This is to certify that the
dissertation entitled

HIERARCHICAL BAYES MODELS FOR BEEF CATTLE
GENETIC EVALUATION UNDER EXTENSIVE
MANAGEMENT CONDITIONS

presented by

FERNANDO FLORES CARDOSO

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Animal Science

Robert J. Tempelman
Major Professor's Signature

August 14, 2003
Date

# HIERARCHICAL BAYES MODELS FOR BEEF CATTLE GENETIC EVALUATION UNDER EXTENSIVE MANAGEMENT CONDITIONS

By

Fernando Flores Cardoso

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Animal Science

2003

# ABSTRACT

## HIERARCHICAL BAYES MODELS FOR BEEF CATTLE GENETIC EVALUATION UNDER EXTENSIVE MANAGEMENT CONDITIONS

By

Fernando Flores Cardoso

The overall aim of this project was to investigate several incompletely resolved issues in statistical modeling applied to quantitative genetic inference of extensively managed multiple-breed beef populations using Bayesian inference based on MCMC methods. A hierarchical animal model (HIER) was developed for inference on genetic merit of livestock with uncertain paternity. This model was compared to a model based on Henderson's average numerator relationship (ANRM) in a simulation study and in an application to growth data from Brazilian Herefords. For both simulated and Hereford data, posterior inference on variance components was similar for ANRM and HIER, and rank correlations on posterior means for genetic effects between the two models exceeded 0.90. Nonetheless, large differences in these posterior means between the two models were observed for some animals. Furthermore, animals with uncertain paternity had generally larger posterior standard deviations of genetic effects using the HIER model likely because this model accounts for the uncertainty on sire assignment probabilities. Bayesian model choice criteria consistently favored the HIER model over the ANRM model in both simulated and Hereford data.

A hierarchical multiple-breed animal model (MBAM) was proposed and applied to estimate genotypic effects, breed-specific additive genetic variances and variances due to the segregation between breeds. Phenotypic records were modeled as function of additive (A), dominance and A × A genetic fixed effects and random animal additive genetic effects using appropriate multiple-breed additive variance-covariance specifications. MBAM was validated on five two-breed simulated datasets and applied to the analysis of post-weaning gain (PWG) records from Nelore-Hereford crosses. MBAM inference on Nelore and Hereford genetic variances differed substantially and a non-zero segregation variance was estimated between these breeds. The Pseudo Bayes Factor (PBF) heavily favored the MBAM over the conventional animal model for both simulated and PWG data. The main advantage of MBAM is the flexibility in modeling heteroskedastic genetic variances of the breed composition groups, hence improving genetic predictions. Finally, the MBAM was extended to allow residual heteroskedasticity and robustness using structural variance models. Six alternative structural specifications were evaluated: Gaussian homoskedastic and heteroskedastic; Student *t* homoskedastic and heteroskedastic; and Slash homoskedastic and heteroskedastic. Based on the PBF, the Student *t* heteroskedastic model provided the best fit to PWG whereas the Gaussian homoskedastic model provided the worst fit. Amongst the fixed factors considered for residual heteroskedasticity (breed proportion, heterozygosity and sex) only heterozygosity appeared to be important. Considerable heteroskedasticity was inferred across random contemporary groups. Inference on genetic variance components changed substantially depending on the structural specification for the residual variance. Furthermore, inference based on the conventional Gaussian homoskedastic model led to significant rerankings of animal genetic effects compared to the better fitting Student *t* heteroskedastic specification, thereby having important implications for genetic improvement programs.

To Magali and Nicole for their love and support

# ACKNOWLEDMENTS

I am deeply grateful for the extraordinary guidance, support and friendship that Dr. Robert J. Tempelman has given me during the last four years, and which have been decisive to my scholarship. I truly valued his never-ending patience and willingness to help the countless times I have come to his office without an appointment. My gratitude is extended to Drs. Dennis Banks, Ricardo Cardellino, Bryan Epperson, Cathy Ernst and David Hawkins for their support to this work and for serving in my guidance committee. I have to specially thank Dr. Ricardo Cardellino; without his encouragement I would not have come to Michigan State University. Appreciation is given to Dr. Guilherme Rosa for kindly revising and providing valuable suggestions, particularly on Chapter 4 of this dissertation. I also thank the assistance given by Dr. Peter Saama at Quantitative Genetics Lab. I greatly appreciated the help and the very productive discussions I had with my fellow graduate students, Dr. Kadir Kizilkaya, Dave Edwards, Juan Steibel and Lan Xiao.

I am in debt with CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Ministry of Education, Brasília, Brazil, for funding tuitions, fees and scholarship during the whole course of my degree. I would also like to thank the College of Agriculture and Natural Resources for support through the dissertation completion fellowship.

Finally, I would like to thank my wife, Magali, my daughter, Nicole, my parents, Francisco and Teresinha, and my brother, Eduardo, for their unconditional love and support and for their great patience and understanding, especially in the toughest days of the last four years.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# INTRODUCTION

Beef products are sources of essential dietary amino acids and microelements (e.g. iron). Supplying the market with these high quality products at a competitive price depends on maximization of efficiency in production systems.

Animal breeders have a key role in the improvement of beef production systems. Selection and planned crossbreeding systems can synergistically create more desirable animal biotypes that match current production systems. The use of heterosis and complementarity between breeds through crossbreeding (Gregory et al., 1999) is a tool available to increase the efficiency of production without large increase in costs. Since this has led to an increasing proportion of beef cattle populations being crossbred, genetic evaluations have been further complicated by the varying genetic backgrounds and degrees of crossbreeding found in these populations.

In order to accurately predict performance and to enhance genetic progress in crossbred populations, it is necessary to develop and apply statistical methodology that accounts for all salient sources of genotypic differences in economically important traits. The complexity of the biological and environmental issues involved requires extensive research effort. Bayesian statistics provides a set of flexible tools and a general modeling framework in this regard (Sorensen and Gianola, 2002). Hierarchical Bayes models (HBMs) can account for virtually any level of complexity that is present in the population of interest and are particularly useful when records are correlated (Hobert, 2000), as is typical of related animals. Moreover, HBMs allow for optimal combination of information present in the data with previous inferences from the literature to estimate the parameters of interest (e.g. genotypic means). The current "state-of-art" multiple-breed genetic evaluation model for beef cattle in the United States uses a HBM to incorporate prior knowledge on heterosis (Klei et al., 1996).

