



HAL
open science

Bayes factor between Student t and Gaussian mixed models within an animal breeding context

Joaquim Casellas, Noelia Ibáñez-Escriche, Luis Alberto García-Cortés, Luis Varona

► **To cite this version:**

Joaquim Casellas, Noelia Ibáñez-Escriche, Luis Alberto García-Cortés, Luis Varona. Bayes factor between Student t and Gaussian mixed models within an animal breeding context. *Genetics Selection Evolution*, 2008, 40 (4), pp.395-413. hal-00894620

HAL Id: hal-00894620

<https://hal.science/hal-00894620>

Submitted on 11 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bayes factor between Student t and Gaussian mixed models within an animal breeding context

Joaquim CASELLAS^{1*}, Noelia IBÁÑEZ-ESCRICHE¹,
Luis Alberto GARCÍA-CORTÉS², Luis VARONA¹

¹Genètica i Millora Animal, IRTA-Lleida, 25198 Lleida, Spain

²Departamento de Mejora Genética Animal, SGIT-INIA, Carretera de la Coruña, km. 7,
28040 Madrid, Spain

(Received 2 April 2007; accepted 19 December 2007)

Abstract – The implementation of Student t mixed models in animal breeding has been suggested as a useful statistical tool to effectively mute the impact of preferential treatment or other sources of outliers in field data. Nevertheless, these additional sources of variation are undeclared and we do not know whether a Student t mixed model is required or if a standard, and less parameterized, Gaussian mixed model would be sufficient to serve the intended purpose. Within this context, our aim was to develop the Bayes factor between two nested models that only differed in a bounded variable in order to easily compare a Student t and a Gaussian mixed model. It is important to highlight that the Student t density converges to a Gaussian process when degrees of freedom tend to infinity. The two models can then be viewed as nested models that differ in terms of degrees of freedom. The Bayes factor can be easily calculated from the output of a Markov chain Monte Carlo sampling of the complex model (Student t mixed model). The performance of this Bayes factor was tested under simulation and on a real dataset, using the deviation information criterion (DIC) as the standard reference criterion. The two statistical tools showed similar trends along the parameter space, although the Bayes factor appeared to be the more conservative. There was considerable evidence favoring the Student t mixed model for data sets simulated under Student t processes with limited degrees of freedom, and moderate advantages associated with using the Gaussian mixed model when working with datasets simulated with 50 or more degrees of freedom. For the analysis of real data (weight of Pietrain pigs at six months), both the Bayes factor and DIC slightly favored the Student t mixed model, with there being a reduced incidence of outlier individuals in this population.

Bayes factor / Gaussian distribution / mixed model / Student t distribution / preferential treatment

* Corresponding author: Joaquim.Casellas@irta.es

1. INTRODUCTION

Genetic evaluations in animal breeding are generally performed using the mixed effects models pioneered by Henderson [9]. Usually, these models assume Gaussian distributions for most random effects, including the residuals, and in absence of contradictory evidence, it is practical to assume normality on the basis of both mathematical convenience and biological plausibility. Nevertheless, departures from normality are common in animal breeding, *e.g.* when more valuable animals receive preferential treatment [14,15]. This preferential treatment could be defined as any management practice that increases or decreases production and is applied to one or several animals, but not to their contemporaries [14]. Amongst others, these practices may include separate housing, better (or worse) or more (or less) feed, or better (or worse) sanitary attentions. Obviously, which animals or productive records receive preferential treatment is not known with any degree of certainty in real populations and this information loss could imply substantial bias in genetic evaluations [14,15]. Other potential causes of outliers or abnormal phenotypic records could be measurement errors, sickness, short-term-changes in herd environment and mismanagement of data [11].

We generally lack *a priori* sufficient information relating to the presence or absence of preferential treatment in our livestock data sets. It has been recently demonstrated that the specification of heavy-tailed residual distributions (such as the Student *t* distribution) instead of the usual Gaussian process in best linear unbiased prediction (BLUP) models may effectively mute the impact of residual outliers, particularly in situations where the preferential treatment of some breed stock may be anticipated [16,21]. As a result, accurate statistical tests are required to compare the mathematical simplicity of the Gaussian mixed model with the improved goodness of fit (under preferential treatment or other unknown sources of outliers) of the Student *t* mixed model.

General statistical tools such as the deviance information criterion (DIC) [20] or other approaches to Bayes factors [6] have been used to make comparisons between Gaussian and Student *t* mixed models. However, they imply high computational demands because both the Gaussian and the Student *t* mixed model must be analysed to calculate the corresponding comparison parameter. Within this context, the Bayes factor developed by García-Cortés *et al.* [5] and Varona *et al.* [23] in the animal breeding context implies a substantial simplification because it compares two models that only differ in terms of a single bounded variable, and therefore only the analysis of the complex model is required. The Student *t* distribution converges with the Gaussian distribution when the number of degrees of freedom tends to infinity. This property can be exploited

to appropriately adapt Varona *et al.* [23] Bayes factor, generating a useful statistical tool for the analysis of field data, especially when used for genetic evaluation purposes. In this paper, we focused our efforts on describing the development of this Bayes factor to make comparisons between Gaussian and Student *t* processes, and we tested its performance on both simulated and real data sets, using DIC as the standard reference criterion.

