Archiv Tierzucht 54 (2011) 1, 93-103, ISSN 0003-9438

Bayesian analysis of ordinal categorical data under a mixed inheritance model

Ewa Skotarczak[1], Krzysztof Molinski[1], Tomasz Szwaczkowski[2] and Anita Dobek[1]

[1]Department of Mathematical and Statistical Methods, [2]Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Poznan, Poland

Abstract

The effectiveness of proposed Gibbs sampling (GS) algorithm to detect single loci determining livestock threshold traits under a different hypothetical breeding and statistical modeling scenarios was examined. The following factors were included into the analysis: the presence of fixed effects, knowledge of one threshold, the size of the population (1 212 and 3 070 pedigreed individuals, respectively) and proportions of individuals in three genotypic classes. Five threshold and one linear unitrait animal model were employed to analysis of these datasets. The GS algorithm was applied to estimate fixed effects (optionally), additive polygenic variance, single allele frequencies, genotypic effects and one threshold (optionally). For each case, 2 000 000 rounds of GS were conducted. The first 1 000 000 steps were discarded as a burn-in-period. The results were collected from every 20th iteration. In general, the accuracy of parameter estimates is not satisfactory. However, taking into account the scant amount of information provided by the ordinal categorical data, it seems that such an analysis is a good first approach. Except for one case in which the estimate was very close to the true value, in all the other cases the estimated gene effect was smaller than the true effect. In general, the algorithm proposed does not provide overestimated effects of single locus.

Keywords: categorical data, major gene effect, Gibbs sampling, threshold model

Zusammenfassung

Bayessche Analyse ordinal skalierter kategorischer Daten im gemischten Vererbungsmodell

Es wurde die Effektivität eines Gibbs sampling (GS) Algorithmus zur Detektion von Genen mit Einfluss of Schwellenmerkmale untersucht. In der Analyse werden folgende Faktoren berücksichtigt: fixe Effekte, Schwellenwerte, Populationsgröße und Genotypen. Es wurden fünf verschiedene Schwellenmodelle und ein lineares Modell analysiert. Mit Hilfe des GS Algorithmus wurden die festen Effekte, die additive polygenetische Variation, die Frequenzen von Allelen des Hauptgens, die Effekte des Hauptgens und die Schwellen geschätzt. Die Analyse wurde anhand von 2 000 000 Durchgängen vorgenommen. Im Ergebnis wurde jede 20. Iteration berücksichtigt. Dabei wurde festgestellt, dass die Schätzungsgenauigkeit unbefriedigend ist. Wenn man aber die begrenzte Menge an Informationen bei ordinalen skalierten kategorischen Variablen betrachtet, scheint die vorgeschlagene Analyse ein guter

Ausgangspunkt für weitere Analysen zu sein. Das wesentliche Merkmal der vorgeschlagenen Methode ist, dass der geschätzte Effekt immer kleiner oder höchstens gleich dem wirklichen Effekt ist, d.h. der Algorithmus überschätzt den Effekt des Hauptgens nicht.

Schlűsselwörter: kategorische Daten, Effekte des Hauptgens, Gibbs sampling, Schwellenwertmodell

Introduction

In the past decades of the 20th century, applied genetic improvement programs in livestock have been targeted at increasing production traits. So far, a considerably phenotypic and genetic gain has been registered. Unfortunately, the production characters are negatively correlated with functional ones, for instance fertility, fecundity, some disease resistances etc. A majority of them is categorical and often binary recorded. In fact, they are difficult to analyze statistically, since two scales (unobserved continuous liability and observed discrete phenotypes) should be statistically modeled. For simplicity classical linear models have been sometimes employed for genetic evaluation in livestock (Szwaczkowski *et al.* 2002). Furthermore, low heritabilities estimated for functional traits have greatly limited its inclusion in selection programs. Hence, new statistical tools are still developed for analysis of categorical data (Gianola 1982, Molinski *et al.* 2003). Comparative studies (Casselas 2007, Silvestre 2007) indicate threshold methodology with Gibbs sampling algorithm to analyze these traits.

