DOI: 10.1111/jbg.12426

EDITORIAL

Journal of Animal Breeding and Genetics WILEY

On the relevance of Bayesian statistics and MCMC for animal models

Quantitative genetics has a historical record of relying on Bayesian statistics, especially so in the field of animal breeding since, for example, the seminal work of Sorensen and Gianola (2002, *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer, New York). One aspect of the appeal of Bayesian statistics is that assigning probabilities to values of parameters and hypothesis allows for a description of the inference process which is very intuitive in the context of scientific research. But beyond the philosophical distinction between Bayesian and frequentist interpretation of statistical inference, Bayesian statistics provide some practical advantages, as well as different constraints (including the need to define prior distributions of parameters and the difficulty to completely solve Bayes' theorem). A first advantage is that having inferences in the form of posterior distributions greatly improves our ability to interpret uncertainty around estimates and, more importantly, to propagate such uncertainty in a series of subsequent analyses. A second, unfortunately often overlooked, the advantage is our ability to incorporate previous, state-of-the-art knowledge into prior distributions to improve our ability to efficiently fit models.

A third, technical advantage comes from the availability of general, flexible and robust algorithms: Markov Chain Monte Carlo (MCMC). These algorithms are often needed in Bayesian statistics, because computing the posterior distribution of the parameters technically requires solving a complex, multivariate integral corresponding to the denominator Bayes' theorem. By sampling directly from the posterior distribution using ratios of posterior probabilities that are easier to compute, MCMC algorithms circumvent this issue. And because they are a very general family of algorithms, with little to no approximations, they can be applied with a high degree of confidence to a very large variety of problems.

Although quantitative genetics has made use of very efficient frequentist algorithms to optimize restricted maximum likelihood (REML) of complex linear mixed models (LMMs), like the animal model (such as the average information algorithm implemented in ASReml, Gilmour et al. 1995 *Biometrics* 51: 1440–1450), such algorithms could also suffer from difficulties to converge and cannot always be used to fit generalized linear mixed models (GLMMs). Yet, not all traits of interest for quantitative geneticists are nicely and normally distributed: some traits, so-called non-Gaussian, are discrete and/or follow highly skewed distribution. Such traits can still be of strong interest and include all kinds of discrete characters, like diseases, survival or fecundity and are often best analysed using binomial or Poisson distribution, sometimes even with zero-inflation, over-/under-dispersion or some kind of truncation. However, historical algorithms for fitting GLMMs, such as the penalized quasi-likelihood (PQL, Breslow and Clayton 1993 *J. Am. Stat. Assoc.* 88: 9–25) have difficulties in correctly estimating variance components (Breslow and Lin 1995 *Biometrika* 82: 81–91). Fortunately, MCMC algorithms being general, they do not suffer from such limitations. This, as well as the other benefits of Bayesian statistics mentioned above (O'Hara et al. 2008 *J. Evol. Biol.* 21: 949–957), led to MCMC implementation of 'generalised animal models', most famously in the R package MCMCglmm (Hadfield 2010). MCMC has been shown to outperform other algorithms for the estimation of the additive genetic variance of, for example, binary traits (de Villemereuil et al. 2013 *Methods Ecol. Evol.* 4: 260–275).

In a decade, the availability of a free and open source package implementing generalized animal models has led, at least in the community of evolutionary quantitative genetics, to an increase in the number of studies including non-Gaussian traits, even in complex settings. The fact that natural selection acts on the actual phenotype, and not breeding values on a virtual 'latent scale', triggered methodological developments to be able to work the actual phenotypic scale when using such generalized animal models (de Villemereuil et al. 2016 *Genetics* 204: 1281–1294). Although the biological interpretation of such models might not be always satisfying (de Villemereuil 2018), these models have helped quantitative geneticists to analyse traits that were difficult to tackle without them. Although skewed traits can often be analysed after a normalizing transformation, MCMC algorithms can be devised to analyse them directly with a skewed error distribution (Varona et al. 2008 *Genet. Res.* 90: 179–190.). Furthermore, quantitative traits can be skewed not only because of the environmental effects, but because of the skewness of the breeding values themselves, for example, when facing local adaptation and migration (Débarre et al. 2015 *Am. Nat.* 186: S37–S47), and such skewness can greatly impact their response to selection (Bonamour et al. 2017 *Evolution* 71: 2703–2713). Although I am not aware