2. MATERIALS AND METHODS

2.1. Statistical background for Student *t* mixed models

Take as a starting point a standard linear model [9] such as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{W}\mathbf{p} + \mathbf{Z}\mathbf{a} + \mathbf{e}, \quad (1)$$

where \mathbf{y} is the vector with n phenotypic data, \mathbf{X} , \mathbf{W} , \mathbf{Z} are the incidence matrices of systematic (\mathbf{b}), permanent environmental (\mathbf{p}) and additive genetic effects (\mathbf{a}), respectively, and \mathbf{e} is the vector of residuals. The probability density of phenotypic data can be modeled under a multivariate Student *t* distribution with ν degrees of freedom (with ν being equal to or greater than 2):

$$p(\mathbf{y}|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \nu) = \prod_{i=1}^n \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\frac{1}{2})\nu^{\frac{1}{2}}} (\sigma_e^2)^{-\frac{1}{2}} \\ \times \left[1 + \frac{(y_i - \mathbf{x}_i\mathbf{b} - \mathbf{w}_i\mathbf{p} - \mathbf{z}_i\mathbf{a})'(y_i - \mathbf{x}_i\mathbf{b} - \mathbf{w}_i\mathbf{p} - \mathbf{z}_i\mathbf{a})}{\nu\sigma_e^2} \right]^{-\frac{1}{2}(\nu+1)}, \quad (2)$$

where \mathbf{x}_i , \mathbf{w}_i and \mathbf{z}_i are the i th row of \mathbf{X} , \mathbf{W} and \mathbf{Z} , respectively, y_i is the i th scalar element of \mathbf{y} , σ_e^2 is the residual variance and $\Gamma(\cdot)$ is the standard gamma function with the argument as defined within parentheses. For small values of ν , the Student *t* distribution shows a Gaussian-like pattern with increased probability in tails, whereas this distribution converges to a Gaussian distribution when ν tends to infinity [16]. For mathematical convenience, we can define $\delta = 2/\nu$ ($0 \leq \delta < 1$) and then, the conditional density (2) reduces to a normal density when $\delta = 0$ (as is, ν tends to infinity).

Following Strandén and Gianola [21], the previous model can be extended to an alternative parameterization if the data vector is partitioned according to

m ‘clusters’ typified by a common factor (*e.g.* animal, maternal environment, herd-year-season at birth), with the previous linear model defined as:

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} \mathbf{b} + \begin{bmatrix} \mathbf{W}_1 \\ \vdots \\ \mathbf{W}_m \end{bmatrix} \mathbf{p} + \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_m \end{bmatrix} \mathbf{a} + \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_m \end{bmatrix}, \tag{3}$$

\mathbf{X}_j , \mathbf{W}_j and \mathbf{Z}_j being the appropriate incidence matrices of records in the j th cluster (\mathbf{y}_j), and \mathbf{e}_j being the corresponding vector of residuals. This reparameterization allows for an alternative description of the conditional density of \mathbf{y} [21]:

$$p(\mathbf{y}|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta) = \prod_{j=1}^m p(\mathbf{y}_j|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, s_j^2) p(s_j^2|\delta), \tag{4}$$

where $p(\mathbf{y}_j|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, s_j^2)$ is a multivariate normal distribution weighted by s_j^2 ,

$$p(\mathbf{y}_j|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, s_j^2) \sim N\left(\mathbf{X}_j\mathbf{b} + \mathbf{W}_j\mathbf{p} + \mathbf{Z}_j\mathbf{a}, \mathbf{I}_{n_j} \frac{\sigma_e^2}{s_j^2}\right), \tag{5}$$

\mathbf{I}_{n_j} being an identity matrix with dimensions $n_j \times n_j$, and the conditional distribution of the mixing parameter (s_j^2) is a Gamma density

$$p(s_j^2|\delta) = \frac{(\frac{1}{2\delta})^{\frac{1}{2\delta}}}{\Gamma(\frac{1}{2\delta})} (s_j^2)^{(\frac{1}{2\delta}-1)} \exp\left(-\frac{s_j^2}{2\delta}\right) \tag{6}$$

with it having an expectation of 1 when $\delta = 0$ [4,21].

2.2. Bayes factor between Student t and Gaussian linear models

The Bayes factor developed by Verdinelli and Wasserman [25], and applied to the animal breeding context by García-Cortés *et al.* [5] and Varona *et al.* [23], contrasts nested linear models that only differ in terms of a bounded variable. We adapted this methodology to compare a Student t mixed linear model with its simplification to the Gaussian mixed linear model when ν tends to infinity or, for mathematical convenience, $\delta = 2/\nu = 0$. Within this context, the posterior distribution of all the parameters of a Student t mixed model can be stated in two ways, with a pure Student t Bayesian likelihood (Model T_1):

$$\begin{aligned} p_{T_1}(\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_p^2, \sigma_a^2, \sigma_e^2, \delta|\mathbf{y}) &\propto p_{T_1}(\mathbf{y}|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta) p_T(\delta) p_T(\mathbf{b}) p_T(\mathbf{p}|\sigma_p^2) \\ &\times p_T(\sigma_p^2) p_T(\mathbf{a}|\mathbf{A}, \sigma_a^2) p_T(\sigma_a^2) p_T(\sigma_e^2), \end{aligned} \tag{7}$$

or with a Gaussian \times Gamma Bayesian likelihood (Model T_2):

$$\begin{aligned}
 p_{T_2}(\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_p^2, \sigma_a^2, \sigma_e^2, \delta, s_{j \in (1,m)}^2 | \mathbf{y}) &\propto \prod_{j=1}^m p_{T_2}(\mathbf{y}_j | \mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, s_j^2) p_{T_2}(s_j^2 | \delta) \\
 &\times p_T(\delta) p_T(\mathbf{b}) p_T(\mathbf{p} | \sigma_p^2) p_T(\sigma_p^2) \\
 &\times p_T(\mathbf{a} | \mathbf{A}, \sigma_a^2) p_T(\sigma_a^2) p_T(\sigma_e^2),
 \end{aligned}
 \tag{8}$$

where \mathbf{A} is the numerator relationship matrix between individuals. Following in part Varona *et al.* [23], the prior distribution assumed for the bounded variable (δ) was assumed

$$p_T(\delta) = \begin{cases} 1 & \text{if } \delta \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}
 \tag{9}$$