A number of authors suggested that utilization of marker-assisted selection (Meuwissen & Van Arendonk 1992, Liu & Mathur 2005) and marker-assisted introgression (Dominik *et al.* 2007), especially for traits with low heritability, can considerably increase selection efficiency and maximize genetic progress. Unfortunately, molecular detection of important loci is still relatively expensive and labor-consuming. Furthermore, the methods used for estimating effects of quantitative trait loci, based on molecular data, often overestimate and or misestimate these effects (Hocking 2005). Similar conclusions were drawn at the 13th Quantitative Trait Locus and Marker Assisted Selection Workshop in 20-21 April, 2009 in Wageningen (Mucha, personal communication) where additionally the usefulness of Bayesian methods was stated. Hence, it seems that advanced molecular works should be preceded by marker-free segregation analysis. A first approach to segregation analysis is described by Elston & Steward (1971). Next, these methods were extended among others by Janss *et al.* (1995) and Guo & Thompson (1994). However, the majority of these approaches are mainly focused on continuous traits. Recently, the Bayesian marker-free segregation analysis has been more widely employed to threshold traits (Kadarmideen & Janss, 2005; Skotarczak *et al.* 2008, Sørensen *et al.* 1995). From the perspective of further molecular and marker segregation analysis, these algorithms can supply a number of useful information on genotype effects, single gene variance, allele frequency and polygenic variance. Furthermore, for the genetic analyses of the threshold characters the fixation and/or estimation (optionally, in case of more than two phenotypic classes) of the thresholds is required. Additionally, sensitivity of statistical inferences about segregation of single genes is influenced by the size and structure of population. It can also be determined by real genotypic effects and their frequencies as

well as numerical properties of applied algorithms. This paper is a continuation of an earlier study by Molinski *et al*. (2003) and Skotarczak *et al*. (2007, 2008) concerning Bayesian detection of single loci under threshold animal model. No literature on statistical effectiveness of proposed tools is available.

The main objective of this paper is to check the effectiveness of the described algorithm in detecting single loci determining livestock threshold traits.

Material and methods

Data simulation

Simulation studies enabled us to examine the influences of several factors, such as: the presence of fixed effects, a knowledge of one threshold, the size of the dataset and the proportion of individuals in the three classes of the analyzed variable. More details are listed in Table 1.

Table 1
Analyzed models

Model	Description
1.1.	Three categories for data, fixed effects excluded, fixed threshold i.e. $t_1=0$
1.2.	Three categories for data, fixed effects excluded, both thresholds unknown
2.1.	Three categories for data, fixed effects included, fixed threshold i.e. $t_1=0$
2.2.	Three categories for data, fixed effects included, both thresholds unknown
3.	Two categories for data, fixed effects included, fixed threshold i.e. t=0
4.	Continuous variable

To verify the correctness of the described algorithm some simulation studies have been provided. To be as close to reality as possible, the calculations were done for two different real pedigrees. The first one (P1), contained 1 212 individuals with 307 founders and 905 observed items. The average number of observation per individual was 3. The second pedigree (P2), consisted of 3 070 individuals with 454 founders and 2 040 observed animals. The mean replication number was about 5.5. To simulate the data we took two sets of parameters σ_a^2, f_{A2} and μ_{A1A1}, namely S1={0.4, 0.2, 0.8} and S2={0.8, 0.2, 1.5}. In the case of both simulated data sets the polygenic heritability coefficient equals 0.3. The assumption concerning the allele frequency and the structure of the relationship matrices gave in P1 757 individuals with genotype A_1A_1, 338 of type A_1A_2 and 117 of type A_2A_2. For the P2 these numbers were respectively equal to 2 138, 771 and 161.

In each analyzed case, 2 000 000 rounds of Gibbs sampling were conducted. The first 1 000 000 steps were discarded as a burn-in period. The important results were collected from every 20th iteration. The means of the posterior distributions were calculated as the point estimators of the unknown parameters. The statistical significance of the single gene effect was verified by the 95 % Highest Posterior Density Regions - HPDR (Scott 1992). If the HPDR included the 0 value it was stated that the single gene effect has no statistical meaning as opposed to the case when HPDR did not include 0.

Model and estimation of parameters

In the following analyses we assume that the ordinal categorical data y are described by a model

$$\boldsymbol{u} = \boldsymbol{X\beta} + \boldsymbol{ZW\mu} + \boldsymbol{Za} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{u}$ is the s vector of unobserved variables (i.e. liability), $\boldsymbol{s}$ is the number of recorded individuals, $\boldsymbol{X}$ is the $s \times b$ design matrix relating fixed nongenetic effects to observations, $\boldsymbol{\beta}$ is the b vector of fixed nongenetic effects, $\boldsymbol{Z}$ is the $s \times q$ design matrix relating polygenic and single locus effects to observations, q is the number of all individuals and $\boldsymbol{W}$ is the $q \times 3$ unknown matrix containing information on the genotype of each individual; each row of $\boldsymbol{W}$ 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, $\boldsymbol{\mu}$ is the vector $[\mu_{A_1A_1}, 0, -\mu_{A_1A_1}]'$, where $\mu_{A_1A_1}$ is the effect of genotype A_1A_1, and $-\mu_{A_1A_1}$ is the effect of A_2A_2, $\boldsymbol{a}$ is the q vector of random additive polygenic effects and $\boldsymbol{e}$ is the s vector of random effects.

