Analysis of single gene multitrait effects in livestock by the use of Gibbs sampling

Anita DOBEK[1], Krzysztof MOLIŃSKI[1], Maciej SZYDŁOWSKI[2], Tomasz SZWACZKOWSKI[2]

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

Abstract. The paper presents a method of multivariate data analysis described by a model which involves fixed effects, additive polygenic individual effects and the effects of a major gene. To find the estimates of model parameters, the maximalisation of likelihood function method is applied. The maximum of likelihood function is computed by the use of the Gibbs sampling approach. In this approach, following the conditional posterior distributions, values of all unknown parameters are generated. On the basis of the obtained samples the marginal posterior densities as well as the estimates of fixed effects, gene frequency, genotypic values, major gene, polygenic and error (co)variances are calculated. A numerical example, supplemented to theoretical considerations, deals with data simulated according to the considered model.

Key words: EM algorithm, major gene, Gibbs sampling, pleiotropic effects.

Introduction

A number of single genes with considerable effects were identified over the last decades. An exact estimation of these effects is possible in the absence of polygenes and the so-called loop pedigrees. However, in livestock populations major gene effects are masked by both polygenic and environmental variability. Several classical procedures for detection of major loci for single traits are presented in literature (LE ROY, ELSEN 1992, ELSTON, STEWARD 1971, MORTON, MACLEAN 1974). Recently, the bayesian methods have also been developed for

Received:September 4, 2000. Accepted: October 25, 2000.

Correspondence: A. DOBEK, Department of Mathematical and Statistical Methods, August Cieszkowski Agricultural University, ul. Wojska Polskiego 28, 60637 Poznań, Poland.

this purpose (GUO, THOMPSON 1992, JANSS et al. 1995, MIYAKE et al. 1999, DOBEK et al. 1999). However, the majority of these algorithms are based on unitrait models. On the other hand, it is known that many of these genes have multitrait effects resulting from pleiotropy and/or linkage phenomena (MERAT 1993). Hence, multitrait linear models are necessary to detect the presence of a major gene.

The main purpose of this paper is to present a method for the analysis of data described by a model which involves correlated individual polygenic effects and an effect of a major gene following the Mendelian inheritance rules. The model additionally contains fixed nongenetic effect, which increases the adequacy of the model. To find the estimates of the model parameters the maximalisation of likelihood function method is used. The maximum of likelihood function is calculated on the basis of a Monte Carlo procedure, named Gibbs sampling.

Model

In a mixed model it is assumed that the traits are influenced by the genotype at a single locus (major gene) and by polygenic effects. Moreover, it is assumed that the single locus is an additive biallelic locus with Mendelian transmission probabilities. The alleles at the single locus are A_1 and A_2 with probabilities θ and $1-\theta$, and therefore the genotypes are : A_1A_1,, A_1A_2, , A_2A_2 ($A_2A_1 = A_1A_2$), denoted by 1, 2 and 3, respectively. We further assume a homogeneous population of individuals with genotypes in Hardy-Weinberg equilibrium.

For a given configuration of major genotypes, say **G**, the observation model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{ZWM} + \mathbf{ZU} + \mathbf{E}, \tag{1}$$

where **Y** is the $n \times p$ matrix of observations, p is the number of traits (variables) measured on each individual; **X** is the $n \times r$ design matrix for nongenetic effects; β is the $r \times p$ matrix of fixed effects; **Z** is the $n \times q$ design matrix relating polygenic effects and major gene to observations; **W** is the $q \times 2$ random matrix containing information on the genotype of each individual. Each row of **W** has one of the following forms: $\{1, 0\}$, $\{-1, 0\}$ and $\{0, 1\}$, depending on whether the individual has the genotype A_1A_1, A_2A_2 or A_1A_2; **M** is the $2 \times p$ matrix with elements m_{1j} corresponding to the effect of the genotype A_1A_1 for the j-th trait and m_{2j} corresponding to the effect of the genotype A_1A_2 for the j-th trait; **U** is the $q \times p$ matrix of random polygenic effects, such that $cs(\mathbf{U}) \sim N_{qp}(\mathbf{0}, \Sigma_{\mathrm{U}} \otimes \mathbf{A})$, where Σ_{U} is the $p \times p$ additive genetic covariance matrix and **A** is a $q \times q$ relationship matrix; **E** is the $n \times p$ matrix of random errors, such that $cs(\mathbf{E}) \sim N_{np}(\mathbf{0}, \Sigma_{\mathrm{E}} \otimes \mathbf{I})$, where Σ_{E} is the $p \times p$ residual covariance matrix.

