J. Appl. Genet. 39(4), 1998, pp. 321-330

Simulation study on the application of Gibbs sampling for major gene detection in a population of laying hens

Maciej SZYDŁOWSKI, Tomasz SZWACZKOWSKI

Department of Genetics and Animal Breeding, August Cieszkowski Agricultural University, Poznań, Poland

Abstract. A method for the detection of segregating major genes based on the analysis of estimated marginal posterior major gene variance density was examined. The properties of the method were investigated using data sets simulated for a real population of laying hens consisting of eleven generations. Marginal posterior densities of model parameters were estimated by the Gibbs sampling approach proposed by JANSS et al. (1995). With the data of about 4000 observations it was possible to detect a major gene responsible for one third of the genetic variance and one tenth of the phenotypic variance, irrespectively of the degree of dominance at the major locus. The inference based on the posterior marginal major gene variance can be sensitive to skewness of the data. It was shown that skewness of 0.2 can lead to a false detection of a major gene. The method is robust against a non-genetic mixture of normal distributions.

Key words: animal model, Gibbs sampling, laying hens, major gene, segregation analysis.

Introduction

Complex segregation analysis (ELSTON, STEWART 1971) is considered to be the most powerful method for the detection of a major gene. The inference is based on the comparison of the likelihood of the data under different inheritance models. Under mixed major gene–polygenic model, a trait distribution is described as a mixture of normals with the weights specified according to the mode of inheritance. The use of this method is limited by computational difficulties in likelihood calculations. Complex genetic models and large pedi-

Received: June 1998.

Correspondence: T. SZWACZKOWSKI, Department of Genetics and Animal Breeding, August Cieszkowski Agricultural University, ul. Wołyńska 33, 60-637 Poznań, Poland.

grees, which are usually available in animal populations, require some approximation.

One of the possible numerical tools is Gibbs sampling, which has been adapted to segregation analysis for human pedigrees (GUO, THOMPSON 1994). JANSS et al. (1995) presented a sampling scheme suitable for large animal pedigrees and based his inferences on posterior density estimates. The existence of a major gene is assumed from the shape of the estimated marginal posterior density of the major gene variance.

Segregation analysis relies strongly on the adequacy of the model assumed, and in particular, on the normality of the underlying distributions. Further, it has been shown that skewed distributed data can lead to false major gene detection (MACLEAN et al. 1975). On the other hand, transformation of the data considerably reduces the power of the method and raises the issue of interpretation (DEMENAIS et al. 1986).

Egg production traits show markedly nonnormal distributions (IBE, HILL 1988, BESBES et al. 1993). Hence, the genetic parameter estimates may be biased and the detection of major genes for these traits may be difficult.

The objective of this study was to examine the properties of the inference from mixed inheritance models based on marginal model parameter densities estimated by Gibbs sampling. The properties of the method for the detection of major gene of a different mode of inheritance were studied. Robustness of the method under a skewed distribution of polygenic trait and in the presence of a non-genetic mixture of normal distributions was also investigated.

Material and methods

The following statistical model was used for segregation analysis:

$$\mathbf{y} = \mathbf{I}\mu + \mathbf{Zu} + \mathbf{ZWm} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of observations, μ is an overall mean, $\mathbf{u}$ and $\mathbf{Wm}$ are vectors of polygenic and single gene effects, respectively; $\mathbf{e}$ is a vector of random errors and $\mathbf{Z}$ is an incidence matrix relating the genetic effects to observations. Vector $\mathbf{m}$ contains unknown additive (a) and dominance (d) genotypic values related to three genotypes, $\mathbf{m}' = \{a, d, -a\}$. $\mathbf{W}$ is unknown three-columns matrix which represents genotype configuration. Three columns correspond to three possible genotypes at autosomal biallelic major locus and each row contains two 0 and 1 to indicate the genotype of an individual. The prior distribution for was uniform, defined on $<-\infty, \infty>$. Distributional assumptions