The relation between observed phenotypes y_i and the liability u_i is conditioned by the thresholds t_1 and t_2, namely

$$y_i = \begin{cases} 1 & \text{if} \quad u_i \le t_1 \\ 2 & \text{if} \quad t_1 < u_1 \le t_2 \\ 3 & \text{if} \quad u_i > t_2 \end{cases} \qquad i = 1, \ldots, s \tag{2}$$

For the realization of Gibbs sampling procedure it is necessary to define the prior distributions for all model parameters. Uniform improper distributions were assumed for vectors $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, the random vectors $\boldsymbol{a}$ and $\boldsymbol{e}$ were normally distributed: $\boldsymbol{a} \sim \boldsymbol{N(0, A\sigma_a^2)}$, where $\boldsymbol{A}$ is the $q \times q$ relationship matrix, $\boldsymbol{e} \sim \boldsymbol{N(0, I_s)}$ and an inverted chi-square distribution was assumed for the component σ_a^2. Moreover, the uniform distribution on the interval $[0, t_{max}]$ was supposed for the unknown thresholds and the interval $[0, 1]$ for allele frequency.

The conditional posterior distributions for vectors $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{a}$ in the presented models are normal with means equal to the solutions of the appropriate mixed model equations (Sørensen & Gianola 2002).

The following formulas were used to calculate the expected values and variances at every step of Gibbs sampling procedure:

for the vector of fixed effects $\boldsymbol{\beta}$:

$$\beta_i \sim N\left((\boldsymbol{x}_i'\boldsymbol{x}_i)^{-1}\boldsymbol{x}_i'(\boldsymbol{u} - \boldsymbol{X}_{-i}\boldsymbol{\beta}_{-i} - \boldsymbol{ZW\mu} - \boldsymbol{Za}); (\boldsymbol{x}_i'\boldsymbol{x}_i)^{-1}\right) \tag{3}$$

where $\boldsymbol{x}_i$ is the i-th column of $\boldsymbol{X}$, $\boldsymbol{X}_{-i}$ is matrix $\boldsymbol{X}$ without the i-th column, $\boldsymbol{\beta}_{-i}$ is vector $\boldsymbol{\beta}$ without the i-th element, i=1, …, b;

for the vector of additive genetic effects $\boldsymbol{a}$:

$$a_i \sim N\left((\boldsymbol{z}_i'\boldsymbol{z}_i + \frac{1}{\sigma_a^2}A_{i,i})^{-1}(\boldsymbol{z}_i'(\boldsymbol{u} - \boldsymbol{X\beta} - \boldsymbol{ZW\mu}) - \frac{1}{\sigma_a^2}\boldsymbol{A}_{i,-i}\boldsymbol{a}_{-i}); \ (\boldsymbol{z}_i'\boldsymbol{z}_i + \frac{1}{\sigma_a^2}A_{i,i})^{-1}\right) \tag{4}$$

where $\boldsymbol{z}_i$ is the i-th column of $\boldsymbol{Z}$, $A_{i,i}$ is the element of $\boldsymbol{A}^{-1}$ in the i-th row and i-th column, $\boldsymbol{A}_{i,-i}$ is the i-th row of matrix $\boldsymbol{A}^{-1}$ without the i-th element, $\boldsymbol{a}_{-i}$ is vector a without the i-th element, i=1, ..., q.

Further, the elements of vector $\boldsymbol{\mu}$ were generated according to the following formula:

$$\mu_{A_1A_1} \sim N\left(\begin{matrix}((\boldsymbol{zw})_1'(\boldsymbol{zw})_1-(\boldsymbol{zw})_3'(\boldsymbol{zw})_3)^{-1}((\boldsymbol{zw})_1'(\boldsymbol{zw})_1-(\boldsymbol{zw})_3'(\boldsymbol{zw})_3)(\boldsymbol{u}-\boldsymbol{X\beta}-\boldsymbol{Za});\\ ((\boldsymbol{zw})_1'(\boldsymbol{zw})_1-(\boldsymbol{zw})_3'(\boldsymbol{zw})_3)^{-1}\end{matrix}\right) \quad (5)$$

where $(\boldsymbol{zw})_i$ is the i-th column of matrix $\boldsymbol{ZW}$, i=1, 2, 3. In every step of Gibbs sampling procedure it was fixed that $\mu_{A_1A_1}>0$. As it was mentioned above we assumed that $\mu_{A_1A_2}=0$ and $\mu_{A_2A_2}=-\mu_{A_1A_1}$