The permanent environmental and the additive genetic effects were assumed to be drawn from multivariate normal distributions,

$$p_T(\mathbf{p} | \sigma_p^2) \sim N(\mathbf{0}, \mathbf{I}_p \sigma_p^2),
 \tag{10}$$

$$p_T(\mathbf{a} | \mathbf{A}, \sigma_a^2) \sim N(\mathbf{0}, \mathbf{A} \sigma_a^2),
 \tag{11}$$

with \mathbf{I}_p being an identity matrix with dimensions equal to the number of elements of \mathbf{p} . The prior distributions for the remaining parameters of the model were defined as:

$$p_T(\mathbf{b}) = \begin{cases} k_1 & \text{if } b_l \in \left[-\frac{1}{2k_1}, \frac{1}{2k_1}\right], \\ 0 & \text{otherwise for each level } l \text{ of } \mathbf{b}, \end{cases}
 \tag{12}$$

$$p_T(\sigma_p^2) = \begin{cases} k_2 & \text{if } \sigma_p^2 \in \left[0, \frac{1}{k_2}\right], \\ 0 & \text{otherwise,} \end{cases}
 \tag{13}$$

$$p_T(\sigma_a^2) = \begin{cases} k_3 & \text{if } \sigma_a^2 \in \left[0, \frac{1}{k_3}\right], \\ 0 & \text{otherwise} \end{cases}, \tag{14}$$

$$p_T(\sigma_e^2) = \begin{cases} k_4 & \text{if } \sigma_e^2 \in \left[0, \frac{1}{k_4}\right], \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where k_1, k_2, k_3 and k_4 are four values that were small enough to ensure a flat distribution over the parameter space [23].

The joint posterior distribution of all the parameters in the alternative Gaussian mixed model (Model G) was proportional to

$$p_G(\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_p^2, \sigma_a^2, \sigma_e^2 | \mathbf{y}) \propto p_G(\mathbf{y} | \mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2) p_G(\mathbf{b}) p_G(\mathbf{p} | \sigma_p^2) \times p_G(\sigma_p^2) p_G(\mathbf{a} | \mathbf{A}, \sigma_a^2) p_G(\sigma_a^2) p_G(\sigma_e^2), \tag{16}$$

where the Bayesian likelihood was defined as multivariate normal,

$$p_G(\mathbf{y} | \mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2) \sim N(\mathbf{Xb} + \mathbf{Wp} + \mathbf{Za}, \mathbf{I}\sigma_e^2), \tag{17}$$

and the prior distributions $p_G(\mathbf{b}), p_G(\mathbf{p} | \sigma_p^2), p_G(\sigma_p^2), p_G(\mathbf{a} | \mathbf{A}, \sigma_a^2), p_G(\sigma_a^2)$ and $p_G(\sigma_e^2)$ were identical to the prior distributions of Model T_1 (or Model T_2).

The Bayes factor between Model T_1 (or Model T_2) and Model G ($\text{BF}_{T/G}$) can be easily calculated from the Markov chain Monte Carlo sampler output of the complex model (Student t mixed model). Under Model T_1 , the conditional posterior distribution of all the parameters in the model did not reduce to well-known distributions and generic sampling processes such as Metropolis-Hastings [8] are required. Simplicity was gained under the alternative Model T_2 during the sampling process. In this case, sampling from all the parameters in Model T_2 can be performed using a Gibbs sampler [7], with the exception of δ , which requires a Metropolis-Hastings step [8]. Following García-Cortés *et al.* [5] and Varona *et al.* [23], the posterior density $p_T(\delta = 0 | \mathbf{y})$ suffices to obtain $\text{BF}_{T/G}$,

$$\text{BF}_{T/G} = \frac{p_T(\delta = 0)}{p_T(\delta = 0 | \mathbf{y})} = \frac{1}{p_T(\delta = 0 | \mathbf{y})}, \tag{18}$$

because $p_T(\delta = 0) = 1$ (see equation (9)). Alternatively,

$$\text{BF}_{G/T} = \frac{p_T(\delta = 0|\mathbf{y})}{p_T(\delta = 0)} = p_T(\delta = 0|\mathbf{y}). \quad (19)$$

The $\text{BF}_{T/G}$ can be obtained by averaging the full conditional densities of each cycle at $\delta = 0$ using the Rao-Blackwell argument [26]. At this point, computational simplicity is gained with Model T_1 (or a normal density for $\delta = 0$), whereas Model T_2 tends to computationally unquantifiable extreme probabilities when δ is close to zero. A $\text{BF}_{T/G}$ greater than 1 indicates that the Student *t* mixed model is more suitable, whereas a $\text{BF}_{T/G}$ smaller than 1 indicates that the Gaussian mixed model is more suitable.

From the standard definition of the Bayes factor [13],

$$\text{PO}_{T/G} = \text{BF}_{T/G} \times \text{PrO}_{T/G} = \text{BF}_{T/G} \times \frac{p_T}{p_G}, \quad (20)$$

where $\text{PO}_{T/G}$ is the posterior odds between models, $\text{PrO}_{T/G}$ is the prior odds between models, and p_T and p_G are the *a priori* probabilities for Student *t* mixed model and Gaussian mixed model, respectively. In the standard development of the Bayes factor described above, we assumed that prior odds were 1 and p_T and p_G were both 0.5. Nevertheless, we could modify prior odds depending on our *a priori* knowledge, *e.g.* Student *t* mixed model is a more parameterized model and it could be easily penalized with a smaller-than-1 prior odds. Posterior odds can be viewed as the weighted value of the Bayes factor, conditional to our *a priori* degree of belief.

2.3. Simulation studies

The Bayes factor methodology developed above was validated through simulation. Seven different scenarios were analyzed following a Student *t* residual process, with degrees of freedom equal to 5 ($\delta = 0.4$), 10 ($\delta = 0.2$), 20 ($\delta = 0.1$), 50 ($\delta = 0.04$), 100 ($\delta = 0.02$), 200 ($\delta = 0.01$) and 300 ($\delta = 0.007$), respectively. Twenty-five replicates were simulated for each case and each replicate included five non-overlapping generations with 200 individuals (10 sires and 190 dams) and random mating. Following Model T_2 , each individual had a phenotypic record and was assigned its own independent cluster. Data were generated from a normal density $N(\boldsymbol{\mu}, \mathbf{I}\sigma_e^2)$ weighted by a cluster-characteristic value drawn from equation (6). Note that $\boldsymbol{\mu}$ included a unique systematic effect (10 levels randomly assigned with equal probability and sampled from a uniform distribution between 0 and 1) and a normally distributed additive genetic effect generated

under standard rules [1]. Residual and additive genetic variances were equal to 1 and 0.5, respectively.