for **u** were specified as $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where **A** is the numerator relationship matrix and σ_u^2 is a polygenic variance with a uniform prior distribution on $<0, \infty>$. Genotype probabilities for each founder were equal to Hardy-Weinberg frequencies. For each nonfounder, genotype probabilities were conditioned on the parental genotypes, assuming Mendelian segregation of alleles. Prior distributions for the frequency of positive allele (p), additive and dominance genotypic values were uniform and defined on [0, 1], $[0, \infty>$ and $<-\infty, \infty>$, respectively. Distributional assumptions for **e** were specified as $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where σ_e^2 is a residual variance component with uniform prior distribution on $<0, \infty>$. The complete set of parameters for the specified model was $\theta = (\mu, \mathbf{u}, \sigma_u^2, \mathbf{W}, p, a, d, \sigma_u^2)$. The variance explained by a single gene is defined as $\sigma_m^2 = 2p(1-p)(a + d(1-2p))^2 + (2p(1-p)d)^2$.

Segregation analysis was based on estimated marginal densities of parameters p, a, d, σ_u^2, σ_e^2 and σ_m^2. The densities were estimated from one thousand virtual independent joint samples of the parameters, generated via Gibbs sampling. The method iteratively generates pseudo random values of all model parameters according to their full conditional posterior distributions. The values successively sampled from the full conditional distributions (Gibbs chain) converge to drawings from the marginal distributions. The virtual independent samples were obtained by collecting values from every 5000 iterations. To facilitate the convergence of the Gibbs sampler, the genotype of each sire was sampled in block with the genotypes of its final progeny. Similar blocked sampling was used for poligenic values. Exact details on the sampling scheme can be found in JANSS et al. (1995). Additional improvement in convergence was achieved using relaxation of allele transmission probabilities to slightly non-Mendelian transmission (SHEEHAN, THOMAS 1993). To provide a correct set of samples for the inference on a strict Mendelian model, the probability for non-Mendelian transmission was gradually reduced from 0.05 to 0 before taking each next sample. The samples were generated in five Gibbs chains, obtaining 200 independent Gibbs samples per chain. Convergence of the Gibbs sampler was judged by testing for a significant chain effect with respect to parameters: p, a, d, σ_m^2, σ_u^2 and σ_e^2, using the standard ANOVA F-test. Densities of parameters were estimated using average shifted histogram (SCOTT 1992), and summary statistics of the marginal posterior distributions were computed by numerical integration using the estimated densities. For major gene variance,

the ratios of marginal density at global mode for $\sigma^2_m > 0$ and at $\sigma^2_m = 0$ (odds ratios) were computed. The 95% highest posterior density regions were also estimated for this parameter. Analyses were carried out using Gibbon program (SZYDŁOWSKI 1998).

All analyses were performed for simulated data sets. Trait records were formed for the real pedigree of a Leghorn strain. Hypothetical traits were observed only in females. The population comprised eleven generations and consisted of 338 base individuals, 1205 progeny which were parents themselves, and 3079 final progeny. Using this pedigree structure, different phenotypic data sets were simulated. Traits were generated according to a mixed and polygenic model of inheritance. In each case, the ratio of polygenic variance (σ^2_u) plus major gene variance (σ^2_m) to the total variance was *0.3*.

The effectiveness of the method for the detection of major genes of different inheritance modes was investigated using three data sets generated under a mixed major gene–polygenic model. The major locus was biallelic with the frequency of the positive allele (p) equal *0.3*. The data sets differed in additive (a) and dominance (d) genotypic values used in the simulations. The three data sets were generated under no dominance ($d = 0$), complete dominance ($d = a$) and overdominace ($d = 1.5a$). Expected value for σ^2_m was *1.0*. In each case, a major gene was responsible for a third part of the total genetic variance and a tenth part of the phenotypic variance.