The values of the liability were generated from a truncated normal distribution with the truncation point defined by the actual threshold value:

$$p(u_1|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{a},\boldsymbol{G},\sigma_a^2)=\frac{\Phi(\boldsymbol{x}_i^R\boldsymbol{\beta}+(\boldsymbol{zw})_i^R\boldsymbol{\mu}+\boldsymbol{z}_i^R\boldsymbol{a},\,1)}{\phi(t_j-\boldsymbol{x}_i^R\boldsymbol{\beta}-(\boldsymbol{zw})_i^R\boldsymbol{\mu}-\boldsymbol{z}_i^R\boldsymbol{a})-\phi(t_{j-1}-\boldsymbol{x}_i^R\boldsymbol{\beta}-(\boldsymbol{zw})_i^R\boldsymbol{\mu}-\boldsymbol{z}_i^R\boldsymbol{a})} \quad (6)$$

where $\boldsymbol{x}_i^R$ is the i-th row of matrix $\boldsymbol{X}$, $(\boldsymbol{zw})_i^R$ is the i-th row of matrix $\boldsymbol{ZW}$, $\boldsymbol{z}_i^R$ is the i-th row of matrix $\boldsymbol{Z}$, i=1, ..., s, $\boldsymbol{G}$ is the table of genotypes determining structure of $\boldsymbol{W}$.

Moreover, Φ (.) and ϕ (.) denote the density and the cumulative distribution function of the normal distribution, respectively.

Threshold t_1 was generated from the following uniform distribution:

$$p(t_1|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{a},\boldsymbol{u},\boldsymbol{G},\sigma_a^2)=\frac{1}{min(\boldsymbol{u}\,|\,y=2)-max(\boldsymbol{u}\,|\,y=1)} \quad (7)$$

where $min(\boldsymbol{u}\,|\,y=2)$ denotes the minimum value of the liabilities within observations in the second category; similarly $max(\boldsymbol{u}\,|\,y=1)$ is the maximum value of the liabilities for observations in the first category. Threshold t_2 was generated similarly.

The additive variance component was generated from the following inverted chi-square distributions:

$$p(\sigma_a^2|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{a},\boldsymbol{u},\boldsymbol{G})\propto(\sigma_a^2)^{-\left(\frac{q+v_a}{2}+1\right)}exp\left(-\frac{\boldsymbol{a}'\boldsymbol{A}^{-1}\boldsymbol{a}+v_aS_a^2}{2\sigma_a^2}\right) \quad (8)$$

where v_a, S_a^2 are the hyperparameters (v_a denotes the degrees of freedom and S_a^2 is a scale parameter).

According to Guo and Thompson (1994), the elements of the unknown genotypes in table $\boldsymbol{G}$ were generated from the following formula:

$$p(G_i|\boldsymbol{\beta},\boldsymbol{\mu},\boldsymbol{a},\boldsymbol{u},\sigma_a^2,\boldsymbol{G}_{-i})\propto\left[\prod_{o_i}P(G_{o_i}|G_i,G_{m_i})\right]P(G_i|G_{s_i},G_{d_i})\,exp\left(-\frac{(\boldsymbol{z}_i'(\boldsymbol{u}-\boldsymbol{X\beta}-\boldsymbol{ZW\mu}-\boldsymbol{Za})^2}{2}\right) \quad (9)$$

where G_i is the genotype of the i-th individual, $\boldsymbol{G}_{-i}$ is the table of the genotypes of all individuals excluding the i-th individual, G_{o_i} refers to the genotype of the progeny of the i-th individual, G_{m_i} is the genotype of the i-th individual's mate, G_{s_i}, G_{d_i} are the genotypes of the i-th individual's parents, i=1, ..., q. When the individual is not observed, the last term will be substituted by 1. In the first step of Gibbs sampling, it was assumed that matrix $\boldsymbol{W}$

has the following form: $\boldsymbol{W}$=[**0 : 1 : 0**]. Further, for the single gene Mendelian transmission probabilities were assumed.

To estimate the genotypes for the individuals, the frequency of alleles among the groups of founders is required. The allele frequency was generated from the beta distribution according to the following formula (Kadarmideen & Janss 2005):

$$f\left(f_{A_1} \mid \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{a}, \boldsymbol{u}, \sigma^2_{a,} \boldsymbol{G}\right) \propto f_{A1}^{n_{A_1}} (1 - f_{A_1})^{n_{A_2}} \tag{10}$$

where n_{A_1} and n_{A_2} denote the number of alleles A_1 and A_2 in the group of founders.