This simulation process generated seven different scenarios with 25 data sets which were analyzed twice, through the previously described Bayes factor and through a standard Gaussian model (Model *G*). For each analysis, a single chain was launched that contained 100 000 rounds, after discarding the first 10 000 rounds as burn-in [19]. Comparisons between the two models were performed through three approaches: (a) Bayes factor between nested models, (b) DIC [20], and (c) correlation coefficient between simulated and predicted breeding values ($\rho_{a,\hat{a}}$). Note that DIC is based on the posterior distribution of the deviance statistic [20], which is -2 times the sampling distribution of the data as specified in formula (2) or as the conjugated distribution of (5) and (6), $p(\mathbf{y}|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta)$ and $\prod_{j=1}^m p(\mathbf{y}_j|\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, s_j^2) p(s_j^2|\delta)$, respectively. Computational simplicity is gained with (2), DIC being calculated as $\bar{D}(\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta) - p_D$ where $\bar{D}(\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta)$ is the posterior expectation of the deviance statistic, $p_D = \bar{D}(\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta) - D(\bar{\mathbf{b}}, \bar{\mathbf{p}}, \bar{\mathbf{a}}, \bar{\sigma}_e^2, \bar{\delta})$ is the effective number of parameters, $\bar{D}(\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta)$ is the mean of the deviance statistic and $\bar{\theta}$ is the mean value of θ ($\theta \in (\mathbf{b}, \mathbf{p}, \mathbf{a}, \sigma_e^2, \delta)$).

2.4. Analysis of weight at six months in Pietrain pigs

After editing, 2330 records of live weight at six months in Pietrain pigs were analyzed, with an average weight (\pm SE) of 102.9 (\pm 0.265) kg. These pigs were randomly chosen from 641 litters from successive generations grouped in 135 batches during the fattening period, and their records were collected between years 2003 and 2006 in a purebred Pietrain farm registered in the reference Spanish Databank (BDporc®, <http://www.bdporc.irta.es>). At the beginning of the fattening period (two months of age), batches were created with pigs from different litters in order to homogenize piglet weight, and these groups were maintained up to slaughter (six months of age). Pigs were reared under standard farm management during the suckling and fattening periods. Pedigree expanded up to five generations and comprised 2601 individuals, with 109 boars and 337 dams with known progeny.

The operational model included the additive genetic effect of each individual, the permanent environmental effect characterized by the batch during the fattening period, and three systematic sources of variation: sex (male or female), year \times season with 11 levels, and age at weighing (180.0 ± 0.3 days) treated as a covariate. Data were analyzed by applying the Bayes factor described above and assuming a different cluster for each pig with phenotypic data. To easily

compare this method with a standard Gaussian model, data were also analyzed under Model G . The empirical correlation between estimated breeding values (posterior mean) was calculated in the two models and, as for the simulated data sets, DIC was calculated for Model T and Model G . Each Gibbs sampler ran with a single chain of 450 000 rounds after discarding the first 50 000 iterations as burn-in [19].

3. RESULTS

3.1. Simulated datasets

Summarized results of the 25 replicates for each simulated Student t process (5, 10, 20, 50, 100, 200 and 300 degrees of freedom) are shown in Table I. Estimates for additive genetic variance showed coherent behavior with average estimates slightly greater than 0.5. Average residual variance estimated using the Student t mixed model clearly agreed with the simulated value. Nevertheless, residual variance was clearly over-estimated for simulations with few degrees of freedom in which a Gaussian mixed model was applied, showing higher standard errors in data sets with few degrees of freedom. Simulations with 5 degrees of freedom showed the highest average residual variance under the Gaussian mixed model (1.664 ± 0.038), whereas the average residual variance was reduced to 1.222 ± 0.025 for replicates with 10 degrees of freedom, and converged to one for datasets with 300 degrees of freedom (showing a standard error smaller than 0.020). Under the Student t mixed model, average estimates of degrees of freedom fitted with true values without any noticeable bias, although precision decreased with larger degrees of freedom (Tab. I). Substantial discrepancies were observed between the two models in terms of predicted breeding values in extreme heavy-tailed simulations. Although the correlation coefficients between predicted breeding values in the Student t and Gaussian mixed models increased quickly in line with the degrees of freedom, the empirical correlation in replicates with 5 degrees of freedom was very small (0.377 ± 0.030) and average correlations greater than 0.9 were observed in simulations with 100 or more degrees of freedom (Tab. I).

Empirical correlations between simulated and predicted breeding values increased with degrees of freedom in both the Student t and Gaussian mixed models, although the Student t mixed model reached higher correlations when simulated degrees of freedom were small. As seen in Table II, simulations under extremely heavy-tailed processes (5 degrees of freedom) showed average correlations of 0.420 and 0.377 for Student t and Gaussian mixed models, respectively, suggesting substantial bias for genetic evaluations performed with

Table I. Variance component ($\times 100$), degrees of freedom and breeding value correlation estimates (mean \pm SE).

| Simulation (ν) | Student t mixed model | | | Gaussian mixed model | | $\rho_{T,G}$ |
|----------------------|-------------------------|----------------------|-----------------|----------------------|----------------------|-------------------|
| | $\tilde{\sigma}_a^2$ | $\tilde{\sigma}_e^2$ | $\tilde{\nu}$ | $\tilde{\sigma}_a^2$ | $\tilde{\sigma}_e^2$ | |
| 5 | 50.7 \pm 2.1 | 106.8 \pm 2.5 | 5.0 \pm 0.3 | 55.2 \pm 3.5 | 166.4 \pm 3.8 | 0.377 \pm 0.030 |
| 10 | 55.5 \pm 2.0 | 98.8 \pm 2.1 | 10.4 \pm 0.4 | 56.7 \pm 2.3 | 122.2 \pm 2.5 | 0.438 \pm 0.028 |
| 20 | 52.1 \pm 2.1 | 98.0 \pm 1.4 | 22.3 \pm 0.9 | 52.3 \pm 1.8 | 108.8 \pm 1.5 | 0.632 \pm 0.025 |
| 50 | 51.3 \pm 2.1 | 99.8 \pm 1.7 | 53.9 \pm 1.0 | 50.7 \pm 2.5 | 104.7 \pm 2.0 | 0.862 \pm 0.019 |
| 100 | 51.7 \pm 2.5 | 101.5 \pm 1.7 | 102.5 \pm 1.2 | 50.8 \pm 2.4 | 103.9 \pm 1.7 | 0.961 \pm 0.010 |
| 200 | 50.5 \pm 1.9 | 101.9 \pm 1.8 | 200.2 \pm 1.3 | 51.9 \pm 1.7 | 102.2 \pm 1.8 | 0.997 \pm 0.001 |
| 300 | 51.8 \pm 2.0 | 101.5 \pm 1.8 | 304.1 \pm 1.5 | 52.3 \pm 2.0 | 101.6 \pm 1.9 | 0.999 \pm 0.001 |

