Journal of
Advanced Veterinary Research
Volume 1 (2011) 94-98 **Review Article**

Bayesian Statistics: Concepts and Applications in Animal Breeding – A Review

Lsxmikant- Sambhaji Kokate, G.R. Gowane, Dige M.S.*., Sonawane G.S., C Mishra, R.K. Singh[3]

Indian Veterinary Research Institute, Izatnagar, Bareilly (Uttar Pradesh) - 243122, India.

(Recieved 18 May 2011/ Accepted 22 July 2011)

Abstract

Statistics uses two major approaches- conventional (or frequentist) and Bayesian approach. Bayesian approach provides a complete paradigm for both statistical inference and decision making under uncertainty. Bayesian methods solve many of the difficulties faced by conventional statistical methods, and extend the applicability of statistical methods. It exploits the use of probabilistic models to formulate scientific problems. To use Bayesian statistics, there is computational difficulty and secondly, Bayesian methods require specifying prior probability distributions. Markov Chain Monte-Carlo (MCMC) methods were applied to overcome the computational difficulty, and interest in Bayesian methods was renewed. In Bayesian statistics, Bayesian structural equation model (SEM) is used. It provides a powerful and flexible approach for studying quantitative traits for wide spectrum problems and thus it has no operational difficulties, with the exception of some complex cases. In this method, the problems are solved at ease, and the statisticians feel it comfortable with the particular way of expressing the results and employing the software available to analyze a large variety of problems.

Keywords: Statistics; Bayesian; Animal Breeding; Bayes theorem

Introduction

TThe term Bayesian in the Bayesian statistics was given in honor of Thomas Bayes (1702–1761), who proved a special case of what is now called as Bayes therom. After the discovery of Bayes theorm, Pierre-Simon Laplace (1749–1827) introduced a general version of the theorem and used it to approach problems in celestial mechanics, medical statistics, reliability and jurisprudence. Laplace also introduced primitive version of conjugate priors and the theorem of Von Mises and Bernstein, according to which the posteriors corresponding to initially differing priors ultimately agree, as the number of observations increases. This was a great leap in the statistical approach in the solution of the problem especially when the problem of uncertainty is involved. Usually, mathematical statisticians use two major paradigms- one is conventional (frequentist) and other is Bayesian approach for data analysis. Bayesian approach provides a complete paradigm for both statistical inference and decision making under uncertainty. Bayesian methods may be derived from a self-evident system, and hence provide a general, coherent methodology. It contains, at particular cases, many of the more often used frequentist procedures, solves many of the difficulties faced by the conventional statistical methods and extends the applicability of statistical methods. There are various fields worldwide where Bayesian statistics is used successfully for better prediction of the effects with more precision such as prediction of monsoon, prediction of chances of winning in the unlikely events in the sports and so on.

The most important limitation for more extensive implementation of Bayesian approach in day to day statistics is that obtaining the posterior distribution often requires the integration of high-dimensional functions. Bayesian calculations almost require integration over uncertain parameters. This integration often has no analytical solution and instead requires computationally intensive numerical integration such as Markov chain Monte Carlo method. Until the advent of computer, Bayesian approach was often not flexible. Secondly, Bayesian methods require specifying prior probability distributions, which are often unknown. Bayesian statis-

*Corresponding author: Mahesh Shivanand Dige
Address:.; INDIAN VETERINARY RESEARCH INSTITUTE

*E-mail address:*maheshdige@gmail.com

tics generally assumes so called 'uninformative' priors in such cases. Though Bayes therom is trivially true for random variables X and Y, it is not clear that parameters or hypothesis should be treated as random variables.

Mathematical computation difficulties such as integration of high dimensional functions and over uncertain parameters were solved by several approaches which could bypass this tedious process reported in the literature by Smith (1991), Evans and Swartz (1995) and Tanner (1996). One of the most important approaches is Markov Chain Monte Carlo (MCMC) methods, which attempt to simulate direct draws from some complex distribution of interest. Here one uses the previous sample values to randomly generate the next sample value, generating a Markov chain (as the transition probabilities between sample values are only a function of the most recent sample value). Gelfand and Smith (1990) realized that one particular MCMC method, the Gibbs sampler, is extremely extensively applicable to a broad class of Bayesian problems. This thinking was new and revolutionized the field of Bayesian statistics completely thus initiating a major increase in the application of Bayesian analysis and the efforts are being still continued for its widespread application.