Results

The results of simulation analysis are listed in Table 2 and visualized by Figures 1-4. Let us recall that by the simulation studies we wanted to check the influence of several factors on the precision of the obtained estimators. We have taken into account the following elements: the presence of fixed effects, knowledge of one threshold, the size of the dataset and the proportion of individuals in the three classes of the analyzed variable.

Table 2
Results of simulation studies – true and estimated parameters

Model	Pedigree and data proportion	f_{A_2}	$\hat{f}_{A_2}$	t_1	$\hat{t}_1$	t_2	$\hat{t}_2$
1.1	P1: 960, 834, 948	0.2	0.194	0.0	-	1.0	0.999
1.2	P1: 960, 834, 948	0.2	0.668	0.0	−0.550	1.0	0.452
1.1	P2: 4 687, 3 439, 3 205	0.2	0.230	0.0	-	1.0	1.010
1.2	P2: 4 687, 3 439, 3 205	0.2	0.605	0.0	−0.567	1.0	0.446
2.1	P1: 2 339, 302, 101	0.2	0.776	2.5	-	3.5	1.007
2.1	P1: 1 990, 435, 317	0.2	0.456	2.5	-	3.5	0.993
2.1	P2: 10 646, 541, 144	0.2	0.498	2.5	-	3.5	0.915
2.1	P2: 9 668, 1 157, 506	0.2	0.261	2.5	-	3.5	1.053
2.2	P1: 2 339, 302, 101	0.2	0.416	2.5	2.912	3.5	3.918
2.2	P1: 1 990, 435, 317	0.2	0.400	2.5	0.826	3.5	1.817
2.2	P2: 10 646, 541, 144	0.2	0.598	2.5	1.135	3.5	2.048
2.2	P2: 9 668, 1 157, 506	0.2	0.352	2.5	−0.434	3.5	0.619
3	P1: 2 339, 302, 101	0.2	0.162	-	-	-	-
3	P1: 1 990, 435, 317	0.2	0.489	-	-	-	-
3	P2: 10 646, 541, 144	0.2	0.528	-	-	-	-
3	P2: 9 668, 1 157, 506	0.2	0.425	-	-	-	-
4	P1: 2 339, 302, 101	0.2	0.275	-	-	-	-
4	P1: 1 990, 435, 317	0.2	0.359	-	-	-	-
4	P2: 10 646, 541, 144	0.2	0.054	-	-	-	-
4	P2: 9 668, 1 157, 506	0.2	0.834	-	-	-	-

f_{A_2}: frequency of allele, A, t_1 and t_2: thresholds

The first step in the verification procedure was to detect the influence of the fixed effects. Two analyses (1.1 and 1.2) in which the fixed effects were omitted gave the estimates of the most interesting parameters (σ^2_a, $\mu_{A_1A_1}$) close to the real values proving the correctness of the method and the procedure (Figure 1-2). The only one wrongly estimated parameter was the major gene effect in the case of both unknown thresholds in P1. Looking at these results, the positive influence of the bigger dataset is perceptible.

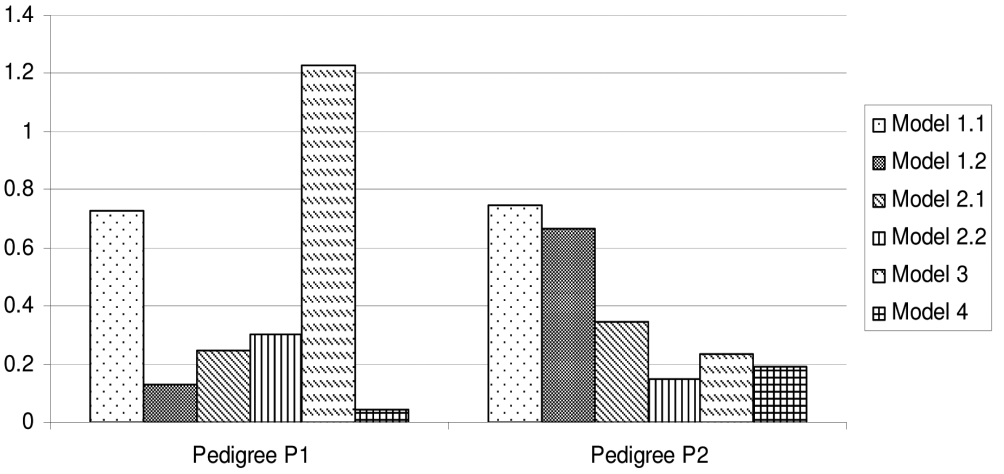

Figure 1
Comparison of major gene effect estimates when the true value is equal to 0.8

