HAL open science

GIBBSTHUR: Software for Estimating Variance Components and Predicting Breeding Values for Ranking Traits Based on a Thurstonian Model

Luis Varona, Andres Legarra

▶ **To cite this version:**

Luis Varona, Andres Legarra. GIBBSTHUR: Software for Estimating Variance Components and Predicting Breeding Values for Ranking Traits Based on a Thurstonian Model. Animals, 2020, 10 (6), 10.3390/ani10061001 . hal-02923496

HAL Id: hal-02923496

https://hal.inrae.fr/hal-02923496v1

Submitted on 27 Aug 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.

Distributed under a Creative Commons Attribution 4.0 International License

animals

MDPI

Article

GIBBSTHUR: Software for Estimating Variance Components and Predicting Breeding Values for Ranking Traits Based on a Thurstonian Model

Luis Varona [1,*] and Andrés Legarra [2]

[1] Departamento de Anatomía, Embriología y Genética Animal, Instituto Agroalimentario de Aragón (IA2), Universidad de Zaragoza, 50018 Zaragoza, Spain
[2] Génétique Physiologie et Systèmes d'Elevage (GenPhySE), Institut National de la Recherche Agronomique de Toulouse, 31326 Castanet-Tolosan, France; andres.legarra@inrae.fr
* Correspondence: lvarona@unizar.es

Received: 29 April 2020; Accepted: 4 June 2020; Published: 8 June 2020

Simple Summary: This article describes a new software (GIBBSTHUR) that provides Bayesian estimation of variance components and predictions of breeding values for ranking traits generated from equine competitions based on a Thurstonian approach. The GIBBSTHUR software was developed in FORTRAN 90 and can be executed in UNIX, OSX, or WINDOWS environments, and is freely available in a public repository (https://github.com/lvaronaunizar/Gibbsthur).

Abstract: (1) Background: Ranking traits are used commonly for breeding purposes in several equine populations; however, implementation is complex, because the position of a horse in a competition event is discontinuous and is influenced by the performance of its competitors. One approach to overcoming these limitations is to assume an underlying Gaussian liability that represents a horse's performance and dictates the observed classification in a competition event. That approach can be implemented using Montecarlo Markov Chain (McMC) techniques with a procedure known as the Thurstonian model. (2) Methods: We have developed software (GIBBSTHUR) that analyses ranking traits along with other continuous or threshold traits. The software implements a Gibbs Sampler scheme with a data-augmentation step for the liability of the ranking traits and provides estimates of the variance and covariance components and predictions of the breeding values and the average performance of the competitors in competition events. (3) Results: The results of a simple example are presented, in which it is shown that the procedure can recover the simulated variance and covariance components. In addition, the correlation between the simulated and predicted breeding values and between the estimates of the event effects and the average additive genetic effect of the competitors demonstrates the ability of the software to produce useful predictions for breeding purposes. (4) Conclusions: the GIBBSTHUR software provides a useful tool for the breeding evaluation of ranking traits in horses and is freely available in a public repository (https://github.com/lvaronaunizar/Gibbsthur).

Keywords: ranking traits; horse; breeding; Thurstonian model; Gibbs sampler

1. Introduction

Ranking traits are used frequently as selection criteria in competitive horse breeding programs [1,2]; however, the position of a horse in a competitive event has a discontinuous nature and is influenced by the performance of its competitors, which makes genetic evaluations based on standard linear models difficult. Several transformations for overcoming that limitation have been proposed—e.g., the squared root of ranks [3], the Snell transformation score [4,5], and the normalized score [6].

Another interesting approach is to assume an underlying Gaussian liability that represents the horse's performance and dictates the observed classification in competition events [1,7]. The original implementation of that approach involved a complex numerical integration [1], but Gianola and Simianer [8] proposed a Bayesian approach called the Thurstonian Model, which can be implemented using McMC methods such as the Gibbs Sampler [9]. In horse breeding, that model has been tested on simulated data [10] and has been applied to trotter horses [11] and the results of endurance competitions [12]. The practical use of the method has been limited, however, partly because of the absence of suitable and user-friendly software.

The objective of this note is to present a new free software (GIBBSTHUR) that implements the Bayesian estimation of variance components and the prediction of breeding values for ranking traits based on a Thurstonian Model. In addition, the software can analyze continuous and threshold traits jointly with the underlying liability of the ranking traits.