$\rho_{T,G}$: Empirical correlation between predicted breeding values in Student t and Gaussian mixed models.

Table II. Comparison criteria (average estimates) between Student t and Gaussian mixed models for each simulation scenario.

| Simulation (ν) | Student t mixed model | | Gaussian mixed model | | DIC _{Diff.} | BF _{T/G} |
|----------------------|----------------------------|------------------|----------------------------|------------------|----------------------|----------------------|
| | $\tilde{\rho}_{a,\bar{a}}$ | DIC _T | $\tilde{\rho}_{a,\bar{a}}$ | DIC _G | | |
| 5 | 0.420 | 3116 | 0.377 | 3606 | -490 | 4.6×10^{92} |
| 10 | 0.443 | 2895 | 0.431 | 3112 | -217 | 1.2×10^{71} |
| 20 | 0.463 | 2854 | 0.460 | 2939 | -85 | 2.2×10^{34} |
| 50 | 0.470 | 2842 | 0.468 | 2887 | -45 | 1.8×10^{14} |
| 100 | 0.469 | 2841 | 0.469 | 2867 | -27 | 5.7×10^4 |
| 200 | 0.470 | 2840 | 0.469 | 2850 | -10 | 0.569 |
| 300 | 0.470 | 2840 | 0.470 | 2841 | -1 | 0.420 |

$\tilde{\rho}_{a,\bar{a}}$: Empirical correlation between simulated and predicted breeding values.

DIC_T: Deviance information criterion for the Student t mixed model.

DIC_G: Deviance information criterion for the Gaussian mixed model.

DIC_{Diff.} = DIC_T - DIC_G.

BF_{T/G}: Bayes factor of the Student t mixed model against the Gaussian mixed model.

standard Gaussian models when normality did not hold. Differences between the two models quickly decreased with increasing degrees of freedom and were almost negligible for $\nu = 20$ and higher values.

As was expected, both the Bayes factor and DIC clearly favored Student t mixed models in simulated scenarios with few degrees of freedom and showed similar behavior throughout the analyzed framework (Fig. 1A). Datasets simulated under a residual Student t process with 5 degrees of freedom reached an average Bayes factor favoring the Student t mixed model of 4.6×10^{92} , and the average difference between DIC was also huge (-490). Although both comparison criteria decreased when degrees of freedom increased, our results suggest that the Student t mixed model was preferable instead of the Gaussian model up to 20 degrees of freedom, and that even for simulations with 50 degrees of freedom, the superiority of the Student t model remained almost total (Tab. III). Substantial discrepancies between the Bayes factor and DIC appeared in the last two scenarios (200 and 300 degrees of freedom). While the average Bayes factor slightly favored the Gaussian model, the DIC continued to produce smaller estimates for the Student t model, although with only a minimal difference in the last scenario (Tab. II). This suggests that the Bayes factor was more conservative, favoring the less parameterized model. This hypothesis was confirmed in Table III where the Bayes factor supported the Gaussian mixed model in 0, 0, 0, 1, 11, 18 and 19 data sets (for simulations with 5, 10, 20,

Table III. Distribution of the Bayes factors ($\log_{10}(\text{BF}_{T/G})$) between Student t and Gaussian mixed models (and the number of Gaussian models favored by the DIC) for each simulation scenario.

| Simulation (v) | $(-10, -1]$ | $(-1, 0]$ | $(0, 1]$ | $(1, 10]$ | $(10, 50]$ | $(50, 100]$ | $(100, 150]$ | $(150, 200]$ | Overall |
|--------------------|-------------|-----------|----------|-----------|------------|-------------|--------------|--------------|---------|
| 5 | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 2 (0) | 14 (0) | 6 (0) | 3 (0) | 25 (0) |
| 10 | 0 (0) | 0 (0) | 0 (0) | 1 (0) | 19 (0) | 4 (0) | 1 (0) | 0 (0) | 25 (0) |
| 20 | 0 (0) | 0 (0) | 0 (0) | 9 (0) | 16 (0) | 0 (0) | 0 (0) | 0 (0) | 25 (0) |
| 50 | 0 (0) | 1 (0) | 1 (0) | 20 (0) | 3 (0) | 0 (0) | 0 (0) | 0 (0) | 25 (0) |
| 100 | 3 (1) | 8 (2) | 6 (0) | 8 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 25 (3) |
| 200 | 3 (2) | 15 (3) | 7 (1) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 25 (6) |
| 300 | 11 (5) | 8 (5) | 6 (1) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 25 (11) |

$(\alpha, \beta] = \alpha < \log_{10}(\text{BF}_{T/G}) \leq \beta$.