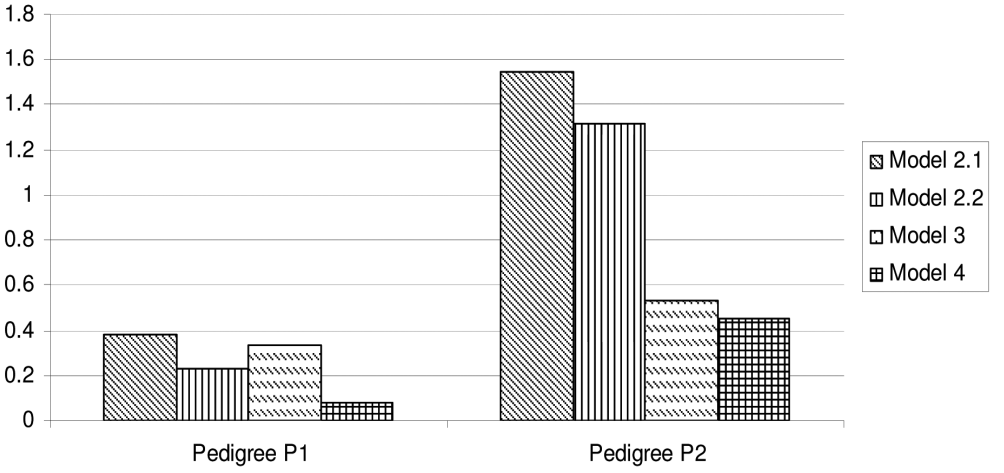

Figure 2
Comparison of major gene effect estimates when the true value is equal to 1.5

Because in the real data the fixed effects are usually present, in the subsequent analysis we have introduced three fixed effects. The estimates of $\boldsymbol{\beta}$ were usually wrong, but what is very interesting and important, in all the analyzed cases, the contrasts between them were correctly estimated.

For the other parameters, a positive influence of an assumption about the zero value of the first threshold, the second being unknown, is observable. Also a positive influence of the magnitude of pedigree and consequently the number of observations is true.

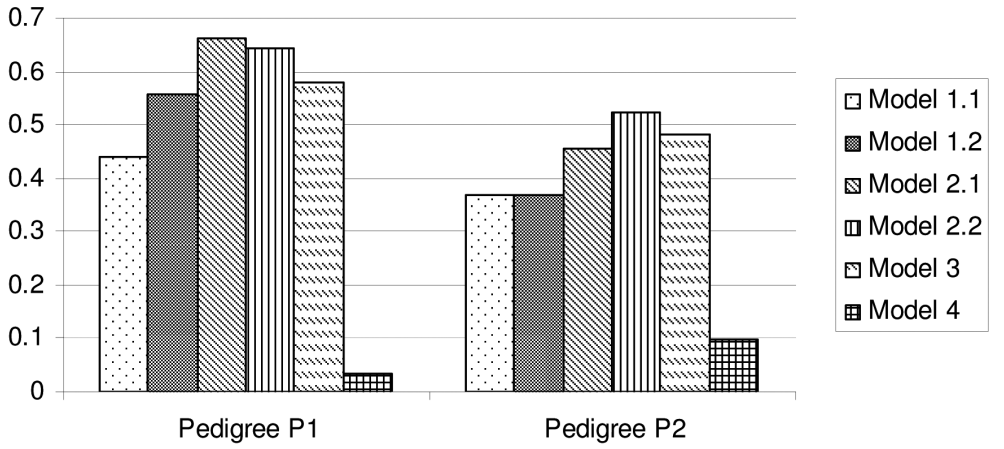

Figure 3
Comparison of genetic additive polygenic variance estimates when the true value is equal to 0.4