### 1. Prediction of performance and genetic merit

The genetic value of an animal can be determined by the mean of its breed-composition or genotypic group plus an individual deviation from its group (Arnold et al., 1992; Elzo, 1994; Klei et al., 1996; Sullivan et al., 1999). Several approaches have been considered to estimate means of breed-composition groups in multiple-breed populations. The simplest strategy involves including breed-composition in the definition of the contemporary group (CG) and estimating heterotic effects jointly with the CG effects. However, this method reduces the number of possible direct comparisons and connectedness in the population, since animals with different compositions are considered to be in different CG even when they are raised together under the same management and environmental conditions (Klei et al., 1996). Parsimonious models are obtained by estimating breed-composition means as a function of additive (breed proportion) and non-additive (degree of allelic and non-allelic interaction) genetic coefficients. If heterosis is primarily due to dominance (allelic interaction) with no epistasis, then it is proportional to heterozygosity (proportion of heterozygotes at individual loci) (Gregory et al., 1999). Dickerson (1969; 1973), however, introduced the concept of "recombination loss" to explain deviations from the heterozygosity found in crossbred individuals. The recombination loss is equal to "the average fraction of independently segregating pairs of loci in the gametes from both parents which are expected to be non-parental combinations" (Dickerson, 1969). The effect of recombination loss is attributable to the loss of favorable epistatic combinations present in the gametes from purebreds as a result of long-term selection. Kinghorn (1987) proposed several hypotheses and models to account for "epistatic loss" in crossbred populations, and Wolf et al. (1995) proposed a general model based on the two-loci theory to account for dominance and epistatic effects.

Confoundedness and multicollinearity between the coefficients for genetic effects complicates the estimation of dominance effects separately from epistatic effects such that most of the models proposed for multiple breed evaluations are only based on dominance effects (Cunningham, 1987; Klei et al., 1996; Miller and Wilton, 1999; Sullivan et al., 1999).

Accounting for additive and heterotic mean effects on genetic evaluations can be accomplished by several approaches, for example: by using information in the literature to pre-adjust records (Roso and Fries, 1998; Sullivan et al., 1999), provided that the published estimates are reliable and applicable to the population being evaluated; by estimating these mean effects solely from the data of the population under investigation (Arnold et al., 1992; Miller and Wilton, 1999); or by simultaneously using information from the literature combined with data information, as in the benchmark model used currently in the U.S. beef industry (Klei et al., 1996; Quaas and Pollak, 1999).

A deviation of the genetic merit of an individual from its group mean is due to additive and non-additive genetic effects. Additive effects or breeding values indicate the deviance from the population mean expected in the offspring of an individual when it is mated at random to other individuals in the population, whereas non-additive effects are useful to determine specific combining abilities between individuals (Falconer and Mackay, 1996). These deviations are determined by the performance of an individual and its relatives; therefore, it is important to properly account for covariances between relatives when predicting genetic value of crossbred animals.

Theory to estimate the covariance between crossbred animals was presented by Lo et al. (1993) for an additive model and by Lo et al. (1995) for an additive and dominance model. Under the additive model, (co)variances are modeled as a function of breed specific additive variances and variances due to the segregation between breeds. These segregation variances represent the additional variance observed in $F_2$ individuals compared to the $F_1$'s (Lo et al., 1993). These methods derive genetic means and covariances between crossbred and purebred individuals from "identity modes" used to determine the probability that related individuals share alleles that are identical-by-descent (IBD). The additive and dominance model is derived for a two-breed and their crosses scenario (Lo et al., 1995). This model has an exact theoretical derivation and can accommodate the presence of inbreeding, but requires a relatively larger number of variance components to be estimated (up to 25 when inbreeding is present). Simplifications arise when the population is composed only by the two pure breeds and F1's (Lo et al., 1997), and this model has been applied to swine data (Lutaaya et al., 2001).

For more general crossbreeding schemes, the dominance model (Lo et al., 1995) can be cumbersome due to the large number of dispersion parameters to be estimated, while the additive model (Lo et al., 1993) can be implemented without great difficulty. An alternative formulation of the additive model with a regression approach to account for non-additive effects and a sire-maternal grandsire model implementation was proposed by Elzo (1994) and applied to multiple-breed data (Elzo et al., 1998; Elzo and Wakeman, 1998). Recently, Birchmeier et al. (2002) proposed an algorithm using restricted maximum likelihood (REML) to estimate additive breed and segregation variances under a typical animal model and general pedigree structure. Yet, several recently proposed models (Klei et al., 1996; Miller and Wilton, 1999; Quaas and Pollak, 1999; Sullivan et al., 1999) assume that all breeds have the same additive genetic variance and there is no variance due to segregation between breeds in advanced crosses. A model including additive and non-additive genotypic effects and random additive individual deviations may offer a parsimonious model for genetic evaluation of multiple-breed populations.

### 2. Multiple-sire mating and uncertain paternity

Extensive beef cattle production systems often rely upon multiple-sire mating to increase the probability of pregnancy, when the size of breeding groups, as a consequence of paddock size, is too large to be sired by a single bull. Breeding cows are exposed to more than one male within the same breeding season and consequently calves born from these matings have uncertain paternity; they are known only to be sired by one of the bulls in the mating group. This situation frequently occurs in pastoral operations such as those found in Argentina, Australia, Brazil and parts of the United States. Other causes of uncertain parentage include the use of artificial insemination followed by natural breeding, accidental/unplanned breeding and insemination with pooled semen. Pedigrees in these herds are uncertain and this can impact genetic evaluations by decreasing accuracy of genetic value prediction and by reducing selection intensity if animals with uncertain paternity are not considered for selection or are not included in the evaluation.