At the beginning it is necessary to assign the true prior distributions to all unknown parameters. Flat or uniform prior distributions are assigned to β and **M**. Further, it is assumed that the rows of **U** follow a normal distribution, i.e.

$$\mathbf{u}_i \sim N_p\left(\mathbf{0}, \Sigma_{\mathbf{U}}\right),$$

when the individual *i* is from the base population, so-called population of founders, and in other cases

$$\mathbf{u}_i \sim N_p\left(\frac{1}{2}\mathbf{u}_{\mathbf{S}_i} + \frac{1}{2}\mathbf{u}_{\mathbf{D}_i}, \frac{1}{2}\Sigma_{\mathbf{U}}\right),$$

where $\mathbf{u}_{\mathbf{S}_i}\left(\mathbf{u}_{\mathbf{D}_i}\right)$ are the polygenic effects of sire (dam).

The probability of genotypic configuration **G** can be written as

$$P(\mathbf{G}) \propto \prod_{i=founders} P(G_i) \prod_{i=nonfounders} P\left(G_i \middle| G_{S_i}, G_{D_i}\right),$$

where $P(G_i)$ is a function of θ and $P\left(G_i \middle| G_{S_i}, G_{D_i}\right)$ is the Mendelian segregation probability and therefore the distribution can be presented as

$$P(\mathbf{G}) \propto h(\mathbf{G})\exp\left\{\left(2\mathbf{1}_F^{'}\mathbf{1}_1 + \mathbf{1}_F^{'}\mathbf{1}_2\right)\ln\left[\theta/(1-\theta)\right] + 2\mathbf{1}_F^{'}\mathbf{1}\ln(1-\theta)\right\},$$

with $h(\mathbf{G})$ not depending on θ, and $\mathbf{1}_F$, $\mathbf{1}_1$ and $\mathbf{1}_2$ being the vectors of 1 and 0 with ones corresponding to founders, individuals with A_1A_1 and A_1A_2, respectively. The joint posterior density has the form

$$P\left(\beta, \mathbf{M}, \mathbf{G}, \mathbf{U}, \Sigma_{\mathbf{E}}, \Sigma_{\mathbf{U}} \middle| \mathbf{y}\right) \propto \left|\Sigma_{\mathbf{U}}\right|^{-\frac{q}{2}} \left|\mathbf{A}\right|^{-\frac{p}{2}} \left|\Sigma_{\mathbf{E}}\right|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} tr\Sigma_{\mathbf{E}}^{-1}\mathbf{V}_{\mathbf{E}}\right\} \tag{2}$$

$$\exp\left(-\frac{1}{2} cs(\mathbf{U})'(\Sigma_{\mathbf{U}} \otimes \mathbf{A})^{-1} cs(\mathbf{U})\right) h(\mathbf{G}) \exp\left[\left(2\mathbf{1}_F{'}\mathbf{1}_1 + \mathbf{1}_F{'}\mathbf{1}_2\right)\ln\frac{\theta}{1-\theta} + 2\mathbf{1}_F^{'}\mathbf{1}\ln(1-\theta)\right]$$

where $\mathbf{V}_{\mathbf{E}} = (\mathbf{Y} - \mathbf{X}\beta - \mathbf{ZWM} - \mathbf{ZU})'(\mathbf{Y} - \mathbf{X}\beta - \mathbf{ZWM} - \mathbf{ZU}) = \mathbf{Y}^{\mathbf{E}}{'}\mathbf{Y}^{\mathbf{E}}$.