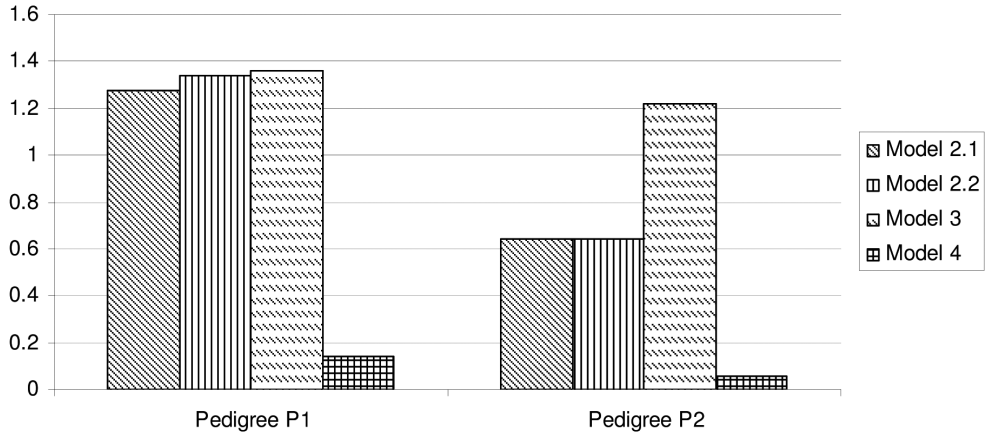

Figure 4
Comparison of genetic additive polygenic variance estimates when the true value is equal to 0.8

The analysis of estimates of f_{A2} indicates model 2.1 as a best solution and again the estimate for P2 being better than that for P1.

A subsequent problem is the division of observations into the three groups. A positive influence of a certain balance is seen only for the bigger set of data (P2), in which the major gene effect occurs to be significant.

When the number of observations in the three observed classes is very unbalanced, the question arises whether it is better to reduce the number of classes. Some comparisons of the results obtained for data grouped into two categories show that it is better not to join the classes and proceed with original data.

As usual in the case of discontinuous variables it may also be interesting to check whether it is not better to adopt one of the well known transformations leading to continuity and

analyze the data as observations of a continuous variable. When the proportion as the transformation is used the obtained results indicate that such procedure deteriorates the estimates.

Discussion

A detection of single loci with very large effects, determining continuous traits, with balanced size of genotypic classes segregating in a big population is very easy. Some major genes (e.g. Booroola gene - *FecB*) were identified many years ago by the use of relatively simple methods. They are effectively implemented into genetic improvement programs in livestock and poultry. König *et al*. (2009) concluded that economic efficiency and increased annual genetic gain in dairy cattle breeding programs are possible due to the replacement of classical procedure based on progeny testing by genome wide selection. It is almost sure that similar tendencies will be registered for other livestock populations.

Although a number of molecular and mathematical methods have been further developed, the detection of important single loci still seems difficult for at least three reasons. Firstly, when a trait of economic interest has a major gene segregating in a given population, the gene can easily be captured by selection. Secondly, the number of important traits (e.g. reproductive ability, disease resistance) has typical discrete vs binary phenotypes. Thirdly, some of the major genes are segregating in small unique populations.

As it has already been mentioned, many single loci have been detected for animal reproductive ability. These genes determining a given trait often have different chromosomal localizations as well as various effects. For instance, eleven different genes are known for fecundity in sheep (Davis 2005). Also single loci have been detected for other binary characters (e.g. disease resistance). However, some of them still have the status of putative quantitative trait loci (Casas & Stone 2005). Physical identification of a single locus with considerable effects requires more efforts and relatively high costs, because genotyping of many individuals is still expensive. Hence, the implementation of molecular information in breeding programs can be effectively preceded by results of the marker-free segregation analysis, mainly hypothetical genotypic effects, allele frequencies as well as quantitative genetic parameters. This corresponds with the results obtained in the present study, in which the proposed Bayesian algorithm does not lead to overestimation of parameters, mainly genotypic effects. Therefore, our results can be used as initiative data prior to classical marker segregation analysis.

Indeed, the proposed method assumed only a biallelic locus without dominance, although these assumptions are quite realistic (Kadarmideen & Janss 2005). Inheritance models for some traits are more complicated, as they include dominance, epistasis or genomic imprinting. Molecular identification of important single loci, especially determined discrete traits, seems to be more difficult (Blasco 2008).

Twenty different scenarios described by six models for two types of populations have been analyzed. To our knowledge, no reports on simulation studies on properties of the Bayesian algorithm to detect single loci by the use of non-marker data under a threshold animal model are available in literature. However, several papers concern the numerical comparison of non-Bayesian methods for both polygenic and mixed inheritance models and influence of data

structure on the accuracy of the estimated genetic parameters under a polygenic inheritance model (Le Roy & Elsen 1995, Kominakis 2008). It is well known that statistical inference based on larger samples leads to more accurate estimation. Such dependence has been confirmed in the present study. Estimates obtained for P2 data are usually more accurate compared to P1 data. Generally, the problem of data structure is connected with field collected data. It varies across populations, species etc. Our investigations are addressed for small livestock populations.