Statistical methods have been developed for genetic evaluation of animals with uncertain pedigree. A simple method that has been used is genetic grouping (Westell et al., 1988), where "phantom parents" are assigned to animals with uncertain sire. Here, "phantom parents" are grouped according to some criteria (e.g. gender, year of birth) such that this group effect, fixed or random, is estimated. This specification is equivalent to the assumption of an infinite number of non-inbred, unrelated possible sires, all having equal probabilities (Perez-Enciso and Fernando, 1992; Sullivan, 1995). However, a finite number of putative sires should be considered, if identification of each sire in each group is known. The average numerator relationship matrix (ANRM), as proposed by Henderson (1988), consists in constructing a relationship matrix based on the probabilities of each putative male being the correct sire. This relationship matrix is the correct covariance between animals when true probabilities are known (Perez-Enciso and Fernando, 1992), and can be used to provide best linear unbiased predictions (BLUP) of genetic merit for all sires based on the records of certain and/or uncertain progeny, records of other relatives, and their own mates and records. A simple and rapid method to compute the inverse of the ANRM is available (Famula, 1992). This method relaxes the assumptions of no inbreeding and no relationship between candidate sires which are required in genetic grouping.

The advantage in terms of selection response of using ANRM compared to genetic grouping, when putative sires are recorded, has been demonstrated through simulation (Perez-Enciso and Fernando, 1992; Sullivan, 1995). Differences in favor of ANRM between the two models were larger when $h^2$ was low and uncertainty high. Nonetheless, this comparison entails that the true probabilities are known. With no prior knowledge, equal probabilities might be assumed for each sire; however information from blood types, genetic markers, records of mating behavior, fertility, breeding period and gestation length could be used to assign probabilities that a given offspring has been sired by different males (Foulley et al., 1987; Henderson, 1988). Nevertheless, it is unrealistic to assume that those are the true probabilities and the BLUP properties of predictions based on ANRM are seldom, if ever, attained.

A less restrictive method that deals with uncertainty on paternity probabilities and different type of data sampling distributions (normal and binomial) is the empirical Bayes procedure proposed by Foulley et al. (1987; 1990). This method uses the data and an approximate algorithm to calculate posterior probabilities of sire assignments. However, its use is limited to sire models, not being developed for more general cases, such as animal models.

Despite the encouraging results found in the literature, methods taking into account uncertain paternity are not broadly used in genetic prediction. This could be due to the fact that ANRM, the most studied and ready to implement method, requires the knowledge of true paternity probabilities or at least "good" approximations of these probabilities. And, on the other hand, methods that overcome such requirements, such as those of Foulley et al (1987) and Im (1992), are not generalized to the animal model, most commonly used in such predictions.

Recent developments in molecular biology, statistics and computational power have provided some solutions. Methods such as blood typing and genotyping could be used to determine true paternity. Despite their precision, it is improbable that they could be used in large scale with the unique purpose of determining paternity due to their costs.

However, genetic markers developed for other purposes may provide a source of prior information that could be combined with the data to increase accuracy on prediction of performance and to maximize selection response on populations undergoing multiple-sire mating. The Bayesian framework facilitates the development of models that fully account for uncertainty on parentage, combining molecular prior information or subjective prior information with performance of the individual, offspring and relatives, to predict genetic merit of individuals pertained to or born from group mating. These predictions would be a function of the posterior probabilities of each bull in the group being the correct sire of the individual given the prior and data information. The proper genetic analysis of animals with uncertain paternity would enhance genetic improvement and consequently economic productivity of large populations raised in pastoral conditions and undergoing multiple-sire mating.

### 3. Heteroskedasticity and robustness

Current methods for genetic evaluation based on Henderson's mixed model equations (Henderson, 1975) require knowledge of variance components to provide BLUP of genetic values. Often, genetic and residual variances are assumed to be constant across environments in these evaluations. However, heteroskedasticity has been reported in beef cattle for growth performance (Garrick et al., 1989; Nunez-Dominguez et al., 1995; Rodriguez-Almeida et al., 1995) and carcass scan traits (Reverter et al., 1997). Region, herd, level of production, herd size, year, sex and class of age of dam are possible sources for heterogeneity. For the case of crossbred populations, breed composition can also be considered (Arnold et al., 1992; Garrick et al., 1989; Rodriguez-Almeida et al., 1995). A parsimonious model for heterogeneous variances is essential, since the number of parameters to be estimated can increase dramatically, making such analyses unfeasible (Foulley and Quaas, 1995).

In general, the phenotypic variation in beef cattle weights increases proportionally with the mean; e g. weaning weights have larger variance than birth weights (Koots et al., 1994). Variance of performance in animals raised on better environmental conditions might be expected to be larger than those grown in poorer environments and males may be expected to have higher variability compared to females due to their typically higher weights.

When several herds in different environments are involved in a genetic evaluation program, the accuracy of selection will depend on a reasonably correct specification of variance components in mixed model equations. The effect of heterogeneity of variances is particularly important for the selection of cows, young bulls and heifers, because these animals have records within one herd or environment and thus their evaluation can be greatly affected by differences in environmental variances (Winkelman and Schaeffer, 1988).

Methods to assess sources of heterogeneity of variance have been proposed by Gianola et al. (1992) and SanCristobal et al. (1993). The method of Gianola et al. (1992) is based on regarding herd residual variances as random variables from a conveniently defined distribution (i.e. scaled inverted chi-square). The estimates obtained represent a compromise between a data based statistic (REML) and parameters of the distribution of variances (hyperparameters). When the amount of information in particular stratum is large, the REML part of the estimator dominates; otherwise the prior distribution is weighted more heavily. Another advantage of this hierarchical specification is in terms of borrowing of information across subclasses as is true for conventional random effects models. In San Cristobal et al. (1993), an extension of the structural linear method for log variances (Foulley et al., 1992) to genetic and residual effects is presented. The method uses a log link for variances i.e. $\ln \sigma_i^2 = \mathbf{m}_i'\boldsymbol{\lambda}$, where $\sigma_i^2$ is the variance per subclass, $\mathbf{m}_i'$is a row incidence vector and $\boldsymbol{\lambda}$, a vector of unknown parameters influencing heteroskedasticity. Procedures to estimate $\boldsymbol{\lambda}$ were presented for marginal likelihood and Bayesian point of view, assuming informative prior information on components of $\boldsymbol{\lambda}$.