Inferences about each of the unknowns are based on their respective marginal densities. Each marginal density is obtained by successive integration of the joint density (2) with respect to parameters other than the one of interest. To overcome the difficulties connected with the integration, the Gibbs sampling procedure is proposed.

The fully conditional posterior densities of all unknowns are needed to implement the Gibbs sampling. Each full conditional can be obtained by regarding all other parameters as known.

The conditional posterior distribution (c.p.d.) for each of the nongenetic effects is

$$\beta_{ij} \mid \text{all other parameters} \sim N\left(y_{(ij)}^{\beta}/n_i, \sigma_{e_j}^2/n_i\right),$$

where $\mathbf{Y}^{\beta} = \mathbf{Y} - \mathbf{ZWM} - \mathbf{ZU}$, $y^{\beta}_{(ij)}$ is the sum of observations at level (i) for the j-th variable, n_i is the number of these observations, and $\sigma^2_{e_j}$ is the diagonal element of $\Sigma_{\mathbf{E}}$.

The conditional probability densities for m_{1j} and m_{2j} are

$$m_{1j} \,|\, \text{all other parameters} \sim N\left(\left(y^{\mathbf{M}}_{(1j)} - y^{\mathbf{M}}_{(3j)} \right) / \left(n_{(1)} + n_{(3)} \right), \frac{1}{n_{(1)} + n_{(3)}} \sigma^2_{e_j} \right),$$

and $$m_{2j} \,|\, \text{all other parameters} \sim N\left(y^{\mathbf{M}}_{(2j)} / n_{(2)}, \frac{1}{n_{(2)}} \sigma^2_{e_j} \right),$$

where $y^{\mathbf{M}}_{(ij)}$ and $n_{(i)}$ denote the sum and the number of elements corresponding to genotype i for the j-th variable in the matrix $\mathbf{Y}^{\mathbf{M}} = \mathbf{Y} - \mathbf{X}\beta - \mathbf{ZU}$.

In establishing the c.p.d. for **G** and **u**, the procedure of blocking was implemented. The idea of blocking sires with its final progeny (i.e. progeny that are not parents themselves) used by JANSS et al. (1995) improves the efficiency of sampling scheme.

The c.p.d. for G_i is

– when i is a dam

$$G_i \,|\text{all other parameters} \propto \exp\left[-\frac{J_D}{2} \mathbf{y}^{\mathbf{E}\prime}_{\mathbf{i}} \Sigma^{-1}_{\mathbf{E}} \mathbf{y}^{\mathbf{E}}_{\mathbf{i}} \right] P\left(G_i \middle| G_{S_i}, G_{D_i} \right) \prod_{p(progeny)} P\left(G_p \middle| G_{Sp}, G_i \right),$$

where $\mathbf{y}^{\mathbf{E}}_{\mathbf{i}}$ is the i-th row of the matrix $\mathbf{Y}^{\mathbf{E}}$ corresponding to individual with genotype G_i and $J_D = 1$ or 0 if the individual is or is not observed,

– when i is a sire

$$G_i \,|\text{all other parameters} \propto \exp\left[-\frac{J_D}{2} \mathbf{y}^{\mathbf{E}\prime}_{(i)} \Sigma^{-1}_{\mathbf{E}} \mathbf{y}^{\mathbf{E}}_{(i)} \right] \prod_{p(nonfinal)} P\left(G_p \middle| G_o, G_{D_p} \right)$$

$$\times \prod_{f(final)} \sum_{G_f} \exp\left(-\frac{J_D}{2} \mathbf{y}^{\mathbf{E}\prime}_{\mathbf{Gf}} \Sigma^{-1}_{\mathbf{E}} \mathbf{y}^{\mathbf{E}}_{\mathbf{Gf}} \right) P\left(G_f \middle| G_i, G_{D_f} \right),$$

– when i is a final progeny

$$G_i \,|\text{all other parameters} \propto \exp\left[-\frac{J_D}{2} \mathbf{y}^{\mathbf{E}\prime}_{\mathbf{i}} \Sigma^{-1}_{\mathbf{E}} \mathbf{y}^{\mathbf{E}}_{\mathbf{i}} \right] P\left(G_i \middle| G_{S_i}, G_{D_i} \right)$$

In the case of founders the conditional probability $P\left(G_i \middle| G_{S_i}, G_{D_i} \right)$ is replaced by $P(G_i)$.

