Journal of
Animal Breeding and Genetics

J. Anim. Breed. Genet. ISSN 0931-2668

SHORT COMMUNICATION

GSEVM v.2: MCMC software to analyze genetically structured environmental variance models

N. Ibáñez-Escriche[1], M. Garcia[2] & D. Sorensen[3]

1 Genètica i Millora Animal- Centre IRTA Lleida, Lleida, Spain
2 Station d'amélioration génétique des animaux, INRA, Castanet-Tolosan cedex, France
3 Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, University of Aarhus, Tjele, Denmark

Keywords
Bayesian analysis; heterogeneous residual variance; MCMC software.

Correspondence
N. Ibáñez-Escriche, Genètica i Millora Animal-Centre IRTA Lleida, 25198 Lleida, Spain.
Tel: (34) 973 00 34 32; Fax: (34) 973 23 83 01;
E-mail: noelia.ibanez@irta.es

Received: 03 July 2009;
accepted: 11 October 2009

Summary

This note provides a description of software that allows to fit Bayesian genetically structured variance models using Markov chain Monte Carlo (MCMC). The GSEVM v.2 program was written in Fortran 90. The DOS and Unix executable programs, the user's guide, and some example files are freely available for research purposes at http://www.bdporc.irta.es/estudis.jsp. The main feature of the program is to compute Monte Carlo estimates of marginal posterior distributions of parameters of interest. The program is quite flexible, allowing the user to fit a variety of linear models at the level of the mean and the logvariance.

Introduction

The aim of this note is to present a statistical program that allows implementing Bayesian-Markov chain Monte Carlo (MCMC) genetically structured environmental variance models. The GSEVM v.2 (genetically structured environmental variance model) program was written in Fortran 90. The DOS (windows application) and Unix executable programs, the user's guide, and some example files are freely available for research purposes at http://www.bdporc.irta.es/estudis.jsp. The program has been tested and a previous version has been used in several studies (Gutiérrez *et al.* 2006; Ibáñez-Escriche *et al.* 2007, 2008, 2009).

Theoretical background

Model

The software fits models with genetically structured residual variance (SanCristobal-Gaudy *et al.* 1998). It is assumed that the sampling distribution of data y is Gaussian, of the form

$$\mathbf{y}|\mathbf{a},\mathbf{b},\mathbf{p},\mathbf{a}^*,\mathbf{b}^*,\mathbf{p}^* \sim N(\boldsymbol{\mu}, diag(\sigma_i^2)_{i=1}^n)$$

where $diag(\sigma_i^2)_{i=1}^n$ is the diagonal matrix with diagonal entries σ_i^2, n is the length of $\mathbf{y}$,

$$\log((\sigma_i^2)_{i=1}^n) = \mathbf{X}^*\mathbf{b}^* + \mathbf{Z}^*\mathbf{a}^* + \mathbf{W}^*\mathbf{p}^*,$$

and

$$\boldsymbol{\mu} = (\mu_{i=1}^n) = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wp}.$$

Vectors $\mathbf{b}$ and $\mathbf{b}^*$ contain systematic effects, $\mathbf{a}$ and $\mathbf{a}^*$ additive genetic effects, $\mathbf{p}$ and $\mathbf{p}^*$ contain other effects such as permanent environmental effects and $\mathbf{X}$, $\mathbf{X}^*$, $\mathbf{Z}$, $\mathbf{Z}^*$, $\mathbf{W}$ and $\mathbf{W}^*$ are known incidence matrices. Star (*) index indicates that parameters and matrices are related to the environmental variance (σ_i^2).

Prior distributions

The GSEVM v.2 program assumes the following prior distributions for the location parameters: a uniform distribution *Unif* $(-\infty, +\infty)$ is assigned to

doi:10.1111/j.1439-0388.2009.00846.x

the systematic effects **b** and **b***, while a normal distribution is assumed for the additive genetic effects **a** and **a***

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{a}* \end{pmatrix} | \mathbf{G} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \mathbf{G} \otimes \mathbf{A} \right)$$

$$\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_{a^*} \\ \rho\sigma_a\sigma_{a^*} & \sigma_{a^*}^2 \end{pmatrix}$$

where **A** is the additive genetic relationship matrix, σ_a^2 is the additive genetic variance at the level of the mean of the trait, $\sigma_{a^*}^2$ is the additive genetic variance at the level of the environmental variance of the trait, ρ is the coefficient of genetic correlation and $\otimes$ denotes the Kronecker product. The chosen prior for σ_a^2 and $\sigma_{a^*}^2$ is a scaled inverted χ^2 distribution and ρ is a priori uniformly distributed in (−1, 1). The vectors **p** and **p*** are assumed to be independent with distributions,

$$\left(\boldsymbol{p} | \ \sigma_p^2\right) \sim \ \mathbf{N}\left(\mathbf{0}, \mathbf{I}_{\mathrm{p}} \otimes \sigma_p^2\right)$$

$$\left(\boldsymbol{p}^* | \ \sigma_{p^*}^2\right) \sim \ \mathbf{N}\left(\mathbf{0}, \mathbf{I}_{\mathrm{p}} \otimes \sigma_{p^*}^2\right)$$

where $\mathbf{I}_{\mathrm{p}}$ is the identity matrix with order equal to the number of random effects, and σ_p^2 and σ_{p*}^2 are variance components. Similarly to the additive variances, scaled inverted χ^2 prior distributions are assumed for σ_p^2 and σ_{p*}^2.

