BESSiE
A PROGRAM FOR MULTIVARIATE LINEAR MODEL BLUP AND BAYESIAN ANALYSIS OF LARGE SCALE GENOMIC DATA.

V. Boerner and B. Tier

Animal Genetics and Breeding Unit[1], University of New England
Armidale, 2351, NSW, Australia

SUMMARY

BESSiE is a software designed for uni- and multivariate analysis of linear mixed models including large scale genomic data.

BESSiE facilitates models allowing for various fixed and random effects, and for observations on continuous or categorical scales, and implements different Bayesian algorithms for the prediction of effects of genetic markers (e.g. BayesA, BayesB, BayesCπ and BayesR), GBLUP and SNP-BLUP.

INTRODUCTION

Various software packages are available for the analysis of phenotypic observations with linear mixed models in quantitative genetics, which can be categorised by the employed algorithm for inferring dispersion and location parameters of the modelled factors: a) Restricted Maximum Likelihood (REML) based software, and b) Bayesian and Markov Chain Monte Carlo (MCMC) based software. While various REML software packages specifically designed for quantitative genetics are widely used and well documented,(e.g. ASREML (Gilmour *et al.* 2009), WOMBAT,(Meyer 2007), DMU (Madsen *et al.* 2014), REMLF90 (Misztal *et al.* 2002), VCE (Groeneveld *et al.* 2010)), software packages employing Bayesian and MCMC methodology are less common (GIBBSF90 and THRGIBBSF90 (Misztal *et al.* 2002),BAYESR (Moser *et al.* 2015), MCMCglmm,(Hadfield 2010)), but only GIBBSF90 AND THRGIBBSF90 are explicitly designed for the application on large data sets and complex models in quantitative genetics, and therefore provide results in a reasonable amount of time. The relatively small number of Bayesian and MCMC software packages for quantitative geneticist may reflect the disadvantage of this methodology in terms of processing time. In addition, large scale genomic marker data emerging in the late 2000 easily fit into existing REML software via approaches like GBLUP or single marker regression. In contrast several Bayesian algorithms for sampling dispersion and location parameters of genomic markers have been proposed (Meuwissen *et al.* 2001; Habier *et al.* 2011; Erbe *et al.* 2012), which differ only slightly but require adjustments in the software source code, thus making it more difficult to develop and maintain a software which covers all.

The aim of this paper is to describe the software BESSiE which is designed for uni- and multivariate BLUP and Bayesian analysis of linear mixed models in quantitative genetics allowing for various factors, algorithms, large scale genomic data and both continuous as well as categorical observations.

SOFTWARE DESCRIPTION

BESSiE is written in FORTRAN90, command line operated, parameter file driven and

[1]A joint venture of the NSW Department of Primary Industry and the University of New England

comes with an extensive manual. It is available for 64bit Unix-like operation systems only, and is optimised for Intel architecture.

The super-set model to be fitted in BESSiE may be written as:

$$\begin{pmatrix} y_1 \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} X_1 & . & 0 \\ . & . & . \\ 0 & . & X_n \end{pmatrix} \begin{pmatrix} b_1 \\ . \\ b_n \end{pmatrix} + \begin{pmatrix} Z_1 & . & 0 \\ . & . & . \\ 0 & . & Z_n \end{pmatrix} \begin{pmatrix} u_1 \\ . \\ u_n \end{pmatrix} + \begin{pmatrix} Q_1M & . & 0 \\ . & . & . \\ 0 & . & Q_nM \end{pmatrix} \begin{pmatrix} g_1 \\ . \\ g_n \end{pmatrix} + \begin{pmatrix} e_1 \\ . \\ e_n \end{pmatrix}$$

where $(y_1, ., y_n)'$, $(b_1, ., b_n)'$, $(u_1, ., u_n)'$, $(g_1, ., g_n)'$ and $(e_1, ., e_n)'$ are vectors of phenotypic observations of linear or categorical scale, fixed effects, random non-marker effects and random marker effects, X, Z and Q are matrices relating the effects to their respective observations, M is a matrix of marker genotypes of dimension "number of genotyped individuals"×"number of markers" and the subscripts are for trait 1 to n. Values in X may be dummy variables or linear co-variables, where for the latter the order of polynomial regression is user-defined. Values in $(u_1, ., u_n)$ are assumed to be distributed $N([0, ., 0]', A \otimes \Sigma)$, $N([0, ., 0]', G \otimes \Sigma)$, $N([0, ., 0]', I \otimes \Sigma)$ or $N([0, ., 0]', K \otimes \Sigma)$, where A is the pedigree derived numerator relationship matrix, G is a relationship matrix derived from genetic markers, I is an identity matrix, K is an unknown matrix of dimension "number of factor levels"× "number of factor levels" provided by the user, and Σ is a co-variance matrix of factors. Note that all random non-marker effects can be fitted together.

The algorithm to obtain dispersion and location parameters when trait observations are of categorical scale is described in Sorensen *et al.* (1995) and Albert and Chib (1993).