Other than heteroskedasticity, the presence of observations influenced by factors not accounted for in the statistical analysis and having potentially extreme influence (i.e. outliers) can severely bias the genetic merit predictions, since most linear mixed models used in animal breeding assume normally distributed residual and random effects. The normal distribution is particularly vulnerable to presence of outliers (Rogers and Tukey, 1972).

Preferential treatment, inappropriate contemporary group formation, record errors and animal misidentification are possible causes of outliers in beef cattle populations. The presence of outliers is normally investigated prior to data analysis. This editing generally consists of deleting observations that are considered extremely far from the phenotypic mean of its class (i.e. greater than three standard deviations). However, the edits used in determining which records are outliers are somewhat ad-hoc in nature and need to be justified, particularly to the breeder(s) affected.

As an alternative to the deletion of observations, some symmetric heavy-tailed distributions, such as Student *t*, Slash and Contaminated Normal, have been suggested and applied in place of the normal distribution for robust estimation (Lange and Sinsheimer, 1993). These distributions are examples of Normal/independent families that can better accommodate extreme observations due to their heavy-tailed feature (Lange and Sinsheimer, 1993; Rogers and Tukey, 1972). They can be defined as the distribution of *k* dimensional random vector $\mathbf{y} = \boldsymbol{\mu} + \frac{\mathbf{e}}{\sqrt{w}}$, where $\boldsymbol{\mu}$ is the location vector of dimension *k*, *w* is a positive random variable with density $p(w|v)$ (depending on the family distribution), where *v* is a robustness parameter (degrees of freedom), and **e** is a *k* dimensional random vector normally distributed with mean **0** and nonsingular covariance matrix $\Sigma$. Conditional to *w*, **y** is normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma / w$. Moreover, the distribution of *w* defines the marginal distribution of **y**.

In animal breeding, Stranden and Gianola (1999) recently introduced a hierarchical Bayes model that specifies the residuals to have Student *t* rather than normal densities. They presented a Monte Carlo Markov Chain (MCMC) strategy to provide inferences on breeding values. Previous results from the same authors indicate that the Student *t* better accommodates data situations that involve a prevalence of preferential treatment, compared to the normal distribution (Stranden and Gianola, 1998).

It is possible to construct hierarchical Bayes models that parsimoniously account for heteroskedasticity, while being robust to outliers. Extensions of the methods proposed by Foulley et al. (1992) and SanCristobal et al. (1993) combined with Normal/independent distributions (Lange and Sinsheimer, 1993) may provide robust tools to identify sources of heterogeneity of variance in multiple-breed populations tested in diverse environments by means of fully Bayesian inference.

## 4. Bayesian inference in animal breeding

The milestone paper that introduces Bayesian inferences to animal breeding research is credited to Gianola and Fernando (1986). The most striking, and perhaps controversial, difference between Bayesian and classical (or frequentist) inference is that Bayesian inference allows the incorporation of prior knowledge (Blasco, 2001). From a practical point of view, if significant prior information is available, then ignoring it seems poorly advised, especially when the inference complexity is high and data information is limited.

Hierarchical or multistage models are used in Bayesian inference to functionally describe complex problems through a series of nested levels or sub-models (Sorensen and Gianola, 2002). Distributional assumptions and parameter values associated with these distributions (hyperparameters in Bayesian terminology) are used to integrate prior knowledge in the analyses. The Henderson's mixed model equations (Henderson, 1973) widely used in animal breeding are a classical example of two stage model.

Inferences (e.g. estimation of genotypic means or prediction of breeding values) are derived from the joint posterior density, which consists of the product of all hierarchically specified stages of the model. There are two primary methods to obtain estimates: 1) an empirical Bayes approach, in which the joint mode of all parameters is obtained by iterative methods, such as the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) and approximate "large sample" standard error derived from an information matrix; 2) a fully Bayes approach, in this case MCMC, a simulation-intensive algorithm, is used to derive marginal densities obtaining "exact" small sample inference on all parameters (Gilks, 1996). The Metropolis-Hasting algorithm (Hasting, 1970; Metropolis, 1953) and the Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984) are the most common MCMC strategies used in animal breeding. A large number of cycles are generated and samples are saved. Eventually, the sampler converges to the joint posterior distribution. Values of each parameter drawn after convergence are considered random samples from its marginal posterior distribution and used to draw inference (e.g. means, modes, medians, standard deviations, credibility sets, etc.) (Sorensen and Gianola, 2002).

Fully Bayesian methods have been used in the last decade for inference in animal breeding problems in several applications, including variance component estimation (Jensen et al., 1994; Wang et al., 1994b), prediction of selection response (Sorensen et al., 1994; Wang et al., 1994a) and in threshold models for categorical data (Sorensen et al., 1995; Wang et al., 1997). The possibility of combining prior and data information, and the ability to provide exact small sample inference, make Bayesian methods attractive for animal breeding and genetics problems, especially when the number of parameters exceeds the number of observations.

### 5. General hypothesis

Precision of genetic merit prediction can be improved by the development and application of Bayesian methods in the genetic evaluation of multiple-breed beef cattle populations.

### 6. Specific aims

This project sought to improve genetic evaluations of multiple-breed populations, taking advantage of the versatility of Bayesian statistics in dealing with complex biological and environmental issues arising in these evaluations. The use of Bayesian methods had been limited in the past by their computational requirements, but is now possible due to the rapid increase in speed and memory capacity of PCs and workstations.

The issues addressed are relevant topics for the beef industry; however the technology being developed is general enough to be applied to other livestock species. More accurate analyses will result in better prediction of breeding values and combining abilities between individuals of different backgrounds, leading to an ultimate improvement of efficiency in current production systems.