Algorithm and implementation

The model is nonstandard and this entails considerable computational challenges. The following strategy results in acceptable MCMC behaviour. A conventional blocking Gibbs sampling (see, Sorensen & Gianola 2002) is used for **b** since it is the only full conditional distribution known. The Metropolis Hasting algorithm with a random walk proposal is used to sample from the posterior distributions of σ_a^2,$\sigma_{a^*}^2$,σ_p^2,$\sigma_{p*}^2$and the correlation coefficient ρ. The posterior distributions of the parameters (**p**, **p***) and (**a**, **a***) are sampled using a Metropolis Hastings with a Langevin Hastings proposal. To avoid problems when sampling additive values (**a**, **a***) with very different posterior variances a reparameterization is performed $(a = \sigma_a^2 \gamma D^{1/2} T^T)$,where γ has a standard normal distribution, and T and D correspond to the factorization $A = TDT^T$ of A (Henderson 1976). A more detailed description of the algorithm can be found in Sorensen & Waagepetersen (2003) and in Waagepetersen *et al.* (2008).

Main characteristics

The GSEVM v.2 software was programmed in Fortran 90. In essence, this program has five main features. First, it provides marginal posterior distributions of parameters of interest. Second, it computes the value of the Deviance Information Criterion (DIC) proposed by (Spiegelhalter *et al.* 2002). Third, the user can assign different models for the mean trait and its variance. Fourth, the program allows defining prior distributions for variance parameters. Fifth, the program runs in Windows and Unix systems.

Inputs

The GSEVM v.2 software is driven by a file called 'parameter.gse'. This file contains the path and name of the parameter file to be used in the analysis. The parameter file defines the input files, attributes, models and the parameters of the prior distributions of the variance components. The input files used by GSEVM v.2 must be prepared as an ASCII file with separated columns, where both tab delimiters and common delimiters are allowed. The program requires three obligatory input files: (i) a parameter file, (ii) a data file (up to 100 000 records), and (iii) a pedigree file. Also, two optional files can be used: (i) an inbreeding file, where the inbreeding coefficient is indicated, and (ii) a generation file where the generation and genetic line are defined for all individuals. A detailed description of the input files, including their structure and format can be found in the user's guide.

Outputs

The GSEVM v.2 program provides six output files. If the names and paths of these files are not specified in the parameter file, the program creates the following default names in the working directory: SUMMARY.out, BREEDINGV.out, VARIANCES.out, SYSTEMATICS.out, SYSTEMATICS1.out and MEGELI.out. Otherwise, the program takes the names and paths given in the parameter file. A detailed summary of the analysis is reported in the file 'SUMMARY.out'. File 'BREEDINGV.out' includes the posterior mean of the additive values **a** and **a*** for all individuals. Samples from the marginal posterior distributions of variances and correlation are collected in file 'VARIANCES.out', and samples from the marginal posterior distributions of the systematic effects **b** and **b*** are found in files 'SYSTEMATICS.out' and 'SYSTEMATICS1.out', respectively. File

'MEGELI.out' contains averages of the sampled additive values of **a** and **a*** for each generation and line in the corresponding iteration.

The output files provided by the GSEVM program gives a high amount of useful information that can be used for statistical analysis of environmental variance models.

Acknowledgments

The authors are grateful to Rasmus Waagepetersen for computational and statistical input over the years.

References

Gutiérrez J.P., Nieto B., Piqueras P., Ibanez N., Salgado C. (2006) Genetic parameters for canalisation analysis of litter size and litter weight traits at birth in mice. *Genet. Sel. Evol.*, **38**, 445–462.

Henderson C.R. (1976) A simple method for the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, **32**, 69–83.

Ibáñez-Escriche N., Varona L., Sorensen D., Noguera J.L. (2007) A study of heterogeneity of environmental variance for slaughter weight in pigs. *Animal*, **2**, 2209–2226.

Ibáñez-Escriche N., Moreno A., Nieto B., Piqueras P., Salgado C., Gutierrez J.P. (2008) Genetic parameters related to environmental variability of weight traits in a selection experiment for weight gain in mice; signs of correlated canalized response. *Genet. Sel. Evol.*, **40**, 279–293.

Ibáñez-Escriche N., Sorensen D., Waagepetersen R., Blasco A. (2009) Selection for environmental variation: a statistical analysis and power calculations to detect response. *Genetics*, **180**, 2209–2226.

SanCristobal-Gaudy M., Elsen J.M., Bodin L., Chevalet C. (1998) Prediction of the response to a selection for canalisation of a continuous trait in animal breeding. *Genet. Sel. Evol.*, **30**, 423–451.

Sorensen D., Gianola D. (2002) *Likelihood, Bayesian and Markov chain Monte Carlo methods in quantitative genetics*. Springer-Verlag, New York.

Sorensen D., Waagepetersen R. (2003) Normal linear models with genetically structured variance heterogeneity: a case of study. *Genet. Res.*, **82**, 207–222.

Spiegelhalter D.J., Best N.G., Carlin B.P., van der Linde A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc., Ser. B*, **64**, 583–639.

Waagepetersen R., Ibáñez-Escriche N., Sorensen D. (2008) A comparison of strategies for Markov Chain Monte Carlo computation in quantitative genetics. *Genet. Sel. Evol.*, **40**, 161–176.