Bayesian detection of single gene determining quantitative trait by threshold model*

A. Dobek[1], T. Szwaczkowski[2], K. Moliński[1], E. Skotarczak[1]

[1]Department of Mathematical and Statistical Methods,
[2]Department of Genetics and Animal Breeding, August Cieszkowski Agricultural University of Poznań, Poland

SUMMARY

A method for the identification of single genes determining quantitative traits is given. The starting point is the analysis of a threshold model with one threshold of an unobservable variable values, the variable which determines the phenotype of an individual. The values of this variable depend, beside the single gene, on fixed effects, polygenic effects and in the case of repeated measurements of these traits also permanent environmental effects.

Key words: Gibbs sampling, quantitative trait, single genes, threshold model.

1. Introduction

In the last years in literature on genetics and animal breeding, two areas are intensively developed. The first one deals with the identification of single loci of animal traits, both in the genetic and physical genome mapping. Several methods of molecular biology and statistical procedures are used for this purpose.

The other area is connected with the revision of genetic improvement strategies. In the last decades the improvement of livestock has focused on an increase in production traits. Unfortunately, intensive production is connected with a bigger

* Paper supported by Ministry of Education and Science grant no. 2 P06D 049 28

metabolism which obviously leads to a weakness of organisms, with several negative consequences. Therefore more attention is now directed towards functional traits, usually not continuous ones, determined by the number of genes. These traits (variables), known in the literature as threshold traits, are rather complicated for statistical analysis. A number of studies concerns an estimation of threshold animal model parameters under polygenic inheritance assumption. For such a model with a number of genetic effects (direct additive, maternal additive), we have already obtained some results concerning the fertility and hatchability of laying hens (Moliński et al. 2003, Dobek et al. 2003, Skotarczak et al. 2004).

The main purpose of this paper is to find a method for the identification of single genes determining these traits. The starting point is the analysis of a threshold model with one threshold of an unobservable variable values, the variable which determines the phenotype of an individual. The values of this variable depend beside the single gene, on fixed effects, polygenic effects and in the case of repeated, measurements of these traits also permanent environmental effects.

2. Model

In the model describing unobservable variable we consider fixed effects, (year, flock), direct additive genetic effects, single genes effects, environmental effects and random errors. The model takes the following form

$$\mathbf{U} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{M}\mathbf{w} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e}$$

where:
U is an *s* vector of unobserved variables (e.g. liability),
X is a *s x b* design matrix of nongenetic effects,
β is a *b* vector of fixed effects,
$\mathbf{Z}_1$ is a *s x q* design matrix relating polygenic and single locus effects to observations,
M is a *s* x 3 random matrix containing information of genotype of each individual; each row of **M** has one of the following forms: [1, 0, 0], [0, 1, 0] or [0, 0, 1] corresponding to genotypes A_1A_1, A_1A_2 or A_2A_2 respectively,
w is the vector [h, d, -h]', where h is the effect of genotype A_1A_1, d is the effect of genotype A_1A_2 and –h is the effect of A_2A_2,
a is a *q* vector of random additive polygenic effects with distribution $N(\mathbf{0}, \sigma_a^2\mathbf{A})$, where **A** is the *q x q* relationship matrix and σ_a^2 is an additive polygenic variance,
$\mathbf{Z}_2$ is a *s x n* design matrix relating environmental effects to observations,
p is an *n* vector of permanent environmental effects with distribution $N(\mathbf{0}, \sigma_p^2\mathbf{I}_n)$, σ_p^2 is a permanent environment variance, and
e is an *s* vector of random effects with distribution $N(\mathbf{0}, \mathbf{I}_s)$.

3. Estimation

There are several methods of detection of major genes described in the literature. The oldest methods are based on the analysis of distributions (Fain 1978, Le Roy et al. 1992, Uimari et al. 1996). These methods are providing very general suggestions about the segregation of single gene if the effect was significantly large and the frequencies of genotypes were similar. More advanced methods, like segregation analysis introduced by Elston and Steward (1971) or Morton and MacLean (1974), modified by Knott et al (1992 a,1992b), allowed for conclusions also about the frequencies of genotypes.

New possibilities are due to the Bayesian analysis. The algorithm described by Guo and Thompson (1994) and Janss et al. (1995), allowed for the detection of major genes and the estimation of model parameters.

These methods are used to estimate all effects describing genotypes as well as the genes frequencies. In the estimation process the Gibbs sampling procedure has been used. The application of this procedure requires the knowledge of the form of marginal distributions, necessary in Bayesian method of parameter estimation. In our case we followed the Gibbs sampling based on the marginal distributions of the form

- the conditional posterior distribution for each of the nongenetic effects is:

$$\beta_i \mid \text{all other parameters} \sim N\left(u^{\beta}_{(i)}/n_i, 1/n_i\right)$$

where $\mathbf{u}^{\beta} = \mathbf{u} - \mathbf{Z}_1\mathbf{M}\mathbf{w} - \mathbf{Z}_1\mathbf{a} - \mathbf{Z}_2\mathbf{p}$, $u^{\beta}_{(i)}$ is the sum of observations in level (i) and n_i is the number of this observations;

- the conditional posterior distribution for h and d are:

$$h \mid \text{all other parameters} \sim N\left((u^{w}_{(1)} - u^{w}_{(3)})/(n_1 + n_3), 1/(n_1 + n_3)\right),$$

$$d \mid \text{all other parameters} \sim N\left(u^{w}_{(2)}/n_2, 1/n_2\right),$$

where $u^{w}_{(i)}$ and n_i denote the sum and the number of elements corresponding to the genotype i in the vector $\mathbf{u}^{w} = \mathbf{u} - \mathbf{X}\beta - \mathbf{Z}_1\mathbf{a} - \mathbf{Z}_2\mathbf{p}$;