of a statistical implementation of animal models accounting for such possible skewness in the breeding values, this exercise should be relatively straightforward using MCMC algorithms. Having a posterior distribution of various parameters such as the breeding values or variances has also been a great help in many situations. A first obvious benefit is the ability to compute standard errors and credible intervals for derivated estimates such as heritability, even in the complex setting of generalized animal models, and even after back-transformation of such parameters (de Villemereuil 2018 *Ann. N. Y. Acad. Sci.* 1422: 29-47). A second benefit is the ability to work with and analyse breeding values while accounting for the (often strong!) uncertainty around them (Hadfield et al. 2010 *J. Stat. Softw.* 33: 1–22).

Yet, despite all the advantages they provide, MCMC algorithms are not the norm for animal models. This might be partly explained by the general reluctance of the scientific community to switch to Bayesian statistics, either because of a philosophical rebuttal of it or more genuinely because of historical inertia. More likely though, it might be explained by the fact that MCMC algorithms are slow algorithms, especially for complex models leading to auto-correlation of MCMC sample chains, and that they require a bit of user expertise to check for convergence and large enough effective sample size. As an example, despite being a very efficient MCMC implementation of the animal model, MCMCglmm can become extremely slow when, instead of a pedigree, it is provided with a genomic relatedness matrix (GRM, which are not sparse matrices). If MCMC algorithms are slow when using such GRM matrices, and more generally so in a world of 'big data', then are they still useful for quantitative genetics?

I would believe that we still need Bayesian statistics and that good algorithms to fit generalized animal models are also still needed. As such, I would expect MCMC to stick around for some time still. A possible advance might come from the next generation of Bayesian simulation algorithms, namely Hamiltonian Monte Carlo (HMC, Hoffman and Gelman 2014 *J. Mach. Learn. Res.* 15: 1593–1623) implemented in the STAN framework. Hamiltonian Monte Carlo differs from plain MCMC in that it relies on a deterministic step in the sampling process, which can strongly reduce the level of auto-correlation between successive iterations. Running a GRM-based animal model with STAN (e.g. using the brms R package, which offers friendly functions for such models, Bürkner 2017 arXiv:1705.11123 [stat.CO]) could be achieved in a reasonable amount of time for example, depending on the size of the dataset. Another possibility is to rely on more approximative but fast methods to obtain Bayesian posterior distribution, such as Integrated Nested Laplace Approximations (INLA, Rue et al. 2009 *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71: 319–392). Specialized packages exists for using INLA with pedigrees, such as the animalINLA R package (Holand et al. 2013 *G3 Genes Genomes Genet.* 3: 1241–1251), which, at least for Gaussian traits, has been shown to perform quite well compared with MCMC implementations such as MCMCglmm (Mathew et al. 2015 *Mol Breeding* 35: 99). Unfortunately, those approximative algorithms, like their frequentist counterparts can lack generality. As an example, for binary traits, the performance of INLA has been shown to be subpar, compared with MCMC (Holand et al., *ibid.*).

Because they are more general than approximation-based algorithms, I thus expect that simulation-based algorithms such as MCMC and HMC will continue to play a strong role in helping quantitative geneticists solve difficult problems, like the ones posed by generalized animal models, or by the skewness of traits (or even of breeding values themselves). By definition, however, such simulation-based algorithms will always be slow, as they require complex schemes to sample directly from the posterior distribution, which I believe is the main point preventing them becoming more popular. For more classical animal models, the availability of fast Bayesian algorithms like INLA could do a lot to popularize the use of Bayesian statistics in quantitative genetics.

ACKNOWLEDGEMENT

I thank Céline Téplitsky and Justine Bisson for their useful comments on the manuscript.

Pierre de Villemereuil

CEFE, CNRS, EPHE, IRD, Université de Montpellier, Université Paul Valéry Montpellier 3, Montpellier, France

Email: bonamy@horus.ens.fr