Four different data sets were generated to examine the effect of the departure from normality on major gene detection. First, a single normally distributed data set was created assuming a pure polygenic model. For this data set, the coefficient of skewness (s) was close to zero. Based on these values, skewed distributed records were generated by transforming each value to a new value using the inverse of the Box-Cox transformation (see, e.g., UIMARI et al. 1996). The transformation parameters were chosen empirically to obtain the desired coefficients of skewness. Three differently skewed data sets were created with the coefficients of skewness close to 0.1, 0.2 and 0.3, respectively.

Three data sets were generated to investigate the reaction of the method to the presence of a non-genetic mixture of normal distributions. The mixture of three normals was created by appending a random effect of three levels with probabilities 0.25, 0.5 and 0.25, to the previously generated values under a pure polygenic model of inheritance. For the three data sets, the additional non-genetic effect was responsible for k = 1/7, k = 3/7 and k = 5/7 of the non-genetic variance, and 1/10, 3/10 and 5/10 of the total variance, respectively. The effect

corresponded to the codominant model with two alleles of equal frequencies, but was nonheritable.

Results

Tests for the convergence of the Gibbs sampler showed no significant ($P > 0.05$) differences between replicated chains for the considered parameters and demonstrated convergence of sampling procedure. The results of all performed analyses are given in Table 1. Means, modes and standard deviations of estimated densities are given relative to their true parameter values used in the simulations. For data sets generated under polygenic model, the estimates of p, a and d were irrelevant and are not shown. As already mentioned, the posterior marginal major gene density was the main criterion to assume a mixed model of inheritance. The statistics which summarise the estimated major gene densities are shown in Figure 1.

The skewness coefficients for the data sets generated under mixed inheritance models were 0.1, 0.12, and 0.16 for additive, complete dominance and overdominance models, respectively. The odds ratios increased slightly with the degree of dominance. For additive and complete dominance model, the 95% highest posterior density regions contained null value of the major gene variance, and only in the case of overdominace the region did not overlap zero. In each case, the mode and mean of marginal distributions of the major gene were close to true value ($\sigma^2_m = 1.0$). However, data generated under complete dominance resulted in the most accurate estimates.

For symmetrically distributed trait records generated under a pure polygenic model the estimated major gene density clearly indicated the absence of a major gene. The same was observed with skewness of 0.1. However, larger skewness led to odds ratios greater than one. It was shown that skewness equal to 0.2 and higher gave the density of major gene variance with the mode greater than zero.

The analyses of the data generated under a non-genetic mixture of distributions yielded the major gene variance densities with the modes at zero, irrespective of the magnitude of variance attributed to the non-genetic factor. The factor increased the estimates of residual variance, but not the major gene and polygenic variance. The increase was proportional to the part of total variability, for which the non-genetic factor was responsible. The results showed that the examined method discriminates a genetic mixture of distributions from that of non-genetic origin.