The conditional posterior density for polygenic effect $\mathbf{u_i}$ is normal, with expected value **E** and variance **V** of the form:

$$\mathbf{E} = \mathbf{V}\left[J_D \Sigma^{-1}_{\mathbf{E}} \mathbf{y}^{\mathbf{U}}_{\mathbf{i}} + J_P \Sigma^{-1}_{\mathbf{U}} \left(\mathbf{u}_{D_i} + \mathbf{u}_{S_i} \right) + \frac{1}{2} J_S \Sigma^{-1}_{\mathbf{U}} \sum_{p(progeny)} \left(2\mathbf{u}_p - \mathbf{u}_{D_p/S_P} \right) \right.$$

$$+\frac{1}{2}J_F\left[\left(2\Sigma_{\mathbf{E}}+\Sigma_{\mathbf{U}}\right)^{-1}\sum_{p(final)}\left(2\mathbf{y}_p^{\mathbf{U}}-\mathbf{u}_{D_p/S_p}\right)-\Sigma_{\mathbf{U}}^{-1}\sum_{p(final)}\left(2\mathbf{u}_p-\mathbf{u}_{D_p/S_p}\right)\right]\Bigg],$$

and

$$\mathbf{V}^{-1}=J_D\Sigma_{\mathbf{E}}^{-1}+\left(J_P+1\right)\Sigma_{\mathbf{U}}^{-1}+\frac{1}{2}J_S N_p\Sigma_{\mathbf{U}}^{-1}+\frac{1}{2}J_F\left[N_f\left(2\Sigma_{\mathbf{E}}+\Sigma_{\mathbf{U}}\right)^{-1}-N_f\Sigma_{\mathbf{U}}^{-1}\right],$$

where $\mathbf{y}_{\mathbf{i}}^{\mathbf{U}}$ is the *i*-th row of the matrix $\mathbf{Y}^{\mathbf{U}}=\mathbf{Y}-\mathbf{X}\beta-\mathbf{ZWM}$, $\mathbf{u}_{D_p/S_p}$ is the vector of polygenic effects of the dam or sire of the *p*-th individual (depending on whether the sire or dam is sampled) and N_p is the total number of progenies, N_f is the number of final progenies. The parameters $J_S = 1$ or 0 if the individual does not have any offspring, $J_P = 0$ or 1 if the individual is or is not a founder, and $J_F = 1$ or 0 if the individual is a sire and has final progeny or not.

For the covariance matrices the densities have inverted Wishart distributions, namely:

$$\Sigma_{\mathbf{U}}\,|\text{ all other parameters}\sim IW_p\left\lfloor\left(\mathbf{U'A}^{-1}\mathbf{U}\right)^{-1},q\right\rfloor,\text{ and}$$

$$\Sigma_{\mathbf{E}}\,|\text{ all other parameters}\sim IW_p\left\lfloor\left(\mathbf{Y}^{\mathbf{E}}{}'\mathbf{Y}^{\mathbf{E}}\right)^{-1},n\right\rfloor.$$

Estimation

The estimation procedure is realized in two steps. In the first step, following JANSS et al. (1995), a long Markov Gibbs chain is produced in the following way:
(i) set arbitrary initial values for β, **M**, **G**, **U**, Σ_{U}, Σ_{E},
(ii) generate β, **G**, **M**, and **U** from the c.p.d.,
(iii) generate Σ_{U} and Σ_{E} from the c.p.d.

