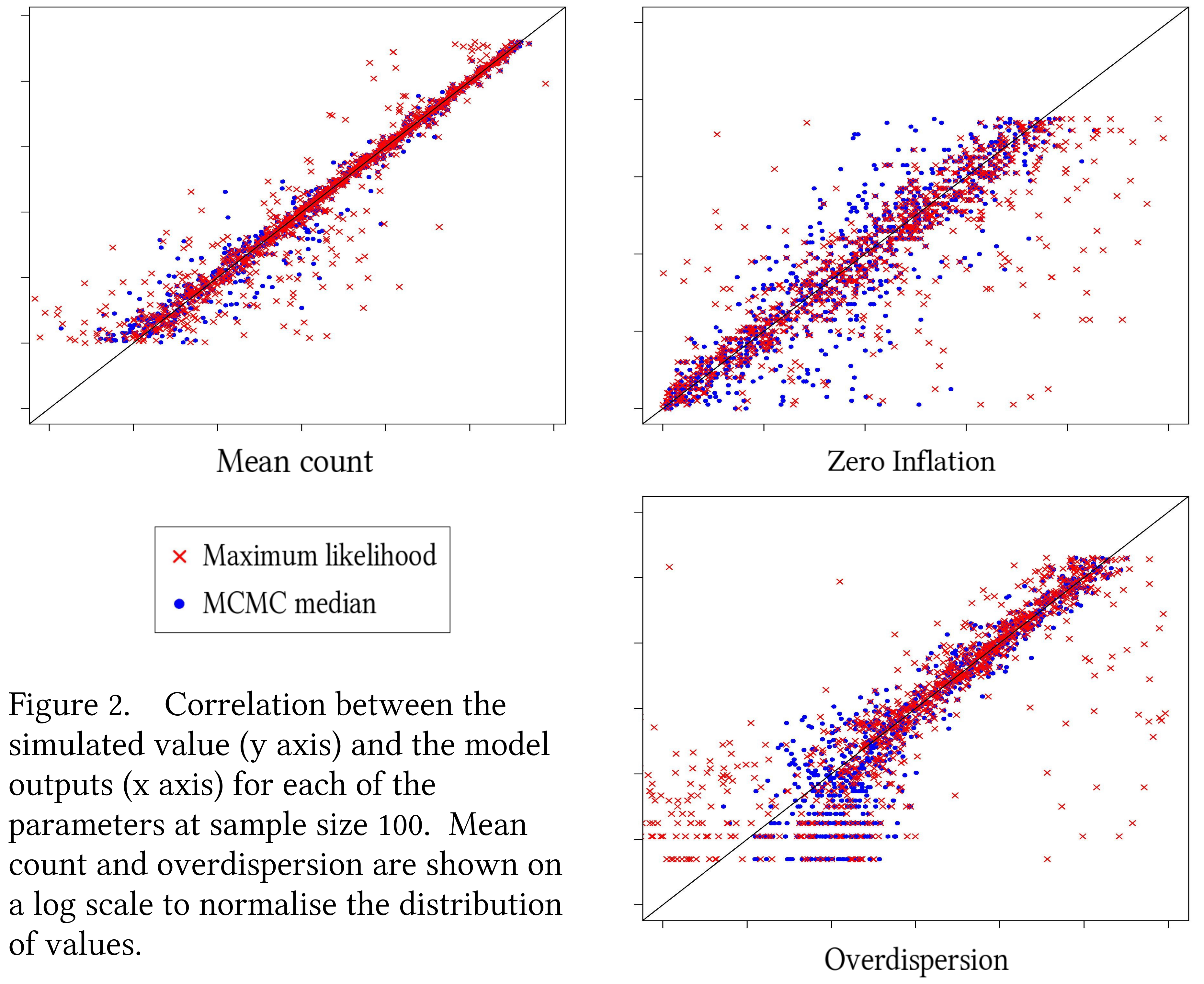

Figure 2. Correlation between the simulated value (y axis) and the model outputs (x axis) for each of the parameters at sample size 100. Mean count and overdispersion are shown on a log scale to normalise the distribution of values.

The root-mean-square-error and proportion of values correctly identified in the 95% confidence interval were both generally superior using MCMC (Figure 1, Table 1). The agreement between the MCMC model output and the simulated value for each parameter was also better than that for the ML model (Figure 2). The increase in the root-mean-square-error for the ML overdispersion at larger sample sizes was most likely due to fewer datasets being rejected because of a lack of information about this parameter, resulting in the most striking difference in accuracy between the two models. The tendency for MCMC to over-estimate overdispersion at very low values is due to the model parameterisation, which has since been re-written to overcome this issue. These results suggest that MCMC should be used increasingly in conjunction with, or even in preference to, ML methods for modelling for this type of data.

	Sample size 10		Sample size 100		Sample size 1000	
	ML	MCMC	ML	MCMC	ML	MCMC
Meancount	0.15	0.05	0.12	0.04	0.15	0.06
Zero Inflation	0.09	0.04	0.12	0.05	0.09	0.07
Overdispersion	0.26	0.17	0.18	0.24	0.17	0.30

Table 1. Proportion of 95% confidence intervals which did not contain the simulation population value for each parameter. Lower numbers represent more correct models.

References
Nødtvedt, A., Dohoo, I., Sanchez, J., Conboy, G., DesCôteaux, L., Keefe, G., Leslie, K. and Campbell, J. (2002) The use of negative binomial modelling in alongitudinal study of gastrointestinal parasite burdens in Canadian dairy cows. The Canadian Journal of Veterinary Research, 66, p249-257.