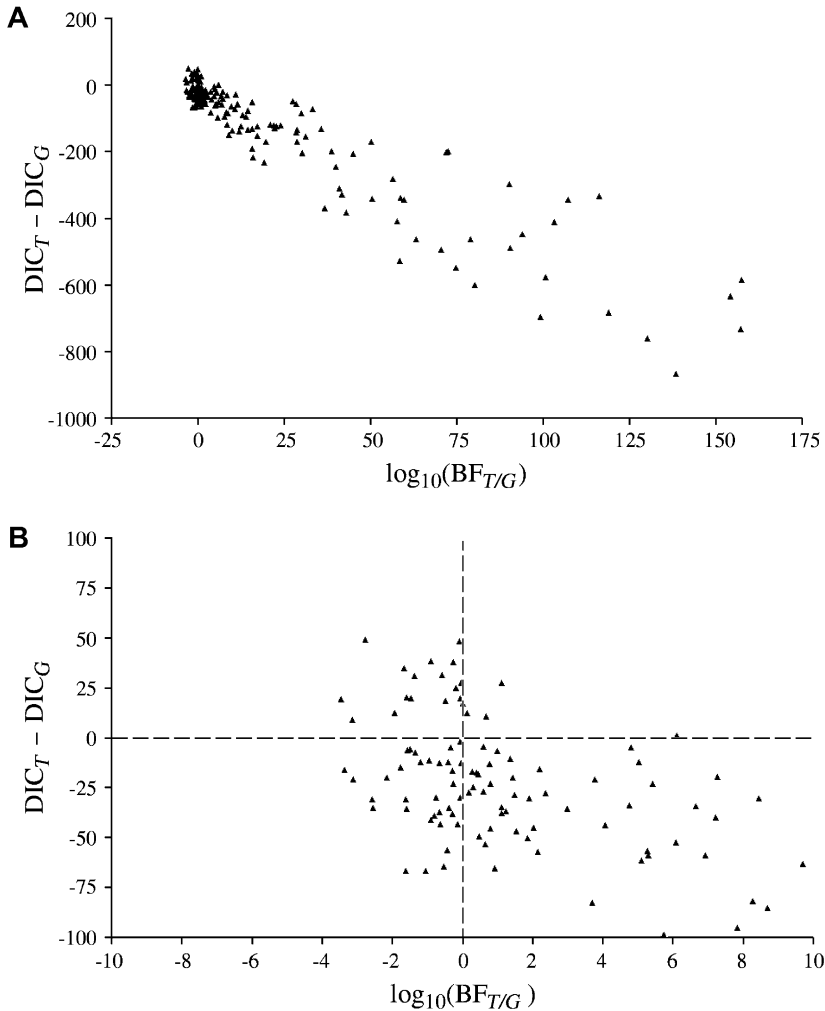


Figure 1. Plot of differences in DIC between Student t and Gaussian mixed models against $\log_{10}(\text{BF}_{T/G})$ for all replicates (A) and for combinations close to 0 (B).

50, 100, 200 and 300 degrees of freedom), whereas DIC favored the Gaussian mixed model in 0, 0, 0, 0, 3, 5 and 10 data sets, respectively (Fig. 1B).

3.2. Analysis of Pietrain pig weight at six months

Analysis under a Student t linear mixed model placed the highest posterior density region at 95% (HPD95) for the δ parameter between 0.004 and 0.009, with the modal value at 0.005 (Tab. IV). Degrees of freedom (ν) determine

Table IV. Summary statistics for the analysis of pig weight at six months under Student *t* and Gaussian mixed models.

| | Mean | Mode | HPD95 |
|------------------------|--------|--------|---------------|
| <i>Student t model</i> | | | |
| $\tilde{\sigma}_a^2$ | 36.87 | 35.92 | 24.81–50.06 |
| $\tilde{\sigma}_p^2$ | 35.16 | 35.86 | 22.33–48.83 |
| $\tilde{\sigma}_e^2$ | 68.72 | 68.63 | 59.97–77.32 |
| \tilde{c}^2 | 0.248 | 0.239 | 0.177–0.321 |
| \tilde{h}^2 | 0.260 | 0.254 | 0.178–0.341 |
| $\tilde{\delta}$ | 0.006 | 0.005 | 0.004–0.009 |
| $\tilde{\nu}$ | 177.95 | 191.12 | 111.20–235.70 |
| <i>Gaussian model</i> | | | |
| $\tilde{\sigma}_a^2$ | 36.21 | 35.84 | 24.67–49.21 |
| $\tilde{\sigma}_p^2$ | 35.88 | 35.81 | 22.32–48.42 |
| $\tilde{\sigma}_e^2$ | 69.54 | 69.01 | 60.91–77.82 |
| \tilde{c}^2 | 0.247 | 0.237 | 0.179–0.318 |
| \tilde{h}^2 | 0.255 | 0.256 | 0.173–0.338 |

HPD95 = highest posterior density region at 95%.

$$\tilde{c}^2 = \tilde{\sigma}_p^2 / (\tilde{\sigma}_a^2 + \tilde{\sigma}_p^2 + \tilde{\sigma}_e^2).$$

$$\tilde{h}^2 = \tilde{\sigma}_a^2 / (\tilde{\sigma}_a^2 + \tilde{\sigma}_p^2 + \tilde{\sigma}_e^2).$$

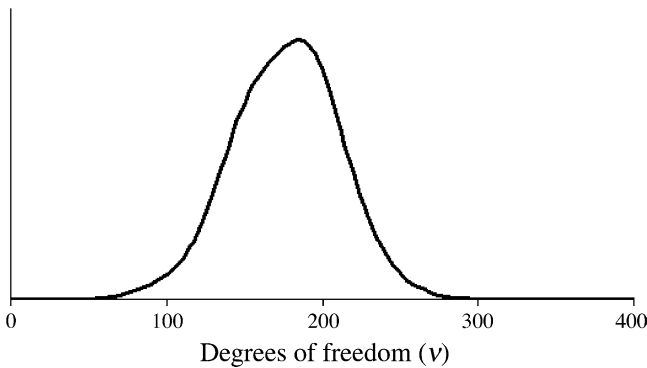


Figure 2. Posterior distribution of degrees of freedom for weight at six months in Pietrain pigs.

the slope of the Student *t* distribution and they were consequently drawn within a wide HPD95 (111.20–235.70) and the mode reached 191.12. The posterior distribution of ν was roughly symmetrical (Fig. 2). The posterior mean of the weights (s_i^2) for the Student *t* mixed model ranged from 0.924 to 1.026, although

Table V. Distribution of the posterior mean value of weights (s_i^2) for the analyses of pig weight at six months under Student *t* mixed models.