Samples obtained in this way are stored. Then a general inference can be made by visualizing the marginal posterior densities for all parameters. The posterior density can be summarized by one or more statistics. For symmetric densities these can be the mean or variance. For the non-symmetric it can be the mode or median. The decision is up to the experimenter.

Having estimates obtained with this method, we are following the analysis in the second step. In this part, using the information about parameters, we are realizing the Monte-Carlo Expectation Maximization algorithm described by GUO and THOMPSON (1994). The algorithm is as follows:
(i) take initial estimates for β, M, Σ_{U}, Σ_{E}, and θ;
(ii) generate **G** and **U** as in the first step,
(iii) calculate the parameters β, M, Σ_{U}, Σ_{E}, and θ on the basis of the formulas:

Table. True, initial values and posterior means of model parameters

Parameter	True	Initial value	Mean
First trait:			
fixed effect (I):			
β_{11}	0.00	0.00	0.00
β_{21}	0.50	0.00	0.75
β_{31}	0.25	0.00	0.44
fixed effect (II):			
β_{41}	0.00	0.00	0.00
β_{51}	0.10	0.00	0.08
σ^2_{u1}	0.60	0.50	0.61
σ^2_{m1}*	0.80	0.50	0.79
σ^2_{e1}	0.20	0.50	0.30
m_{11}	2.00	1.00	1.88
m_{21}	0.00	1.00	0.15
Second trait:			
fixed effect (I):			
β_{12}	0.00	0.00	0.00
β_{22}	0.25	0.00	0.23
β_{32}	0.30	0.00	0.26
fixed effect (II):			
β_{42}	0.00	0.00	0.00
β_{52}	0.30	0.00	0.30
σ^2_{u2}	0.60	0.50	0.66
σ^2_{m2}*	0.40	0.50	0.39
σ^2_{e2}	0.20	0.50	0.22
m_{12}	2.00	1.00	2.02
m_{22}	1.00	1.00	1.06
Third trait:			
fixed effect (I):			
β_{13}	0.00	0.00	0.00
β_{23}	0.00	0.00	-0.06
β_{33}	0.00	0.00	0.07
fixed effect (II):			
β_{43}	0.00	0.00	0.00
β_{53}	0.00	0.00	0.01
σ^2_{u3}	0.60	0.50	0.53
σ^2_{m3}*	0.20	0.50	0.18
σ^2_{e3}	0.20	0.50	0.20
m_{13}	2.00	1.00	2.02
m_{23}	2.00	1.00	1.78

Common parameters:			
θ	0.70	0.50	0.72
Polygenic additive covariances:			
σ_{u12}	0.00	0.00	0.08
σ_{u13}	0.00	0.00	0.02
σ_{u23}	0.00	0.00	0.14
Single gene covariances:			
σ_{m12}**	0.50	0.00	0.51
σ_{m13}**	0.20	0.00	0.17
σ_{m23}**	0.20	0.00	0.17
Residual covariances:			
σ_{e12}	0.10	0.00	0.05
σ_{e13}	0.00	0.00	–0.03
σ_{e23}	0.10	0.00	0.05

* denotes the major gene variances for first, second and third trait, respectively.
** denotes the major gene covariances between these traits.

$$\hat{\beta}_{ij} = y^{\beta}_{(ij)} / n_i$$

$$\hat{m}_{1j} = \left(y^{\mathrm{M}}_{(1j)} - y^{\mathrm{M}}_{(3j)}\right) / \left(n_{(1)} + n_{(3)}\right) \qquad \hat{m}_{2j} = y^{\mathrm{M}}_{(2j)} / n_{(2)}$$

$$\hat{\Sigma}_{\mathrm{U}} = \frac{1}{q}\mathbf{Y}^{\mathbf{U}\prime}(\mathbf{ZAZ'})^{-1}\mathbf{Y}^{\mathbf{U}} \qquad \hat{\Sigma}_{\mathrm{E}} = \frac{1}{n}\mathbf{Y}^{\mathbf{E}\prime}\mathbf{Y}^{\mathbf{E}}$$

$$\hat{\theta} = \frac{1}{2n_F}\left(2\mathbf{1}_F\mathbf{1}_1 + \mathbf{1}_F\mathbf{1}_2\right),$$

where n_{F} is the number of founders and r(**X**) denotes the rank of the matrix **X**, (iv) with estimates from (iii) go back to (ii) and repeat the calculation until the prechosen number of N steps are completed.