The overall aim was to investigate some incompletely resolved questions in statistical modeling applied to the estimation of genetic parameters of extensively managed multiple-breed populations. These issues included the partition of genetic variances, uncertain paternity, heterogeneity of residual variances and robustness to outliers. As part of this aim, fully Bayes genetic evaluation software, to be applied in multiple-breed beef cattle populations was developed. The specific objectives were:

1) To develop and apply a hierarchical Bayes model for genetic evaluation of animals originated from multiple-sire mating systems;
2) To develop and apply a hierarchical Bayes model for genetic evaluation of animals in multiple-breed populations;
3) To extend the genetic evaluation models in 1) and 2) to account for heterogeneity of residual variances across environments and provide greater robustness to outliers.

# CHAPTER 1

# BAYESIAN INFERENCE ON GENETIC MERIT UNDER UNCERTAIN PATERNITY

**ABSTRACT:** A hierarchical animal model is developed for inference on genetic merit of livestock with uncertain paternity. Fully conditional posterior distributions for fixed and genetic effects, variance components, sire assignments and their probabilities are derived to facilitate a Bayesian inference strategy using MCMC methods. We compare this model to a model based on Henderson's average numerator relationship (ANRM) in a simulation study with 10 replicated datasets generated for each of two traits. Trait 1 had a medium heritability ($h^2$) for each of direct and maternal genetic effects whereas Trait 2 had a high $h^2$ attributable only to direct effects. The average posterior probabilities inferred on the true sire were between 1 and 10% larger than the corresponding priors (the inverse of the number of candidate sires in a mating pasture) for Trait 1 and between 4 and 13% larger than the corresponding priors for Trait 2. The predicted additive and maternal genetic effects were very similar using both models; however, model choice criteria (Pseudo Bayes Factor and Deviance Information Criterion) decisively favored the proposed hierarchical model over the ANRM model.

**Key Words:** Uncertain paternity, Multiple-sire, Genetic merit, Bayesian inference, Reduced animal model.

## 1. Introduction

Multiple-sire mating is common on large pastoral beef cattle operations in Argentina, Australia, Brazil and parts of the United States, for example. Here, groups of cows are exposed to several males within the same breeding season. Consequently, pedigrees in these herds are uncertain, adversely affecting accuracy of genetic evaluations and selection intensities.

A number of statistical models have been proposed for genetic evaluation of animals with uncertain paternity. One simple solution appears to be genetic grouping (Westell et al., 1988), whereby "phantom parents" groups are assigned to animals within the same mating pasture. In genetic grouping, phantom parent groups are typically defined to be a contemporary cluster of unknown parents in order to minimize bias on breeding value predictions due to genetic trend (Cantet et al., 1993; Quaas, 1988). The use of genetic grouping for multiple-sire mating, however, is equivalent to assuming an infinite number of non-inbred, unrelated candidate sires within each group, each candidate having the same probability of being the correct sire (Perez-Enciso and Fernando, 1992; Sullivan, 1995) of the animal with uncertain paternity. However, only the candidate sires actually used within a group or pasture should be considered.

This requirement is more aptly handled with the average numerator relationship matrix (ANRM) proposed by Henderson (1988). The ANRM is based on knowledge of true probabilities of each candidate male being the correct sire. The ANRM helps specify the correct genetic variance-covariance matrix when these probabilities are presumed known (Perez-Enciso and Fernando, 1992), thereby facilitating best linear unbiased predictions (BLUP) of genetic merit. A simple and rapid algorithm to compute the inverse of the ANRM is available (Famula, 1992) and the advantage in selection response of using ANRM versus genetic grouping, when candidate sires are recorded, has been demonstrated by simulation studies (Perez-Enciso and Fernando, 1992; Sullivan, 1995). Equal prior probabilities might be assumed for each sire; however information from blood typing, genetic markers, mating behavior, fertility, breeding period and gestation length could also be used to make these probabilities more distinctive (Foulley et al., 1987; Henderson, 1988).

A novel empirical Bayes procedure to infer upon uncertain paternity was proposed by Foulley et al. (1987; 1990). Their sire model implementation combines data and prior information to determine the posterior probabilities of sire assignments for each animal with uncertain paternity. With the advent of Markov chain Monte Carlo (MCMC) techniques in animal breeding (Wang et al., 1994b), it is now possible to extend their method to an animal model and allow a more formal assessment of statistical uncertainty on genetic merit and of probabilities of sire assignments.

The objectives of this study were to: 1) develop a hierarchical animal model and Bayesian MCMC inference strategy for the prediction of genetic merit on animals having uncertain paternity; 2) use this model to estimate posterior probabilities of paternity, by combining prior and data information; and 3) compare the performance of the proposed model with a model based on the use of Henderson's ANRM having equal prior probability assignments for all candidate sires.

## 2. The Bayes hierarchical model

### 2.1. The reduced animal model with maternal effects

Consider an $n \times 1$ data vector $\mathbf{y} = \{y_{ij}\}$, $i$=1,2,...,$n$; $j$=1,2...,$q$. Here $i$ identifies the record and $j$ the animal associated with the $i$th record. We allow for the possibility of any animal $j$ having no record; nevertheless, a genetic evaluation may be desired on that same animal if it is related to other animals having data. In the reduced animal model (RAM) of Quaas and Pollak (1980), **y** is partitioned into two major subsets:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_p \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} \mathbf{X}_p \\ \mathbf{X}_t \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_{1p} \\ \mathbf{Z}_{1t}\mathbf{P}_t \end{bmatrix} \mathbf{a}_p + \begin{bmatrix} \mathbf{Z}_{2p} \\ \mathbf{Z}_{2t} \end{bmatrix} \mathbf{m}_p + \begin{bmatrix} \mathbf{e}_p \\ \mathbf{Z}_{1t}\boldsymbol{\gamma}_t + \mathbf{e}_t \end{bmatrix}. \qquad [1]$$