| | <i>n</i> | % |
|----------------------------|----------|--------|
| $0.900 < s_i^2 \leq 0.925$ | 5 | 0.22 |
| $0.925 < s_i^2 \leq 0.950$ | 15 | 0.64 |
| $0.950 < s_i^2 \leq 0.975$ | 88 | 3.78 |
| $0.975 < s_i^2 \leq 1.000$ | 799 | 34.29 |
| $1.000 < s_i^2 \leq 1.025$ | 1419 | 60.90 |
| $1.025 < s_i^2 \leq 1.050$ | 4 | 0.17 |
| Overall | 2330 | 100.00 |

the majority were located between 0.975 and 1.025 (95.19%; Tab. V). In this sense, values smaller than 0.925 were in a minority (0.22%), although they could have had a substantial influence as outliers. It is important to highlight the fact that differences in variance component estimation between the Gaussian and Student *t* mixed models were minimal (Tab. IV), with similar values for heritability (0.256 and 0.254, respectively) and for the coefficient of common environment (0.237 and 0.239, respectively).

The Bayes factor favored the Student *t* model rather than the Gaussian model ($BF_{T/G} = 2.532$), although the small size of this value is not worth more than a bare mention according to Jeffreys' [12] scale of evidence. He/she suggested that differences could be minimal, although DIC reported substantial discrepancies between the Student *t* mixed model (16 445) and the Gaussian mixed model (16 450); this difference was smaller than the average difference between DIC in simulations with 200 degrees of freedom. On the contrary, the empirical correlation coefficient between predicted breeding values in each model was 0.999 (this was 0.993 if only breeding animals were considered) and showed that, although the posterior probabilities of both models were slightly (Bayes factor) or substantially (DIC) different, genetic evaluations performed with a Student *t* or a Gaussian mixed model provided an almost identical genetic ranking for this data set.

4. DISCUSSION

The Bayes factor as originally proposed by Verdinelli and Wasserman [25] has been applied to various models used in animal breeding and the genetics research field. Although the method was initially developed to test for the genetic background of linear traits [5] and the location of quantitative trait loci (QTL) [23], this Bayes factor has been recently modified to discriminate

between linked and pleiotropic QTL [24], to test for the genetic background of threshold traits [2,18], and to compare different structures of random genetic groups [3], all with encouraging results. Previous research highlighted its computational stability [5]. Although two nested models were compared, only analysis of the complex model was required, which allowed a substantial reduction in computational requirements. Within this context, these properties highlight an interesting Bayesian tool that could be extended to compare a wide range of models.

Analysis of simulated data sets showed correct fitting of residual variances under a Student t mixed model. On the contrary, Gaussian mixed models suffered from substantial over-estimations of residual variance when phenotypic records were drawn from a Student t process that involved relatively few degrees of freedom. It confirmed previous research [11] and this phenomenon could be directly related to heavy-tailed characteristics of Student t density. Nevertheless, it is important to highlight the fact that in each simulation scenario, the average estimate of additive genetic variance across 25 replicates was placed close to the true value (Tab. 1) in both the Student t and Gaussian mixed models. A slight overestimation was suggested for average additive genetic variance, but this is commonly observed in Bayesian animal models in which modal estimates of σ_a^2 are unbiased, whereas the mean tends towards slight or moderate overestimation [22]. This stability in the estimation of genetic variance contrasted with the substantial discrepancies observed between the two models in terms of empirical correlation between predicted breeding values (between the two models), as well as empirical correlations between simulated and predicted breeding values (within a given model). For small degrees of freedom, the Student t mixed model reached higher correlations and verified the results previously reported by Strandén and Gianola [21] and Jamrozik *et al.* [11], who suggested that Gaussian models had only a limited ability to mute the effects of outliers in genetic evaluations. As expected, the Bayes factor confirmed the superiority of the Student t mixed models when data were generated from a Student t process with few degrees of freedom, and its performance was comparable to that provided by DIC (Fig. 1A). Our study shows that, although differences between Student t and Gaussian mixed models may be small when degrees of freedom tend towards large values, Student t mixed models could be preferable within a wide range of degrees of freedom. This confirmed previous results reported by Jamrozik *et al.* [11]. Indeed, the Bayes factor provided substantial evidence favoring the Student t model for six data sets in the replicate with 300 degrees of freedom. Slight discrepancies appeared between the Bayes factor and DIC in simulations with 50 or more degrees of freedom, where the former

could be viewed as a more conservative method. As suggested in Figure 1B and Table II, both the Bayes factor and DIC favored the Gaussian mixed model in 16 cases, whereas they produced contradictory results in a further 20 cases. This does not invalidate our methodology, although it suggests that for large degrees of freedom, the more parameterized model (Student t) would tend to be more penalized using the Bayes factor.

The analysis of Pietrain pig weight at six months provided an example of the Bayes factor applied to real data. Estimated variance components and their ratios were similar for both the Student t and the Gaussian mixed models. This was not surprising given the mode of degrees of freedom in the Student t mixed model (191.12; Tab. IV). Heritability for weight at six months was around 0.25, which was similar to the values reported by Hovernier *et al.* [10] and Noguera *et al.* [17]. There are no comparable results in the animal breeding literature for Student t mixed models with high values of degrees of freedom, whereas Jamrozik *et al.* [11] and Chang *et al.* [4] reported extremely heavy-tailed distributions (2.4 and 8.5 degrees of freedom, respectively) for mastitis in dairy cattle. This suggests a clearly higher incidence of unknown sources of outliers in dairy cattle than in fattening pigs, as well as a greater probability of the influence of preferential treatment. As seen in Table V, the weight parameters (s_j^2) of the Student t mixed model were placed around 1, showing less dispersion than in other research involving a smaller estimate of degrees of freedom [4,11]. Nevertheless, some outlier individuals have been observed (*e.g.* 20 individuals with s_j^2 lower than 0.95) and this result confirms the slight advantage of the Student t model over the Gaussian model suggested by the Bayes factor and DIC. Moreover, whereas the Bayes factor showed a small value that was close to 1, DIC showed substantial differences close to 5 units. This confirmed previous results obtained under simulation, in which the Bayes factor was suggested as being the more conservative. Nevertheless, given the extreme correlation obtained between the predicted breeding values in the two models and the computational simplicity provided by Model G , it would seem preferable to use a standard Gaussian mixed model for genetic evaluations. The high correlation between breeding values predicted by the two models confirmed the results obtained under simulation. Nevertheless, these results are conditional to our data set and they can not be taken as a general rule for genetic evaluations in swine.