2. Theoretical Background

The Thurstonian Model assumes an underlying variable or liability ($\boldsymbol{l}$) that is transformed into ranking in competitions ($\boldsymbol{t}$). The model for the underlying variable can include systematic, random environmental and additive genetics effects, and a residual effect whose variance has to be predefined. Systematic effects generally include, among others, sex, year, or type of competition, but for ranking traits, it is interesting to include an "event" effect in order to take into account the heterogeneity of the level of the competitors linked with the prestige of the competitive events. The liability can be modelled alone or in conjunction with other continuous or threshold traits through the additive genetic, random environmental, or residual covariances between them.

To illustrate the model, suppose a simple bivariate model composed of $\boldsymbol{l}$ and a continuous Gaussian trait ($\boldsymbol{y}$) which includes systematic and random additive genetic effects only. The probability distribution of $\boldsymbol{l}$ and $\boldsymbol{y}$, given the remaining parameters in the model, is the following multivariate Gaussian distribution (MVN):

$$\begin{pmatrix} \boldsymbol{l} \\ \boldsymbol{y} \end{pmatrix} = MVN\left(\begin{matrix} \boldsymbol{X}_l\boldsymbol{b}_l + \boldsymbol{Z}_l\boldsymbol{a}_l \\ \boldsymbol{X}_y\boldsymbol{b}_y + \boldsymbol{Z}_y\boldsymbol{a}_y \end{matrix}, \boldsymbol{R} \otimes \boldsymbol{I} \right),$$

where $\boldsymbol{b}_l$ and $\boldsymbol{b}_y$ and $\boldsymbol{a}_l$ and $\boldsymbol{a}_y$ are the systematic and additive genetic effects for the liability ($\boldsymbol{l}$) and the continuous ($\boldsymbol{y}$) traits, respectively. Furthermore, $\boldsymbol{X}_l$, $\boldsymbol{X}_y$, $\boldsymbol{Z}_l$, and $\boldsymbol{Z}_y$ are the corresponding incidence matrices, and $\boldsymbol{R}$ is the residual covariance matrix. The prior distribution of the additive genetic effects is the following multivariate Gaussian distribution:

$$\begin{pmatrix} \boldsymbol{a}_l \\ \boldsymbol{a}_y \end{pmatrix} = MVN\left(\begin{matrix} 0 \\ 0 \end{matrix}, \boldsymbol{G} \otimes \boldsymbol{A} \right),$$

where $\boldsymbol{A}$ is the numerator relationship matrix and $\boldsymbol{G}$ is the additive genetic covariance matrix. In addition, the prior distributions of $\boldsymbol{R}$ and $\boldsymbol{G}$ are Inverted Wishart. The program uses a Gibbs Sampler approach [9] to obtain the posterior marginal estimates of all the unknowns in the model. The Gibbs sampler is an updating sampling scheme that requires sampling from the full conditional posterior distributions. The conditional distributions of systematic, random environmental, and additive genetic effects are univariate Gaussian, while the conditional distributions of the variance covariance matrices are Inverted Wishart. As with threshold traits, the Bayesian implementation of the model based on the Gibbs Sampler, after conditioning on liability, is similar to the bivariate model that has two continuous traits [13].

Therefore, the key step for the implementation within a McMC Bayesian approach is the defining of the conditional predictive distribution of the liabilities given the observed ranking data, as follows:

$$p\left(l_{kj}\middle|ELSE\right) \propto \prod_{j=1}^{J}\left[N\left(l_{kj}\middle|\mu_{kj},\sigma_{kj}^2\right)\right]^{\delta_{kl}}\boldsymbol{T_k},$$

where l_{kj} is the liability of the jth individual at the kth completion event and ELSE denotes the liabilities at the same and other competition events and the systematic and additive genetic effects for the ranking trait and the correlated Gaussian trait. Further,

$$\mu_{kj} = \boldsymbol{x_{lkj}b_l} + \boldsymbol{z_{lkj}a_l} - \left(y_{kj} - \boldsymbol{x_{ykj}b_y} - \boldsymbol{z_{ykj}a_y}\right) \times \frac{r^{12}}{r^{22}},$$