The first $n_p \times 1$ subset $\mathbf{y}_p$ of **y** is observed on $q_p$ animals that are identified as *parents* or ancestors of other animals having data. In [1], $\mathbf{y}_p$ is a linear function of a $p \times 1$ vector of "fixed" effects $\boldsymbol{\beta}$, a $q_p \times 1$ vector of additive direct genetic effects $\mathbf{a}_p$, and a $q_p \times 1$ vector of additive maternal genetic effects $\mathbf{m}_p$. Here, $\mathbf{a}_p$ and $\mathbf{m}_p$ correspond to effects on the $q_p$ parents. The design matrices connecting $\mathbf{y}_p$ to $\boldsymbol{\beta}$, $\mathbf{a}_p$ and $\mathbf{m}_p$ are $\mathbf{X}_p$, $\mathbf{Z}_{1p}$, and $\mathbf{Z}_{2p}$, respectively. The remaining $n_t \times 1$ data subset $\mathbf{y}_t$ is recorded on *terminal* or non-parent animals who are not parents of any other animals with data. As with $\mathbf{y}_p$, $\mathbf{y}_t$ is modeled similarly as a function of $\boldsymbol{\beta}$, and $\mathbf{m}_p$ except that $t$ rather than $p$ is used as the subscript index for the respective design matrices in [1]. Furthermore, $\mathbf{y}_t$ is modeled as a linear function (through $\mathbf{Z}_{1t}\mathbf{P}_t$) of $\mathbf{a}_p$. Here $\mathbf{Z}_{1t}$ is a $n_t$ x $q_t$ design matrix and $\mathbf{P}_t$ is a $q_t \times q_p$ matrix connecting the genetic effects of $q_t$ non-parent animals to that of their parents. That is, in $\mathbf{P}_t$, row $j$, indexed $j = q_p$+1,$q_p$+2,...,$q$, connects the genetic effect of non-parent animal $j$ to that of its sire $s_j^*$ and dam $d_j^*$ such that the $j, s_j^*$ and $j, d_j^*$ elements of $\mathbf{P}_t$ for identified parents of animal $j$ are equal to 0.5. The "residual" vector is composed of error terms $\mathbf{e}_p$

and $\mathbf{e}_t$, respectively of parent and terminal animals, and additionally, for terminal animals, of additive Mendelian genetic sampling terms in the vector $\boldsymbol{\gamma}_t$, which is connected to $\mathbf{y}_t$ through $\mathbf{Z}_{1t}$.

We assume that the variance covariance matrix of the RAM residual vector is:

$$\mathbf{R} = \mathrm{var}\begin{bmatrix} \mathbf{e}_p \\ \mathbf{Z}_{1t}\boldsymbol{\gamma}_t + \mathbf{e}_t \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{q_p}\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{1t}\boldsymbol{\Omega}_{tt}\mathbf{Z}_{1t}'\sigma_a^2 + \mathbf{I}_{q_t}\sigma_e^2 \end{bmatrix},$$

where $\boldsymbol{\Omega}_{tt} = diag\left\{\omega_j\right\}_{j=q_p+1}^{q}$ is a $q_t \times q_t$ diagonal matrix, with the $j$th element corresponding to the proportion of the additive genetic variance ($\sigma_a^2$) on animal $j$ that is due to Mendelian sampling (Quaas, 1988); and $\sigma_e^2$ is the residual variance.

The structural prior specifications on the genetic effects are defined accordingly to include only parent terms; i.e.

$$p\left(\begin{bmatrix} \mathbf{a}_p \\ \mathbf{m}_p \end{bmatrix} | \mathbf{G}\right) = N\left(\mathbf{0}, \mathbf{G} \otimes \mathbf{A}_{pp}\right), \qquad [2]$$

where $\mathbf{G} = \begin{bmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{bmatrix}$ is the genetic variance-covariance matrix for direct and maternal genetic effects with $\sigma_m^2$ being the maternal genetic variance, and $\sigma_{am}$ the covariance between direct and maternal genetic effects. Furthermore, $\mathbf{A}_{pp}$ is the numerator relationship matrix amongst all $q_p$ parent animals and $\otimes$ is the Kronecker or direct product. For conjugate convenience, a joint bounded uniform or normal prior $p(\boldsymbol{\beta})$ may be specified for $\boldsymbol{\beta}$, an inverted Wishart prior density $p(\mathbf{G})$ specified for $\mathbf{G}$ and an inverted gamma density $p(\sigma_e^2)$ specified for $\sigma_e^2$.

In addition we have that

$$\begin{bmatrix}\mathbf{a}_t\\ \mathbf{m}_t\end{bmatrix}=[\mathbf{I}_2\otimes\mathbf{P}_t]\begin{bmatrix}\mathbf{a}_p\\ \mathbf{m}_p\end{bmatrix}+\begin{bmatrix}\boldsymbol{\gamma}_t\\ \boldsymbol{\delta}_t\end{bmatrix}. \qquad [3]$$

where $\mathbf{a}_t$ and $\mathbf{m}_t$ are respectively, the $q_t \times 1$ vectors of additive and maternal genetic effects associated with terminal animals. Furthermore, $\boldsymbol{\gamma}_t$ and $\boldsymbol{\delta}_t$ are each $q_t \times 1$ vectors of additive and maternal Mendelian genetic sampling terms, respectively, also associated with terminal animals and such that

$$\begin{bmatrix}\boldsymbol{\gamma}_t\\ \boldsymbol{\delta}_t\end{bmatrix}|\,\mathbf{G}\sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\mathbf{G}\otimes\boldsymbol{\Omega}_{tt}\right).$$

### 2.2. Modeling uncertain paternity

In populations undergoing multiple-sire mating, a number of males are possible candidate sires for each of several animals. This translates into uncertainty on various elements of $\mathbf{P}_t$ for non-parent animals and on various elements of $\mathbf{A}_{pp}$ for parent animals.