It was way back around 1950's when the MCMC methods' seeds were sown through the origin of the Metropolis algorithm (Metropolis and Ulam, 1949 and Metropolis et al., 1953). In the year of 1984, the field of image processing gave rise to the most important method of MCMC i.e. the Gibbs sampler (Geman and Geman, 1984). It is therefore important to see that this field is not as new as it seems to be. It is also ironic to see that in spite of many efforts by research workers, this method neither was not quickly and widely accepted, nor is it used today as extensively as it should be looking towards the potential of the method. The transformation in the field of Bayesian statistics was use of computer packages and programs for using the Bayesian approaches. Today many softwares such as SAS, WINBUGS, OpenBUGS are used for the general purpose-Bayesian analysis of the data (SAS 2006, Thomas et al., 2006, Lunn et al., 2000). Application of Bayesian approach in statistical design of the solutions for problems in breeding was a novel idea. Harville (1977) offered a Bayesian interpretation of REML successfully for the first time.

Application in Animal Breeding

Taking lessons from the history and the background of the Bayesian statistics and the initiation of this method in the application for analysis of animal breeding problems we turn towards its practical application in the field. One of the most important problems in the breeding data analysis is estimation of variance components. In the past, ANOVA was used for this purpose. Henderson (1953) developed analogue techniques for unbalanced data. Because of the use of vector notation, those techniques became popular for use in computer programmes (Harvey, 1977 and SAS, loc. cit). In essence, techniques are the same as in balanced data, using an ANOVA table with the sum of squares for the different effects and their expectations. Maximum likelihood approach was used and still being used for computation of the genetic parameters which was surpassed by the Residual Maximum Likelihood (REML) and upgraded to Derivative Free (DF) REML. However, these procedures are basically meant for the data which is linear and normally distributed. Many a times we come across several parameters which are not normally distributed and are basically categorical or binary in nature. Many a times they follow a binomial distribution. Bayesian approach is better way for the data which is categorical, not normally distributed and also for the normally distributed data. Usually MCMC Gibbs sampling procedure is followed for the data analysis in animal breeding while applying Bayes statistics. In the MCMC Gibbs sampling, we would obtain a point estimate of genetic variance and a single measure of uncertainty, which, technically speaking is only meaningful in large samples and if the data are normally distributed. Estimation of responses can be derived from an animal model, but their properties are unknown. An alternative is to adopt the Bayesian approach. The Bayesian approach resides in arriving at the marginal posterior distribution of the unknown of interest. This distribution provides an exact account of the uncertainty about the unknown parameter. Although Bayesian methods were theoretically powerful, they usually led to formula in which multiple integrals had to be solved in order to obtain the marginal posterior distributions used for a complete Bayesian inference. Because these integrals could not be calculated, even using approximate methods, Bayesian inference was based

on the mode of posterior distributions, often giving results rather similar to the REML approaches. However, till today people do not readily change to Bayesians from frequentists due to many reasons.

The frequentists' way of inference is based on how a large number of estimates would be distributed around the true value if a large number of samples were taken, whereas Bayesians examine the probability distribution of the true value, given the data. For a frequentist, the true value is usually fixed and the sample is variable, whereas for a Bayesian the sample is fixed and the parameter of interest is a random variable. Some statistical concepts currently used in animal breeding do not have a Bayesian interpretation, for example, "bias" and the difference between fixed and random effects. In a Bayesian context "bias" does not exist, because conceptual repetitions of the experiment are not considered. Also, all effects are random because the Bayesian way of expressing uncertainty is to draw density functions of all unknowns, and thus all unknowns are considered random variables. This can be surprising to an animal breeder working with BLUP (Henderson, 1975) or REML (Patterson and Thompson, 1971), but the property of unbiasedness has been discussed even within the frequentist methods. Inferences obtained from both schools are not always coincident, particularly for small samples and when the Bayesian analysis uses prior information. Some problems that have no solution (or have only a rough approximation) in the frequentist methods can be solved unambiguously with the Bayesian approach. There are many cases in animal breeding in which the frequentist approach gives an accurate and rapid answer and Bayesian methods are not needed (BLUP). Problems faced by frequentists are difficulties with obtaining multivariate REML estimation of the variance components when the database is large, and with the problem of taking into account the error of estimation of variance components in the prediction of breeding values.