and

$$\sigma_{kj}^2 = \frac{1}{r^{11}},$$

where r^{ij} is the (i,j) element of $\boldsymbol{R}^{-1} = \begin{pmatrix} r^{11} & r^{12} \\ r^{12} & r^{22} \end{pmatrix}$, and $\boldsymbol{x_{lkj}}$, $\boldsymbol{x_{ykj}}$, $\boldsymbol{z_{lkj}}$, and $\boldsymbol{z_{ykj}}$ are the corresponding rows of the $\boldsymbol{X_l}$, $\boldsymbol{X_y}$, $\boldsymbol{Z_l}$, and $\boldsymbol{Z_y}$ matrices that link the liability (l_{kj}) and the Gaussian (y_{kj}) trait for the jth data at the kth event with the vectors of systematic ($\boldsymbol{b_l}$ and $\boldsymbol{b_y}$) and additive genetic ($\boldsymbol{a_l}$ and $\boldsymbol{a_y}$) effects. In addition, δ_{kj} is 1 if the jth competitor is present at the kth event and 0 otherwise, and $\boldsymbol{T_k}$ is an indicator set that denotes the order of liabilities within the kth competition. Therefore, the liabilities within each event are not independent, because the sample space is restricted by the ranking represented by $\boldsymbol{T_k}$. The conditional sampling of liabilities is performed sequentially. First, the liability of the horse that finishes in position one is set to 0 to achieve the identifiability of the model, whereas an "event" effect is estimated. An alternative, equivalent option used in [10] is to fix the "event" effect to 0 and sample the first horse from a truncated normal conditional on the other competitors. Second, the liabilities of the remaining individuals in the competition event are calculated by sampling from the truncated Gaussian distributions (TN) that have a mean μ_{kj} and a variance σ_{kj}^2, with limits defined by the liabilities of the preceding horse and of the horse classified just after it.

3. The Software: Input and Output Files

The software (GIBBSTHUR) was developed in FORTRAN90 and is based on the TM software [14]. It is available at https://github.com/lvaronaunizar/Gibbsthur. Running the program requires a parameter file ('parameters.txt'), which must include the names of the pedigree and data file; the number of additional traits and their nature (continuous, threshold, or ranking); the description of the model of analysis for all the traits (systematic, random, and genetic effects); and the hyperparameters of the Gibbs sampler implementation that comprise the total number of iterations, the number of burn-in iterations to discard, and the thin interval. The pedigree file must include a list of the individual, sire, and dam entries numbered from 1 to the number of individuals. The program allows the user to include genetic groups for unknown sires or dams. The data file must have an entry for each data, and should include the level for each systematic, random, and additive genetic effect, numbered from 1 to the maximum number of levels for each effect; and the performance for the continuous, threshold, and ranking traits. Missing data should be coded as 0. There is no limit to the number of competitors per competition. Empty positions are admitted in the results of competitions.

The execution of the program generates the following files: "Results.txt", "Breeding.txt", "Solutions.txt", and "Iterations.txt". The file "Results.txt" includes a summary of the posterior mean and standard deviations for the variance and covariance components; the heritabilities; the ratios of variance; and the additive genetic, permanent environmental, and residual correlations. They are computed with the output of the Gibbs sampler iterations after the burn-in. The file "Breeding.txt" includes the posterior mean and standard deviations of the breeding values for the ranking traits and

the additional traits, and they can be directly used for selection purposes. The file "Solutions.txt" includes the posterior mean and standard deviations of the systematic and random environmental effects for the ranking and the additional traits. The file "Iterations.txt" includes the samples of each iteration of the Gibbs sampler for the variance and covariance components, which can be used to analyze in detail the marginal posterior distributions. A more detailed description is available at https://github.com/lvaronaunizar/Gibbsthur.

4. Simulation Example

A simple pedigree comprising 100 sires and 500 dams generates 2000 horses that compete randomly in 2000 events of 10 horses each. On average, each horse competes in 10 events. The liability was generated by a model that included a general mean, one additive genetic effect, and a residual. The additive genetic and residual variances were set to 0.25 and 1 (h_l^2 = 0.2), respectively. In addition, we simulated a correlated continuous trait that had additive and residual variances of 1 and 2 (h_l^2 = 0.33), respectively, and additive genetic and residual correlations with the liability of the ranking trait of 0.3 and −0.2, respectively. The model of analysis with the GIBBSTHUR software included two systematic effects (a general mean and an event effect), and the additive genetic and residual random effects. We used a Gibbs sampler chain of 100,000 iterations after discarding the first 10,000. The initial values of variance components were established by dividing the raw phenotypic variance by the number of variance components in the model. The convergence was checked by a visual inspection (see Figure 1). The computing time was 51 min and 23 s with a i9-7960X processor (2.80 GHz).