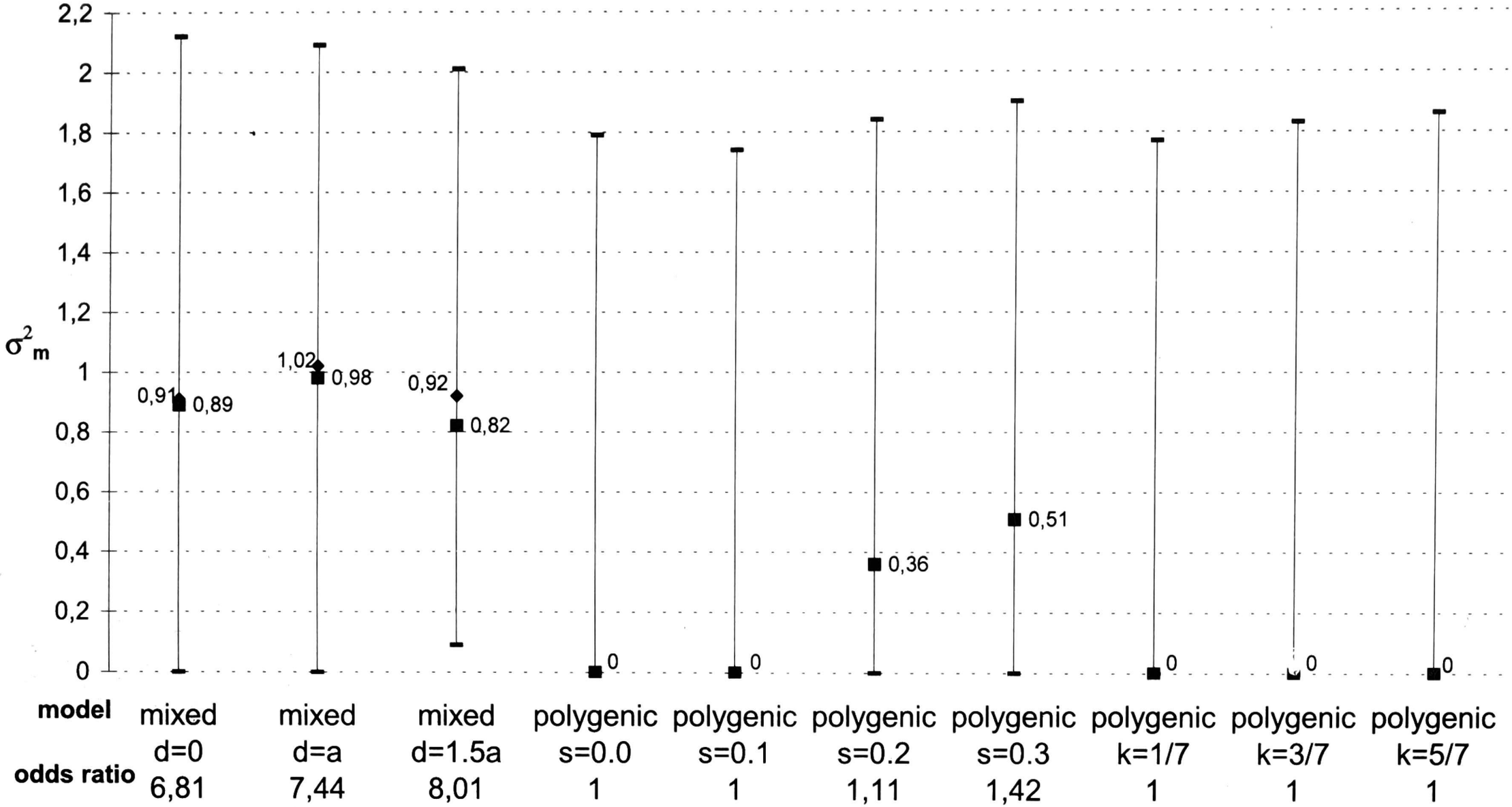

Figure 1. Estimates for posterior mean (◆), mode (■), the 95% highest posterior density region and odds ratio for major gene variance density obtained for data sets generated under mixed major gene-polygenic inheritance (model for additive (*d=0*), complete dominance (*d=a*) and overdominance (*d=1.5a*) gene action) and under pure polygenic model with different coefficients of skewness (s) in trait distribution and different parts (k) of error variance explained by unidentified non-genetic effect.

Table 1. Relative means, modes and standard deviations of posterior marginal model parameter densities obtained for data sets generated under mixed major gene–polygenic inheritance (model for additive ($d = 0$), complete dominance ($d = a$) and overdominance ($d = 1.5a$) gene action) and under pure polygenic model with different coefficients of skewness (s) in trait distribution and different parts (k) of error variance explained by unidentified non-genetic effect. All statistics are shown relative to true parameter values used in simulations.