In conclusion, a straightforward Bayes factor was developed to compare Gaussian and Student t mixed linear models, proposing two alternative parameterizations of the Student t model. This methodology could be of special interest to check preferential treatment or other sources of outlier data, which is a common phenomena in livestock data.

ACKNOWLEDGEMENTS

The authors are indebted to Dr. J.L. Noguera and COPAGA for field data on pig weight at six months, and to Dr. J. Piedrafita and Dr. G. Caja for providing additional field data sets during preliminary tests of the Bayes factor. The research contract of J. Casellas was partially financed by Spain's *Ministerio de Educación y Ciencia (Programa Juan de la Cierva)*.

REFERENCES

- [1] Bulmer M.G., *The Mathematical Theory of Quantitative Genetics*, Clarendon Press, Oxford, UK, 1980.
- [2] Casellas J., Piedrafita J., Bayes factor for testing the genetic background of quantitative threshold traits, *J. Anim. Breed. Genet.* 123 (2006) 301–306.
- [3] Casellas J., Piedrafita J., Varona L., Bayes factor for testing between different structures of random genetic groups: a case study using weaning weight in *Bruna dels Pirineus* beef cattle, *Genet. Sel. Evol.* 39 (2007) 39–53.
- [4] Chang Y.-M., Gianola D., Heringstad B., Klemetsdal G., A comparison between multivariate slash, Student's *t* and probit threshold models for analysis of clinical mastitis in first lactation cows, *J. Anim. Breed. Genet.* 123 (2006) 290–300.
- [5] García-Cortés L.A., Cabrillo C., Moreno C., Varona L., Hypothesis testing for the background of quantitative traits, *Genet. Sel. Evol.* 33 (2001) 3–16.
- [6] Gelfand A., Dey D.K., Bayesian model choice: asymptotics and exact calculations, *J. R. Stat. Soc. B Biol. Sci.* 56 (1994) 501–514.
- [7] Gelfand A., Smith A.F.M., Sampling based approaches to calculating marginal densities, *J. Am. Stat. Assoc.* 85 (1990) 398–409.
- [8] Hastings W.K., Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* 57 (1970) 97–109.
- [9] Henderson C.R., Sire evaluation and genetic trends, in: *Proc. Anim. Breed. Genet. Symp. in Honor of Dr. Jay L. Lush*, ASAS and ADSA, Champaign, IL, USA, 1973.
- [10] Hovemier R., Kanis E., van Asseldonk T., Westerink N.G., Genetic parameters of pig meat quality traits in a halothane negative population, *Livest. Prod. Sci.* 32 (1992) 309–321.
- [11] Jamrozik J., Strandén I., Schaeffer L.R., Random regression test-day models with residuals following a Student's-*t* distribution, *J. Dairy Sci.* 87 (2004) 699–705.
- [12] Jeffreys H., *Theory of Probability*, 3rd edn., Clarendon Press, Oxford, UK, 1961.
- [13] Kass R.E., Raftery A.E., Bayes factors, *J. Am. Stat. Assoc.* 90 (1995) 773–795.
- [14] Kuhn M.T., Boettcher P.J., Freeman A.E., Potential biases in predicted transmitting abilities of females from preferential treatment, *J. Dairy Sci.* 77 (1994) 2428–2437.

- [15] Kuhn M.T., Freeman A.E., Biases in predicted transmitting abilities of sires when daughters receive preferential treatment, *J. Dairy Sci.* 78 (1995) 2067–2072.
- [16] Lange K.L., Little R.J.A., Taylor J.M.G., Robust statistical modelling using a *t* distribution, *J. Am. Stat. Sci.* 84 (1989) 881–896.
- [17] Noguera J.L., Varona L., Babot D., Estany J., Multivariate analysis of litter size for multiple parities with production traits in pig: I. Bayesian variance component estimation, *J. Anim. Sci.* 80 (2002) 2540–2547.
- [18] Quintanilla R., Varona L., Noguera J.L., Testing genetic determinism in rate of hoof growth in pigs using Bayes factors, *Livest. Sci.* 105 (2006) 50–56.
- [19] Raftery A.E., Lewis S.M., How many iterations in the Gibbs sampler?, in: Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M. (Eds.), *Bayesian Statistics IV*, Oxford University Press, Oxford, UK, 1992, pp. 763–774.
- [20] Spiegelhalter D.J., Best N.G., Carlin B.P., Vanderlinde A., Bayesian measures of model complexity and fit (with discussion), *J. R. Stat. Soc. B* 64 (2002) 583–639.
- [21] Strandén I., Gianola D., Mixed effects linear models with *t*-distributions for quantitative genetic analysis: a Bayesian approach, *Genet. Sel. Evol.* 31 (1999) 25–42.
- [22] Varona L., Aplicaciones del muestreo de Gibbs en modelos de genética cuantitativa: Análisis de un caso de heterogeneidad de varianzas, Ph.D. thesis, Universidad de Zaragoza, Zaragoza, Spain, 1994.
- [23] Varona L., García-Cortés L.A., Pérez-Enciso M., Bayes factors for detection of quantitative trait loci, *Genet. Sel. Evol.* 33 (2001) 133–152.
- [24] Varona L., Gómez-Raya L., Rauw W.M., Clop A., Ovilo C., Noguera J.L., Derivation of a Bayes factor to distinguish between linked or pleiotropic quantitative trait loci, *Genetics* 166 (2004) 1025–1035.
- [25] Verdinelli I., Wasserman L., Computing Bayes factors using a generalization of the Savage-Dickey density ratio, *J. Am. Stat. Assoc.* 90 (1995) 614–618.
- [26] Wang C.S., Rutledge J.J., Gianola D., Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs, *Genet. Sel. Evol.* 26 (1994) 91–115.