The results obtained in the last step are taken as the Monte-Carlo EM estimates.

Simulation study

The simulated data used in this study were created on the basis of a mixed inheritance model (major gene plus polygenes). A non-inbred population consisted of 259 recorded and 41 base individuals. Two fixed effects (with three and two levels, respectively) were considered. Three traits observed in each recorded individual were included into the analysis. More details concerning the population and initial values for estimated parameters are listed in the Table. All individuals are assumed to be heterozygotes. The Gibbs sampler was run 2 500 000 rounds and each 500-th sample was stored. Finally, 5000 cycles were chosen to outline a posterior density of estimated parameters.

Since the parameters' posterior distributions are approximately symmetric, modes are not shown in the Table. Generally, posterior means of estimated parameters are satisfactory. The main goal of this investigation, the detection of three trait single gene effects has been achieved. However, some estimates of fixed effects are different from their true values. Moreover, the residual variances of the first two traits are slightly overestimated and the residual covariances between studied traits are slightly underestimated.

Implications

Although, as already mentioned, the simulation study gave optimistic results, it should be noted that the effects of a single gene (and their variances) were relatively large, whereas the polygenic and residual effects were rather moderate in this example. For simplicity, null polygenic additive covariances between these traits were also assumed. Thus, a detection of major gene effects was evident. It seems that the proposed method requires further studies, including more complex genetic models and different aspects of data modelling. An investigation (for unitrait model analysed via Gibbs sampling) conducted by SZYDŁOWSKI and SZWACZKOWSKI (1998) showed for example that skewed trait distribution leads to false inference about the presence of a major gene.

From a practical point of view this algorithm offers a number of genetic parameters (as functions of estimated variance-covariance components) used in livestock improvement. Apart from classical parameters (heritability and genetic correlation), the (co)variance function estimates for a single locus (e.g. ratio of major gene variance to total variance as well as respective genetic correlations) are also possible to estimate.

Acknowledgements. This work was supported by The State Committee for Scientific Research, grant No. 5P06D 03612

REFERENCES

DOBEK A., MOLIŃSKI K., SZYDŁOWSKI M., SZWACZKOWSKI T. (1999). Mixed models with fixed nongenetic and random major gene-polygenic effects. Anim. Sci. Pap. Rep. 17(2): 5-14.

ELSTON R.C., STEWARD J. (1980). A general model for the genetic analysis of pedigree data. Hum. Hered. 21: 523-542.

GUO S.W., THOMPSON E.A. (1994). Monte Carlo estimation of mixed model 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. Theor. Appl. Genet. 91: 1137-1147.

LE ROY P., ELSEN J.M. (1992). Simple test statistics for major gene detection: a numerical comparison. Theor. Appl. Genet. 83: 635-644.

MERAT P. (1993). Pleiotropic and associated effects of major genes. In: Poultry Breeding and Genetics (R.D. Crawdford, ed.). Elsevier, Amsterdam. Netherlands.

MIYAKE T., GAILLARD C., MORIYA K., SASAKI Y. (1999). Accuracy of detection of major genes segregating in outbred population by Gibbs sampling using phenotypic values of quantitative traits. J. Anim. Breed. Genet. 116: 281-288.

MORTON N.E., MACLEAN C.J. (1974). Analysis of family resemblance. III. Complex segregation of quantitative traits. Amer. J. Hum. Genet. 26: 489-503.

SZYDŁOWSKI M., SZWACZKOWSKI T. (1998). Simulation study on the application of the Gibbs sampler for major gene detection in laying hen population. J. Appl. Genet. 39: 321-330.