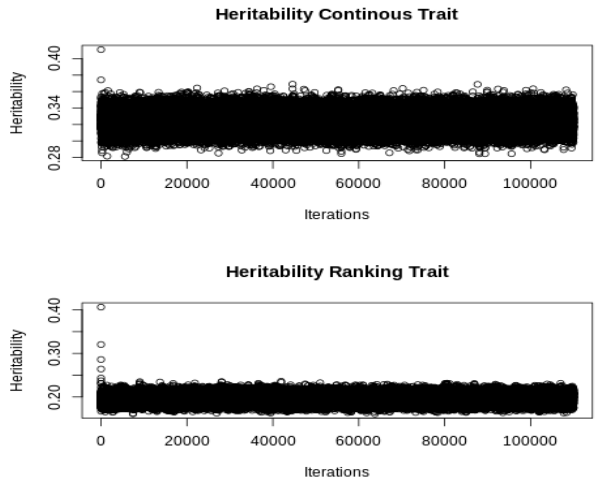

Figure 1. Plot of the Gibbs sampler iterations of the heritabilities of the continuous and ranking traits.

5. Results and Discussion

The posterior distribution of the heritabilities and the genetic and residual correlations is presented in Figure 2.

The simulated values were within the Highest Posterior Density (HPD) at 95% of the posterior distributions, which confirmed that the software can retrieve the simulated parameters for continuous and ranking traits. Slight differences between the mode of the posterior distributions and the simulated values can be expected, given the small size of the dataset. Figure 3 shows the predicted and simulated breeding values for the continuous and ranking traits.

The correlation between the predicted and the simulated breeding values was higher for the continuous (0.89) than it was for the liability of the ranking trait (0.83), although, in both cases, the correlation was very strong because the amount of information available for each horse was large (on average, 10 phenotypic records).

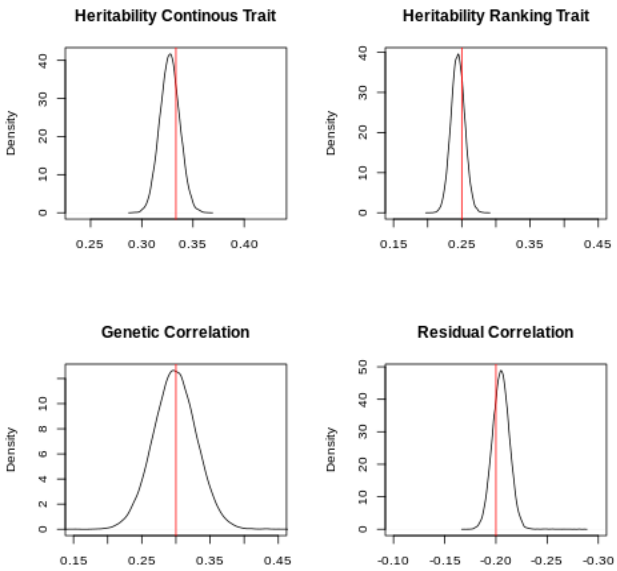

Figure 2. Posterior distribution and simulated parameters (in red) of the heritabilities of the continuous and ranking traits and of the genetic and residual correlations.

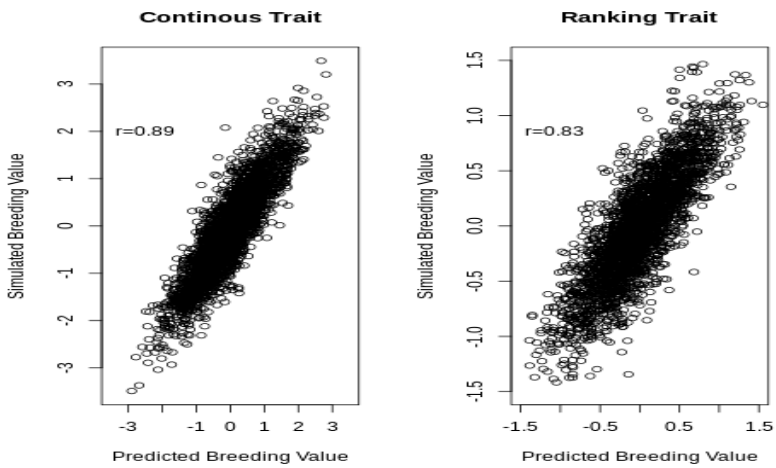

Figure 3. Two-dimensional plots of the simulated and predicted breeding values for the continuous and the liability of the ranking traits.