From the previous remarks it is obvious that the proposed analysis should be treated as a first step analysis. The accuracy of estimates is not satisfactory. However, taking into account the scant amount of information provided by the categorical data, it seems that such an analysis is a good first approach in the estimation procedure.

A very important result is connected with the estimate of the major gene effect. Except for one case in which the estimate was very close to the true value, in all the other cases the estimated gene effect was less than the true effect. Consequently, it seems to us, that an estimated gene effect, proved to be significant by, for example, the Highest Posterior Density Regions is indicative of the presence of a segregating gene which conditions the analyzed trait.

References

Blasco A (2008) The role of genetic engineering in livestock production. Livest Sci 113,191-201

Casas E, Stone RT (2006) Putative quantitative trait loci associated with the probability of contracting infectious bovine keratoconjunctivitis. J Anim Sci 84, 3180-3184

Casellas J, Caja G, Ferret A, Piedrafita J (2007) Analysis of litter size and days to lambing in the Ripollesa ewe. I. Comparison of models with linear and threshold approaches. J Anim Sci 85, 618-624

Davis GH (2005) Major genes affecting ovulation rate in sheep. Genet Sel Evol 37, 11-23

Dominik S, Henshall J, O'Grady J (2007) Factors influencing the efficiency of a marker-assisted introgression programme in Merino sheep. Genet Sel Evol 39, 495-511

Elston RC, Steward J (1971) Segregation analysis. Curr Dev Anthropol Genet 1, 327-354

Gianola D (1982) Theory and analysis of threshold characters. J Anim Sci 54, 1079-1096

Guo SW, Thompson EA (1994) Monte Carlo estimation of mixed model for large complex pedigrees. Biometrics 50, 417-432

Hocking PM (2005) Review of QTL mapping results in chickens. World Poultry Sci J 61, 215-226

Janss LLG, Thompson R, Van Arendonk JAM (1995) Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theor Appl Genet 91, 1137-1147

Kadarmideen HN, Janss LLG (2005) Evidence of a major gene from Bayesian Segregation Analyses of Liability to osteochondral diseases in pigs. Genetics 171, 1195-1206

Kominakis AP (2008) Effect of unfavourable population structure on estimates of heritability, systematic effects and breeding values. Arch Tierz 51, 601-610

König S, Simianer H, Willam A (2009) Economic evaluation of genomic breeding programs. J Dairy Sci 92, 382-391

Le Roy P, Elsen JM (1995) Numerical comparison between powers of maximum likelihood and analysis of variance methods for QTL detection in progeny test designs: the case of monogenic of monogenic inheritance. Theor Appl Genet 90, 65-72

Liu Y, Mathur PK (2005) Simplifications of marker-assisted genetic evaluation and accounting for non-additive interaction effects. Arch Tierz 48, 460-474

Meuwissen THE, Van Arendonk JAM (1992) Potential improvement in rate of genetic gain from marker assisted selection in dairy cattle breeding schemes. J Dairy Sci 75, 1651-1659

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

Scott DW (1992) Multivariate density estimation. Theory, Practice and Visualisation. John Wiley & Sons, New York

Silvestre AM, Ginja MMD, Ferreira AJA, Colaco J (2007) Comparison of estimates of hip dysplasia genetic parameters in Estrela Mountain Dog using linear and threshold models. J Anim Sci 85, 1880-1884

Skotarczak E, Moliński K, Szwaczkowski T, Dobek A (2007) Bayesian analysis of genetic backgrounds of twinning rate of Thorougbred horses. J Anim Feed Sci 16, 527-538

Skotarczak E, Szwaczkowski T, Moliński K, Dobek A (2008) Mixed Model Studies on Inheritance of Reproductive Traits in Laying Hens – a Bayesian Approach. Poultry Sci 87, 878-884

Sørensen DA, Andersen S, Gianola D, Kørsgaard I (1995) Bayesian inference in threshold models using Gibbs sampling. Genet Sel Evol 27, 229-249

Sørensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics, Springer-Verlag, New York

Szwaczkowski T, Moliński K, Szydłowski M, Skotarczak E, Piotrowski P, Dobek A (2002) Comparison of heritability estimates of hatchability layer traits obtained with linear and threshold model. 7th World's Congress on Genetics Applied to Livestock Production, August 19-23, 2002, Montpellier, France, Book of Abstracts, 86

Received 8 October 2009, accepted 4 October 2010.

Corresponding author:

Tomasz Szwaczkowski
email: tomasz@jay.au.poznan.pl

Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Wolynska 33, 60-637 Poznan, Poland