Estimation of different variance components (VC) from the total phenotypic variance in a trait is very important in animal breeding. Many workers in the past have tried to find out various approaches for estimating the VC through statistical techniques. According to Henderson (1975) and Schaeffer (1984), accurate estimates of VC are important because prediction of error variances for predicted random effects (e.g., breeding values) increases as differences between estimated and true values of VC increase. These days REML is considered as the method of choice for estimating VC (Meyer, 1990). The use of REML and DF-REML in animal breeding has increased significantly as various softwares for REML procedure were made available by research workers. Some examples of these programs are DFREML 3.0.beta (Meyer, 1998), MTDFREML (Boldman et al., 1995), VCE (Groeneveld, 1994) and Wombat (Meyer, 2010). The use of Bayesian analysis in animal breeding did not have many takers in the past research community. A bold effort by Van Tassell and Van Vleck in 1995 for constructing a program for variance component estimation by Gibbs sampling resulted in the formation of multiple-trait Gibbs sampler for animal models (MTGSAM) programs, which are developed to implement the Gibbs sampling (GS) algorithm for Bayesian analysis of a broad range of animal models (Van Tassell and Van Vleck, loc. cit). These programs expand the methods available for statistical analysis of animal breeding data.
The number of minimum iterations which are required as the burn in and the total number of iterations for estimation of the correct means of posterior distributions are very important while performing the Gibbs sampling analysis in animal breeding. Raftery and Lewis made efforts successfully to put forth the idea in the form of an algorithm which was converted to the Fortran program (Gibbsit) for calculation of burn in and minimum number of iterations required for the correct estimates of posterior means in the Gibbs sampling (Raftery and Lewis, 1996).

Research Experiences

Using Bayesian method, we can integrate over varying degree of uncertainty in the different aspect of the analysis. It is useful in the post genomic world of analyzing large, noisy biological data sets. Bayesian methods do not require any particular regularity conditions on the probability model, do not depend on the existence of sufficient statistics of finite dimension, do not rely on asymptotic relations and do not require the derivation of any sampling distribution.

Van Tassell and Van Vleck (1996) have described in their paper that if the Gibbs sampling is used giving flat priors and compared with the

REML procedure estimates then the outputs are similar to each other. However, they may vary a little and non-significantly if priors are used in the Gibbs sampling procedure. They used MTGSAM computer program for the computation of genetic parameters. This software is a leap ahead in the analysis of animal breeding data by the Bayesian approach. A further attempt to use another software for animal breeding data was done by Damgaard (2007). He concluded in his paper that Winbugs can be used to make inferences in small-sized, quantitative, genetic data sets applying a wide range of animal models that are not yet standard in the animal breeding literature. However, not many people are using the Winbugs for genetic data analysis, but its use can be explored in the future. The use of Gibbs sampler for estimation of genetic parameter was done in Jack Russell Terrier (JRT) dog by Famula et al., (2007). The estimation of heritability of deafness in the JRT was 0.22 when deafness was considered a binary (normal/deaf) trait and 0.31 when deafness was considered a three-category trait (normal/unilateral/bilateral deafness). The influence of coat colour in the incidence of JRT deafness was statistically significant, indicating that dogs with more white are more likely to be deaf. Complex segregation analysis revealed a model of a single locus with a large effect on the binary measure of hearing loss is not supported. Stock et al. (2007) used Gibbs sampling in horse population. Their results showed that bias of heritability estimates was -6% to +6% for the continuous trait, -6% to +10% for the binary traits of moderate heritability, and -21% to +25% for the binary traits of low heritability. Additive genetic correlations were mostly underestimated between the continuous trait and binary traits of low heritability, under- or overestimated between the continuous trait and binary traits of moderate heritability and overestimated between two binary traits. Use of trait information on two subsequent generations of animals increased effective sample size (ESS) and reduced bias of parameter estimates more than mere increase of the number of informative animals from one generation. Consideration of genotype information as a fixed effect in the model resulted in overestimation of polygenic heritability of the QTL trait, but increased accuracy of estimated additive genetic correlations of the QTL trait. Yague et al., (2009) estimated genetic parameters for days to first insemination (DFI), days from first insemination to conception (FIC), number of inseminations per conception (IN), days open (DO), gestation length (GL) and calving interval (CI) by multitrait Bayesian procedures. In his literature, estimates of the mean of posterior distribution of the heritability of DFI, FIC, IN, DO, GL and CI were 0.050, 0.078, 0.071, 0.053, 0.037 and 0.085 respectively and the corresponding estimates for repeatability of these traits were 0.116, 0.129, 0.147, 0.138, 0.082 and 0.132 respectively. No significant genetic correlations associated to DFI or GL were found. However, genetic correlations between the other four analyzed traits were high and significant. Genetic correlations between FIC and IN, DO and CI were similar and higher than 0.85. Genetic correlations of IN–DO and IN–CI were over 0.65. The highest genetic correlation was estimated for the pair DO–CI (0.992) that can be considered the same trait in genetic terms. Details of this study are given here to give an importance of the Gibbs sampling as the method of choice for recent trends in the genetic parameter estimation.