The ability to discriminate between the performances of the individuals in competition events is illustrated in Table 1, which provides a detailed description of the phenotypic information for the ranking traits of the best and worst predicted breeding values. The best individual won 8 of 13 races, and the horse with the worst breeding values came last in 5 of 8 races.

One of the main limitations of using ranking traits for the breeding evaluation of competitive horses is that competition events often are structured into categories based on their prestige or monetary reward and, therefore, the participation of horses in those events is not random, because the most prestigious events attract the best horses and the worst horses tend to participate in other events, which have a lower average performance. Gianola and Simianer [8] proposed including an "event" effect, although Legarra and Ricard [10] pointed out that "event" effects cannot be estimated from ranking data because the event does not have any influence on rankings, and they recommended removing them to achieve identifiability. Here, we used a different (but equivalent) approach that zeroed the liability of the best horse in each event, which forced the liabilities of the horses from the second to the last position to be negative. Therefore, the estimates of the event effects were also forced to be negative. In that approach, the systematic effect of the event is not an effect of the race per se

(i.e., weather or track conditions), but is linked to the performance of the best horse and provides an indicator of the average level of the competitors in the event.

Table 1. Phenotypic performance in the ranking trait of the horses with the highest and lowest predicted breeding value (EBV) for the liability of the ranking trait.

Place	Highest EBV	Lowest EBV
1	8	-
2	4	-
3	-	-
4	-	-
5	-	-
6	1	-
7	-	1
8	-	-
9	-	2
10	-	5

The ability to capture the level of the competition events is illustrated by the results shown in Figure 4, which represent the estimates of the event effects and the average true additive genetic effect for the liabilities of the horses competing in each event.

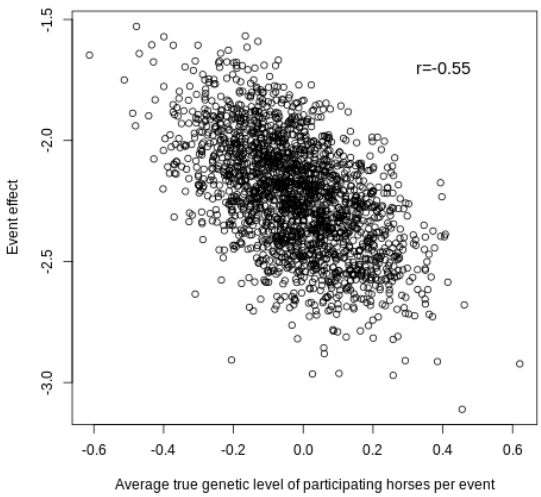

Figure 4. Two-dimensional plot of the estimate of the event effects and the average true genetic level for the liability of the ranking trait of the horses competing in an event.

As expected, there was a negative correlation (−0.55), which indicated that the most exigent events required a higher breeding value for the winner of the competition because the liability of the winner was set to zero. Therefore, competitions with the best horses (average genetic level) are associated with low "event" effects, while competitions of the worst horses have high "event" effects. In addition, Table 2 shows the predicted breeding values and the percentile within their ranking of the horses that competed in the highest and lowest estimated events in the simulation study.

In the event with the highest estimated effect, the predicted breeding values of 5 of 10 horses were ranked in the top 10%, and 8 out of 10 had a positive predicted breeding value. In contrast, in the event with the lowest estimated effect, all the horses had a negative predicted breeding value, and five had breeding values that were among the worst 10% of individuals. Nevertheless, the proposed approach might work well if the distribution of horses in events is random, as it was in the simulated example, or if the degree of structuring is small. In cases in which the structuring is high, Legarra and Ricard [10] proved with a simulation study that the results can be biased. Therefore, some other

alternatives might be appropriate—e.g., the approach suggested by these authors, who proposed a mixture model in which horses are classified into several groups based on their performance. Further research is needed to adapt the GIBBSTHUR software for use in those scenarios.