Model	Statistics	Parameter				
		p	a	d	σ_u^2	σ_e^2
Mixed						
$d = 0$	mean	0.93	1.11		1.06	1.12
	mode	0.89	1.08		0.98	1.07
	s.d.	0.15	0.04		0.24	0.09
$d = a$	mean	0.94	1.04	0.91	1.02	1.13
	mode	0.91	1.06	0.95	0.99	1.06
	s.d.	0.14	0.04	0.06	0.21	0.10
$d = 1.5a$	mean	0.95	1.08	0.93	1.05	1.14
	mode	0.97	1.10	0.94	1.02	1.13
	s.d.	0.15	0.05	0.07	0.23	0.09
Polygenic						
s = 0.0	mean				0.85	1.09
	mode				0.95	1.02
	s.d.				0.24	0.12
s = 0.1	mean				0.94	1.10
	mode				0.87	1.03
	s.d.				0.15	0.11
s = 0.2	mean				0.88	1.09
	mode				0.93	1.01
	s.d.				0.23	0.14
s = 0.3	mean				0.81	1.08
	mode				0.97	1.06
	s.d.				0.22	0.14
k = 1/7	mean				0.91	1.12
	mode				0.97	1.08
	s.d.				0.20	0.13
k = 3/7	mean				0.94	1.35
	mode				0.91	1.34
	s.d.				0.22	0.14
k = 5/7	mean				0.97	1.61
	mode				1.20	1.57
	s.d.				0.24	0.17

Discussion

A variety of statistical approaches to major gene detection have been proposed (see, e.g., LE ROY, ELSEN 1992). Simple indicators of major gene segregations and computationally inexpensive methods based on mixture models allow data permutation to determine appropriate threshold values for test statistics (CHURCHILL, DOERGE 1994). However, these methods ignore many relationships among individuals and their application is limited to simple models. Gibbs sampling, in combination with the Bayesian approach, provides a very flexible tool. It makes the use of complex and looped pedigrees possible and gives the opportunity to fit many random genetic and non-genetic effects. These advantages are accompanied by large computational requirements and expensive permutation tests are not available.

To reduce the error of falsely accepting presence of a single gene, JANSS et al. (1995) proposed to consider σ^2_m to be significant only when the odds ratio exceeds 20. This value corresponds to a 5% significance level and prevents accepting a major gene unless abundant evidence is available. However, this criterion may be quite stringent. Alternatively, it was suggested to assume a mixed mode of inheritance as soon as the odds ratio exceeds one. As it has been shown in this paper on the sample size of about 4000 observations one can not expect odds ratio to exceed 20 even in the case of quite substantial magnitude of major gene variance. For such sample size, the effectiveness of the method is limited to major genes of very large effects. Hence, it seems better to perform a single analysis on a simulated trait for the population under consideration to get some idea about expected results of analysis for a trait affected by a major gene. The simulated single gene effect should be the smallest effect which is still of interest, e.g., required by marker free methods to estimate the genotype of an individual at major locus with a desired precision.

The method and type of inference applied in this paper appeared to be sensitive to skewness in the sense that skewed data lead to a positive mode of the major gene variance. For a classical segregation analysis based on likelihood ratio, it was demonstrated that skewness larger than 0.2 may lead to a false detection of major gene (DEMANAIS et al. 1986). As it was already mentioned, the skewed distribution is expected for egg production traits even when no major gene is segregating. Removing all skewness with the use of the Box-Cox transformation can lead to a considerable reduction in power (DEMENAIS et al. 1986). When dealing with a maximum likelihood approach, the transformation parameter can be estimated simultaneously with other parameters (MACLEAN et al. 1984). However, it was observed that it can lead to a tremendous change

in likelihood surface and in parameter estimates (UIMARI et al. 1997). Some skewness can be explained by the so-called scale effect. In this case, the analysis with different variances within genotypes instead of the common variance is a possible option.