We first consider uncertain paternity on the $q_t$ non-parent animals indexed $j = q_p+1, q_p+2,\ldots,q$, and associated with $n_t$ records in $\mathbf{y}_t$. Let $\mathbf{Z}_1=\begin{bmatrix}\mathbf{Z}_{1p}\\ \mathbf{Z}_{1t}\mathbf{P}_t\end{bmatrix}$. Then if non-parent $j$ has uncertain paternity, this uncertainty translates into the $j, s_j^*$ element of $\mathbf{P}_t$ being unknown or, equivalently, the $s_j^*$ element of $\mathbf{z}'_{1ij}$ being unknown, where $\mathbf{z}'_{1ij}$ is the row of $\mathbf{Z}_1$ matching with the address of $y_{ij}$ in $\mathbf{y}$. Suppose, that for animal $j$, there are $v_j$ possible candidate sires with identifications listed in $\mathbf{s}_j=\left\{s_j^{(1)},s_j^{(2)},\ldots,s_j^{(v_j)}\right\}$. The distribution of $y_{ij}$, conditional on a given sire assignment $s_j^*=s_j^{(k)}$, $1\le k\le v_j$, on animal $j$ and all other parameters is given by:

$$y_{ij} \mid \boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, s_j^* = s_j^{(k)}, \sigma_a^2, \sigma_e^2 \sim N\left(\mathbf{x}'_{ij}\boldsymbol{\beta} + 0.5a_{s_j^{(k)}} + 0.5a_{d_j^*} + \mathbf{z}'_{2ij}\mathbf{m}_p, \sigma_e^2 + \omega_j^{(k)}\sigma_a^2\right),$$

$$i = n_p+1, n_p+2, \ldots, n; j = q_p+1, q_p+2, \ldots, q. \quad [4a]$$

Here $\mathbf{x}'_{ij}$, and $\mathbf{z}'_{2ij}$ are, respectively, the rows of $\mathbf{X}$ and $\mathbf{Z}_2$ matching the address of $y_{ij}$ in $\mathbf{y}$. When animal $j$ has certain paternity, $v_j = 1$ such that then $s_j^*$ is not random. Note that the conditioning on known $d_j^*$ (dam identification) is implied for all animals throughout this chapter whereas the conditioning on $s_j^* = s_j^{(k)}$ is explicitly provided given that $s_j^*$ may be uncertain. This uncertainty is further reflected in [4a] by the term $\omega_j^{(k)} = \omega_j\big|_{s_j^* = s_j^{(k)}}$ indicating that fraction $\omega_j^{(k)}$ of genetic variance attributable to Mendelian sampling for animal $j$ is a function of its inbreeding coefficient and hence of the sire assignment $s_j^* = s_j^{(k)}$.

Now consider the possibility that at least one of the parent animals, indexed from 1 to $q_p$, has uncertain paternity such that elements of $\mathbf{A}_{pp}$ are also uncertain. The sampling distribution of $y_{ij}$, on parent animal $j$, $j = 1,2,\ldots,q_p$, is not conditioned on uncertainty on sires, that is,

$$y_{ij} \mid \boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \sigma_e^2 \sim N\left(\mathbf{x}'_{ij}\boldsymbol{\beta} + a_j + \mathbf{z}'_{2ij}\mathbf{m}_p, \sigma_e^2\right),$$

$$i = 1,2,\ldots,n_p; \; j = 1,2,\ldots,q_p. \quad [4b]$$

Uncertain paternity on parent animals is modeled with the second stage structural prior on $\mathbf{a}_p$ and $\mathbf{m}_p$ in [2]. A useful decomposition of $\mathbf{A}_{pp}^{-1}$ as shown by Henderson (1976) and Quaas (1988) is $\mathbf{A}_{pp}^{-1} = \mathbf{T}_p\boldsymbol{\Omega}_{pp}^{-1}\mathbf{T}'_p$, where $\mathbf{T}_p$ is a $q_p$ × $q_p$ lower triangular matrix and

$\mathbf{\Omega}_{pp} = diag\left\{\omega_j\right\}_{j=1}^{q_p}$ is $q_p \times q_p$ diagonal matrix analogous to $\mathbf{\Omega}_{tt}$, but with elements corresponding to the fraction of $\sigma_a^2$ due to Mendelian sampling on each parent animal *j*. All of the diagonal elements of $\mathbf{T}_p$ are equal to 1 with at most two other elements per row, say $j, s_j^*$ and $j, d_j^*$ , being equal to -0.5, if the corresponding parents $s_j^*$ and $d_j^*$ of animal *j* are identified, for $j = 1,2,\ldots,q_p$. Consequently, $\left|\mathbf{A}_{pp}^{-1}\right| = \left|\mathbf{T}_p\right|\left|\mathbf{\Omega}_{pp}^{-1}\right|\left|\mathbf{T}_p'\right| = \left|\mathbf{\Omega}_{pp}^{-1}\right|$ since $\left|\mathbf{T}_p\right| = 1$. Given this result, the joint prior density of $\mathbf{a}_p$ and $\mathbf{m}_p$ conditioned on $\mathbf{A}_{pp}$, can be written as,

$$p\left(\begin{matrix}\mathbf{a}_p \\ \mathbf{m}_p\end{matrix}\middle|\mathbf{G}, \mathbf{A}_{pp}\right) \propto \left|\mathbf{G}\right|^{-\frac{p}{2}}\left|\mathbf{\Omega}_{pp}^{-1}\right|$$
$$\times \exp\left(-0.5\left(\mathbf{a}_p'\mathbf{T}_p\mathbf{\Omega}_{pp}^{-1}\mathbf{T}_p'\mathbf{a}_p g^{11} + 2\mathbf{a}_p'\mathbf{T}_p\mathbf{\Omega}_{pp}^{-1}\mathbf{T}_p'\mathbf{m}_p g^{12} + \mathbf{m}_p'\mathbf{T}_p\mathbf{\Omega}_{pp}^{-1}\mathbf{T}_p'\mathbf{m}_p g^{22}\right)\right), \qquad [5]$$

where $g^{ij}$ is the $(i,j)$th element of $\mathbf{G}^{-1}$ for $i,j = 1,2$.