Random effects of genetic markers $(g_1, ., g_n)'$ can be obtained from BayesA and BayesB (Meuwissen *et al.* 2001), BayesCπ (Habier *et al.* 2011), BayesR (Erbe *et al.* 2012) or ridge regression SNP-BLUP (Piepho 2009). For BayesA, BayesB and BayesCπ, all relevant parameters of the algorithms and the prior distributions of marker variances are taken from the related publications, but can also be set by the user.

Residuals are assumed to be distributed $N([0, ., 0]', I \otimes R)$, where R is the residual co-variance matrix of dimension $n \times n$. However, to account for observations with different residual variances (e.g. de-regressed breeding values), a co-variance Ω can be modelled, where Ω is a block-diagonal matrix containing $\omega_1\sigma^2_{e_1}$ to $\omega_n\sigma^2_{e_n}$ in the diagonal elements of the diagonal blocks, and $\sqrt{\omega_1\omega_n}\sigma_{e_{1,n}}$ in the diagonal elements of the off-diagonal block which links trait 1 and trait n, where ω_1 and ω_n are the weights of trait 1 and n, and $\sigma^2_{e_1}$, $\sigma^2_{e_n}$ and $\sigma_{e_{1,n}}$ are the residual variances and co-variance of both the traits.

In multivariate analysis using BayesA, BayesB, BayesCπ or BayesR effects of genetic markers are estimated from

$$\left(\begin{bmatrix} Q_1M & . & 0 \\ . & . & . \\ 0 & . & Q_nM \end{bmatrix}' R^{-1} \begin{bmatrix} Q_1M & . & 0 \\ . & . & . \\ 0 & . & Q_nM \end{bmatrix} + \begin{bmatrix} \sigma^2_1 & . & 0 \\ . & . & . \\ 0 & . & \sigma^2_n \end{bmatrix}^{-1} \right) \begin{pmatrix} g_1 \\ . \\ g_n \end{pmatrix} =$$

$$\begin{pmatrix} Q_1M & . & 0 \\ . & . & . \\ 0 & . & Q_nM \end{pmatrix}' R^{-1} \left(\begin{bmatrix} y_1 \\ . \\ y_n \end{bmatrix} - \begin{bmatrix} X_1 & . & 0 \\ . & . & . \\ 0 & . & X_n \end{bmatrix} \begin{bmatrix} b_1 \\ . \\ b_n \end{bmatrix} - \begin{bmatrix} Z_1 & . & 0 \\ . & . & . \\ 0 & . & Z_n \end{bmatrix} \begin{bmatrix} u_1 \\ . \\ u_n \end{bmatrix} \right)$$

where σ^2_1 to σ^2_n are diagonal matrices of dimension "number of markers"× "number of markers" of which elements contain the marker variances generated according to the Bayesian method specified for trait 1 to n. The co-variances between the effects of a genetic marker

on trait 1 to n are assumed to be zero.

BESSiE has no hard coded limitations in terms of number of traits, factors, genotypes and markers, and has been tested on very large data sets.

As an example, a bi-variate analysis with 4,420 individuals genotyped for 510,174 single nucleotide polymorphism (SNP), 19,549 individuals in the pedigree, 7 fixed effects and a polygenic random effect per trait, and SNP effects modelled according to BayesR with 4 distributions requires 4.3GB of RAM and about 7 real time seconds on an Intel(R) Core(TM) i7-3770 processor to sample all location and dispersion parameters once.

BESSiE comes without any warranties and can be used by the scientific community free of charge. It can be downloaded from http://turing.une.edu.au/~agbu-admin/BESSiE/.

REFERENCES

Albert J. H. and Chib S. (1993) *J. Am. Stat. Assoc.* **88**:669 .

Erbe M., Hayes B. J., Matukumalli L. K., Goswami S., Bowman P. J., Reich C. M., Mason B. A. and Goddard M. E. (2012) *J. Dairy Sci.* **95**(7):4114.

Gilmour A. R., Gogel B. J., Cullis B. R. and Thompson R. (2009) ASReml User Guide Release 3.0, VSN International Ltd, Hemel Hempstead.

Groeneveld E., Kovac M. and Mielenz N. (2010) VCE user's guide and reference manual version 6.0.

Habier D., Fernando R. L., Kizilkaya K. and Garrick D. J. (2011) *BMC Bioinformatics* **12**:186.

Hadfield J. D. (2010) *J. Stat. Softw.* **33**(2).

Madsen P., Jensen J., Labouriau R., Christensen O. F. and Sahana G. (2014) in Proc. 10th. WCGALP, Vancouver, Canada.

Meuwissen T. H., Hayes B. J. and Goddard M. E. (2001) *Genetics* **157**(4):1819.

Meyer K. (2007) *J. Zhejiang Univ. Sci. B* **8**(11):815.

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T. and Lee D. (2002) in Proc. 7th. WCGALP, Montpellier, France.

Moser G., Lee S. H., Hayes B. J., Goddard M. E., Wray N. R. and Visscher P. M. (2015) *PLoS Genet* **11**(4):e1004969.

Piepho H. P. (2009) *Crop Sci.* **49**(4):1165.

Sorensen D. A., Andersen S., Gianola D. and Korsgaard I. (1995) *Genet. Sel. Evol.* **27**:229 .