- the conditional posterior distribution for genotype G_i is :

$$G_i \mid \text{all other parameters}$$

$$\propto \exp\left(\frac{-J_D}{2}u_i^e\right) P\left(G_i \mid G_{S_i}, G_{D_i}\right) \prod_{p(\text{progeny})} P\left(G_p \mid G_i, G_{\text{spouse}(p)}\right),$$

where u_i^e is the element of the vector $\mathbf{u^e} = \mathbf{u} - \mathbf{X\beta} - \mathbf{Z_1Mw} - \mathbf{Z_1a} - \mathbf{Z_2p}$ corres-ponding to an individual with genotype G_i and J_D =1 or 0 as the individual is, or is not observed; for the final progeny the last term disappears;

- the conditional posterior distribution for additive effects is:

$$a_i \mid \text{all other parameters} \sim N\left(\alpha_i, (n_i + a^{i,i}\lambda)^{-1}\right)$$

where $\alpha_i = \left(n_i + a^{i,i}\lambda\right)^{-1}\left(u_i^a - \lambda\sum_{j\neq i} a^{ij}a_j\right)$, $\lambda = \left(\sigma_a^2\right)^{-1}$, a^{ij} are the elements of the inverse of the relationship matrix **A** and u_i^a is the sum of elements of the vector $\mathbf{u^a} = \mathbf{u} - \mathbf{X\beta} - \mathbf{Z_1Mw} - \mathbf{Z_2p}$ corresponding to the individual i and n_i is the number of these elements;

- the conditional posterior distribution for environmental effects is:

$$p_i \mid \text{all other parameters} \sim N\left(\rho_i, (n_i + \psi)^{-1}\right)$$

where $\rho_i = \left(n_i + \psi\right)^{-1}u_i^p$, $\psi = \left(\sigma_p^2\right)^{-1}$, u_i^p is the sum of elements of the vector $\mathbf{u^p} = \mathbf{u} - \mathbf{X\beta} - \mathbf{Z_1Mw} - \mathbf{Z_1a}$ corresponding to the individual i and n_i is the number of these elements.

For the variance components the densities are:

$$\sigma_a^2 \mid \text{all other parameters} \sim \mathbf{a'A^{-1}a}/\chi^2_{q-2},$$
$$\sigma_p^2 \mid \text{all other parameters} \sim \mathbf{p'p}\ /\chi^2_{n-2}$$

In the case of threshold traits we observe in fact **y**, an *s* vector of bivariate variables corresponding to **u**. The elements of **u** are also sampled from posterior conditional distributions:

$$u_i \mid \text{all other parameters and } y_i = 1 \sim N\left(u_i^e, 1\right) \text{ (left-truncated)},$$
$$u_i \mid \text{all other parameters and } y_i = 0 \sim N\left(u_i^e, 1\right) \text{ (right-truncated)}$$

where u_i^e is the appropriate element of the vector $\mathbf{u^e} = \mathbf{X\beta} + \mathbf{Z_1Mw} + \mathbf{Z_1a} + \mathbf{Z_2p}$.

Values obtained by sampling from posterior conditional distributions are stored and the general inference is made by visualizing the marginal posterior densities for all the parameters.

REFERENCE

Dobek A., Szydłowski M., Szwaczkowski T. , Skotarczak E., Moliński K. (2003). Genetic variability of fertility and hatchability estimated by the Gibbs sampling under a threshold model. Journal of Animal and Feed Sciences 12: 307-314.

Elston R.C., Steward J. (1971). A general model for the genetic analysis of pedigree data. Human Heredity 21: 523-542.

Fain P.R. (1978). Characteristics of simple sibship variance tests for the detection of major loci and application to height, weight and spatial performance. Annales of Human Genetics 42: 109-120.

Guo S.W., Thompson E.A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. Biometrics 50, 417-432.

Janss L.L.G., Thompson R., Van Arendonk J.A.M. (1995). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theoretical and Applied Genetics 91: 1137-1147.

Knott S., Halley C.S., Thompson R. (1992a). Methods of segregation analysis for animal breeding data: a comparison of power. Heredity 68: 299-311.

Knott S., Halley C.S., Thompson R. (1992b). Methods of segregation analysis for animal breeding data: parameter estimates. Heredity 68: 313-320.

Le Roy P., Elsen J.M. 1992. Simple test statistics for major gene detection: numerical comparison. Theoretical and Applied Genetics 83: 635-644.

Moliński K., Szydłowski M., Szwaczkowski T., Dobek A., Skotarczak E.. (2003). An algorithm for genetic variance estimation of reproductive traits under a threshold model. Archiv fuer Tierzucht.46, 1: 85-91.

Morton N.E., MacLean C.J. (1974). Analysis of family resemblance. III. Complex segregation of quantitative traits. American Journal of Human Genetics 26: 489-503.

Skotarczak E., Szydłowski M., Dobek A., Moliński K., Szwaczkowski T. (2004). The algorithm of Bayesian estimation of maternal genetic and permanent maternal environmental variances in a two-trait binary threshold model. Czech Journal of Animal Science 49(2): 58-63.

Uimari P., Kennedy B.W., Dekkers J.C.M. (1996). Power and sensitivity of some simple test for detection of major genes in outbreed populations. Journal of Animal Breeding and Genetics 113: 17-28.