In many problems of statistical inference, objective and universally agreed contextual information is available on the parameter values. This information is usually very difficult to handle within the framework of conventional statistics, but it is easily incorporated into a Bayesian analysis by simply restricting the prior distribution to the class of priors which are compatible with such information. Animal breeding data which is categorical can be very well evaluated by using the Gibbs sampling procedure. In India, there are no good reports of use of Gibbs sampling for analysis of the animal breeding data till now. With the advent of the computer packages such as Gibbsit, MTGSAM, WINBUGS, the computation part is made easy. Breeders with access to the scientific records for the breeding data can give a brave attempt in this direction, which may lead to a new era of breeding data analysis in our country.

References

Boldman, K.G., Kriese, L.A., Van Vleck, L.D., Kachman, S.D., 1995. A Manual for Use of MTDFREML. A set of programs to obtain estimates of variances and covariances [Draft]. Lincoln, Department of Agriculture, p. 120.

Damgaard, L.H., 2007. How to use Winbugs to draw inferences in animal models. Journal of Animal Science

85, 1363-1368.

Evans, M., Swartz, T., 1995. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. Statistical Science 10, 254-272.

Famula, T.R., Cargill, E.J., Strain, G.M. 2007. Heritability and complex segregation analysis of deafness in Jack Russell Terriers. BMC Veterinary Research 3, 31.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398-409.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. IEE Transactions on Pattern Analysis and Machine Intelligence 6, 721-741.

Groeneveld, E., 1994. VCE-A multivariate multimodel REML (co)variance component estimation package. In: Proc. 5th World Congr. Genet. Appl. Livest. Prod. 22, 46.

Harvey, W.R., 1977. Users Guide for mixed model least squares and maximum likelihood computer program. Columbus: The Ohio State University, p. 76.

Harville, D.A., 1977. Maximum Likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72, 320-338.

Henderson, C.R., 1953. Estimation of variance and covariance components Biometrics, 9, 226-252.

Henderson, C.R., 1976. Multiple Trait sire evaluation using the relationship matrix. Journal of Dairy Science 59, 769-774.

Lunn, D.; Thomas, A.; Best, N., Spiegelhalter, D., 2000. WinBUGS: A Bayesian modeling framework: concepts, structure, and extensibility, Statistics and Computing, 10, 325-337

Metropolis, N., Ulam, S., 1949. The Monte Carlo method. Journal of the American Statistical Association 44, 335-341.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller. E., 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1091.

Meyer, K., 1990. Present status of knowledge about statistical procedures and algorithms to estimate variance and covariance components. Proc. 4th World Cong. Appl. Livest. Prod., Edinbugh, Scotland, XIII, 407-418.

Meyer, K., 1998. DFREML programs to estimate variance components by restricted maximum likelihood using derivative free algorithm. Version 3.0. Beta -user notes.

Meyer, K., 2010. WOMBAT. A program for mixed model analyses by restricted maximum likelihood. Version 1.0 -user notes, AGBU.

Patterson, H.D., Thompson, R., 1971. Recovery of interblock information when block sizes are unequal. Biometrika 58, 545-554.

Raftery, A.E., Lewis, S.M., 1996. The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms. In: Gilks, W.R., Spiegelhalter, D.J., Richardson, S. eds. Practical Markov Chain Monte Carlo. Chapman & Hall, London, UK.

SAS Institute Inc., 2006. Preliminary Capabilities for Bayesian Analysis in SAS/STAT Software, SAS Institute Inc., Cary, NC, USA.

Schaeffer, L.R., 1984. Sire and cow evaluation under multiple trait models. Journal of Dairy Science 67, 1567.

Smith, A.F.M., 1991. Bayesian computational methods. Philosophical Transactions of the Royal Society London, 337, 369–386.

Stock, K.F., Distl, O., Hoeschele, I., 2007. Bayesian estimation of genetic parameters for multivariate threshold and continuous phenotypes and molecular genetic data in simulated horse populations using Gibbs sampling. BMC Genetics 8, 19.

Tanner, M.A., 1996. Tools for statistical inference, 3rd ed. Springer-Verlag, New York.

Thomas, A., O'Hara, B., Ligges, U., Sturtz, S., 2006. Making BUGS Open, R. News 6, 12-17.

Van Tassell, C., Van Vleck, L.D., 1996. Multiple-trait Gibbs sampler for animal models: Flexible programs for Bayesian and likelihood-based (co)variance component inference. Journal of Animal Science 74, 2586-2597.

Van Tassell, C.P., Van Vleck, L.D., 1995. A Manual for Use of MTGSAM. A set of Fortran programs to apply Gibbs sampling to animal models for variance component estimation. U.S. Department of Agriculture, Agricultural Research Service.

Yague, G., Goyache, F., Becerra, J., Moreno, C., Sanchez, L., Altarriba, J., 2009. Bayesian estimates of genetic parameters for pre-conception traits, gestation length and calving interval in beef cattle. Animal Reproduction Science 114, 72–80.