When analysing skewed data, one can expect odds ratio to be greater than one. This may be caused by a major gene, a noninherited effect or both. To determine whether a major gene really exists, it can be useful to perform a segregation analysis on permuted data. A permuted data set is created from original data by random shuffling trait observations (with categorical variables and covariables for non-genetic effects) over individuals. The degrees of skewness are equal in original and permuted data sets. However, when analysing permuted data, significant results stem totally from nongenetic factors. As was shown in this paper for traits with approximately equal skewness, data skewed by a major gene tend to provide a larger odds ratio than a data set skewed by nongenetic effects. So, if a major gene really segregates, the odds ratio estimated from permuted data is expected to be lower than the odds ratio estimated from the original data set. If no major gene is segregating, there should be no substantial difference between these two estimates. Some caution is needed in density estimation from permuted data. The permutation of the data makes both the major gene variance and the polygenic variance not estimable. As the marginal posterior distribution of σ^2_m takes into account uncertainty in all other parameters in the model, the shape of density of a major gene variance can also be influenced by an increased uncertainty of polygenic variance. To reduce this effect, one can assume no polygenic background and compare the shapes of major gene densities estimated from a monogenic model or use the original observations for sampling polygenic values.

The results suggest that the method considered here is quite robust against a non-genetic mixture of normal distributions. However, sensitivity for a non-genetic mixture of distributions can depend on the mixing weights used. Such dependence was found by UIMARI et al. (1996) for simpler methods when two mixing proportions differed substantially. Hence, the robustness is limited to the cases, in which non-genetic mixture is not the source of skewness.

In conclusion, the simulated studies discussed in this paper showed that with a data set of about four thousand observations it is possible to detect a major gene responsible for one third of the genetic variance and one tenth of the total variance. However, skewness of the data can lead to a false major gene detection. To reduce the possibility of falsely accepting a mixed mode of inheritance, additional analysis based on a permuted data set is required.

Acknowledgements. This work was supported by The State Committee for Scientific Research, grant No. 5 PO6D 030 12.

REFERENCES

BESBES B., DUCROCQ V., FOULLEY J.L., PROTAIS M., TAVERNIER A., TIXER-BOICHARD M., BEAUMONT C. (1993). Box-Cox transformation of egg-production traits of laying hens to improve genetic parameter estimation and breeding value evaluation. Livest. Prod. Sci. 33: 313-326.

CHURCHILL G.A., DOERGE R.W. (1994). Empirical threshold values for quantitative traits mapping. Genetics 138: 963-971.

DEMENAIS F., LATHROP M., LALOUEL J.M. (1986). Robustness and power of the unified model in the analysis of quantitative measurements. Am. J. Hum. Genet. 38: 228-234.

ELSTON R.C., STEWART J. (1971). 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 models for large complex pedigrees. Biometrics 50: 417-432.

IBE S.N., HILL W.G. (1988). Transformation of poultry egg production data to improve normality, homoscedasticity and linearity of genotypic regression. J. Anim. Breed. Genet. 105: 231-240.

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.

MACLEAN C.J., MORTON N.E., LEW R. (1975). Analysis of family resemblance. IV Operational characteristics of segregation analysis. Am. J. Hum. Genet. 27: 365-384.

MACLEAN C.J., MORTON N.E., YEE S. (1984). Combined analysis of genetic segregation and linkage under an oligogenic model. Comput. Biomed. Res. 17: 471-480.

SCOTT D.W. (1992). Multivariate Density Estimation. Wiley and Sons, New York.

SHEEHAN N., THOMAS A. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics 49: 163-176.

SZYDŁOWSKI M. (1998). Gibbon – C++ program for Monte Carlo estimation of mixed models. Proceedings of 6th WCGALP, vol 27: 479-480, Armidale, Australia.

UIMARI P., KENNEDY B.W., DEKKERS J.C.M. (1996). Power and sensitivity of some simple tests for detection of major gene in outbred populations. J. Anim. Breed. Genet. 113: 17-28.