Table 2. Predicted breeding values and percentile within their ranking (between brackets) of the participants in the highest or lowest estimated events.

Position	Highest	Lowest
1st	1.55 (100.00)	−0.62 (9.64)
2nd	1.23 (99.50)	−0.69 (7.27)
3rd	0.69 (92.60)	−0.03 (48.19)
4th	0.86 (96.20)	−0.06 (46.19)
5th	0.73 (93.73)	−0.64 (9.03)
6th	0.46 (82.88)	−0.12 (40.58)
7th	0.03 (53.92)	−0.92 (2.65)
8th	0.32 (75.80)	−0.51 (13.73)
9th	−0.35 (22.00)	−0.31 (24.50)
10th	−0.15 (34.15)	−0.89 (3.07)

6. Conclusions

The GIBBSTHUR software provides a useful tool for the breeding evaluation of ranking traits in horses. It can analyze ranking traits jointly with other categorical or continuous traits.

Author Contributions: Conceptualization, L.V.; methodology, L.V. and A.L.; software, L.V. and A.L.; validation, L.V.; writing—original draft preparation, L.V.; writing—review and editing, L.V. and A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tavernier, A. Genetic evaluation of horses based on ranks in competitions. *Genet. Sel. Evol.* **1991**, *23*, 159–173. [CrossRef]
2. Ricard, A.; Chanu, I. Genetic parameters of eventing horse competition in France. *Genet. Sel. Evol.* **2001**, *33*, 175–190. [CrossRef] [PubMed]
3. Lührs-Behnke, H.; Röehe, R.; Kalm, E. Genetic analysis of riding test and their connections with traits of stallion performance and breeding mare tests. *Züchtungskunde* **2006**, *78*, 119–128.
4. Snell, E.J. A scaling procedure for ordered categorical data. *Biometrics* **1964**, *20*, 592–607. [CrossRef]
5. Gómez, M.D.; Cervántes, I.; Bartolomé, E.; Molina, A.; Valera, M. Genetic evaluation of show jumping performance in young Spanish Sport Horse Breed. In *Book of Abstracts of the 57th Annual Meeting of the EAAP*; Wageningen Academic Publishers: Antalya, Turkey, 2006.
6. Aldridge, L.I.; Kelleher, D.L.; Reilly, M.; Brophy, P.O. Estimation of the genetics correlation between performances at different levels of show jumping competitions in Ireland. *J. Anim. Breed. Genet.* **2000**, *117*, 65–72. [CrossRef]
7. Tavernier, A. Estimation of breeding value of jumping horses from their ranks. *Lives. Prod. Sci.* **1990**, *26*, 277–290. [CrossRef]
8. Gianola, D.; Simianer, H. A thurstonian model for quantitative genetic analysis of ranks: A Bayesian approach. *Genetics* **2006**, *174*, 1613–1624. [CrossRef] [PubMed]
9. Gelfand, A.E.; Smith, A.F.M. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **1990**, *85*, 398–409. [CrossRef]
10. Ricard, A.; Legarra, A. Validation of models for analysis of ranks in horse breeding evaluation. *Genet. Sel. Evol.* **2010**, *42*, 3–10. [CrossRef] [PubMed]
11. Gómez, M.D.; Varona, L.; Molina, A.; Valera, M. Genetic evaluation of racing performance in trotter horses by competitive models. *Lives. Sci.* **2011**, *140*, 155–260. [CrossRef]

12. García-Ballesteros, S.; Varona, L.; Valera, M.; Gutierrez, J.P.; Cervántes, I. Cross-validation analysis for genetic evaluation models for ranking in endurance horses. *Animal* **2018**, *12*, 20–27. [CrossRef] [PubMed]
13. Varona, L.; Misztal, I.; Bertrand, J.K. Threshold-linear versus linear-linear analysis of birth weight and calving ease using an animal model: I. Variance component estimation. *J. Anim. Sci.* **1999**, *77*, 1994–2002. [CrossRef] [PubMed]
14. Legarra, A.; Varona, L.; López de Maturana, E. Threshold Model (TM). Available online: http://snp.toulouse.inrafr/~{}alegarra/manualtm.pdf (accessed on 10 March 2020).

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).