Let $\mathbf{t}_j'$ denote the *j*th row of $\mathbf{T}_p$. Then it can be readily shown that the additive and maternal Mendelian sampling terms are respectively $\gamma_j = \mathbf{t}_j'\mathbf{a}_p = a_j - .5a_{s_j^*} - .5a_{d_j^*}$ and $\delta_j = \mathbf{t}_j'\mathbf{m}_p = m_j - .5m_{s_j^*} - .5m_{d_j^*}$ for $j = 1,\ldots,q_p$. If there are no known candidates for $s_j^*$ and $d_j^*$ then the corresponding parental contributions of $a_{s_j^*}$ and $a_{d_j^*}$ to $\gamma_j$ and $m_{s_j^*}$ and $m_{d_j^*}$ to $\delta_j$ are equal to 0, as would be true for each of the base population animals $j=1,2,\ldots,q_b \leq q_p$. Let $\mathbf{s}_p^* = \left\{s_j^*\right\}_{j=1}^{q_p}$ denote the vector of random sire assignments on parent animals and $\mathbf{s}_p^{(\mathbf{k})} = \left\{s_j^{(k)}\right\}_{j=1}^{q_p}$ be a particular realization of $\mathbf{s}_p^*$ from the set

$\mathbf{S}_p = \left\{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \ldots, \mathbf{s}_{q_p}\right\}$ such that the *j*th element of $\mathbf{s}_p^{(\mathbf{k})}$ is one of the $v_j$ elements chosen from $\mathbf{s}_j = \left\{s_j^{(1)}, s_j^{(2)}, \ldots.., s_j^{(v_j)}\right\}$ for $j = 1,2,\ldots,q_p$. Note that for the $q_b$ base animals, $\mathbf{s}_j$ is an empty set. We can then rewrite [5], explicitly conditioning on sire assignments as follows:

$$p\left(\begin{matrix}\mathbf{a}_p \\ \mathbf{m}_p\end{matrix}\middle|\mathbf{G}, \mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})}\right) \propto \left|\mathbf{G}\right|^{-\frac{q_p}{2}}$$

$$\times\prod_{j=1}^{q_p}\left(\left(\omega_j^{(k)}\right)^{-1}\exp\left(-0.5\left(\omega_j^{(k)}\right)^{-1}\left(\left(\gamma_j^{(k)}\right)^2 g^{11} + \left(\delta_j^{(k)}\right)^2 g^{22} + 2\gamma_j^{(k)}\delta_j^{(k)}g^{12}\right)\right)\right), \qquad [6]$$

where $\delta_j^{(k)} = \delta_j\big|_{s_j^* = s_j^{(k)}}$ and $\gamma_j^{(k)} = \gamma_j\big|_{s_j^* = s_j^{(k)}}$, indicating the natural dependence of Mendelian sampling terms on the sire assignment $s_j^* = s_j^{(k)}$. As there is no need to infer upon uncertain paternity for the $q_b$ base animals, $\omega_j^{(k)} = 1$ for $j = 1,2,\ldots,q_b$ with $\left\{s_j^{(k)}\right\}_{j=1}^{q_b}$ being an empty subset of $\mathbf{s}_p^{(\mathbf{k})}$.

The third stage of the model specifies the prior probability for each of $v_j$ males being the correct sire of an animal *j*. As we do similarly for parents, we let $\mathbf{s}_t^* = \left\{s_j^*\right\}_{j=q_p+1}^{q}$ denote the vector of random sire assignments on the non-parent animals and $\mathbf{s}_t^{(\mathbf{k})} = \left\{s_j^{(k)}\right\}_{j=q_p+1}^{q}$ denote a particular realization of $\mathbf{s}_t^*$ from the set $\mathbf{S}_t = \left\{\mathbf{s}_{q_p+1}, \mathbf{s}_{q_p+2}, \ldots, \mathbf{s}_q\right\}$. For all $q$ animals, we then write $\mathbf{s}^{(\mathbf{k})} = \begin{bmatrix}\mathbf{s}_p^{(\mathbf{k})} \\ \mathbf{s}_t^{(\mathbf{k})}\end{bmatrix} = \left\{s_j^{(k)}\right\}_{j=1}^{q}$ as being a realization of $\mathbf{s}^* = \begin{bmatrix} \mathbf{s}_p^* \\ \mathbf{s}_t^* \end{bmatrix} = \left\{s_j^*\right\}_{j=1}^q$ from the set $\mathbf{S} = \left\{\mathbf{S}_p, \mathbf{S}_t\right\}$. The probability that $s_j^{(k)}$ is the sire of animal *j* is defined as $\pi_j^{(k)} = \text{Prob}\left(s_j^* = s_j^{(k)}\right)$ for $k = 1,2,...,v_j$ such that $\sum_{k=1}^{v_j} \pi_j^{(k)} = 1$. For animals with certain paternity, there is only one candidate $s_j^* \equiv s_j^{(1)}$ such that $\pi_j^{(1)} = 1$ and hence is constant. For each of the $q_b$ base animals, $\pi_j^{(k)}$ is not specified since there are no candidate sires. The set of probabilities $\boldsymbol{\pi}_j = \left\{\pi_j^{(1)}, \pi_j^{(2)}, ..., \pi_j^{(v_j)}\right\}$ for each one of $v_j$ candidate sires for non-base animal $j$ ($j = q_b+1, q_b+2, \ldots, q$) may be conceptually elicited using external information (e.g. genetic markers). The entire set of probabilities $\boldsymbol{\pi} = \left\{\boldsymbol{\pi}_j\right\}_{q_b+1}^q$ is rarely known with absolute certainty, and so we might regard them as random quantities from a Dirichlet distribution:

$$p\left(\boldsymbol{\pi}_j \mid \boldsymbol{\alpha}_j\right) \propto \prod_{k=1}^{v_j} \left(\pi_j^{(k)}\right)^{\alpha_j^{(k)}} \qquad [7]$$

where $\boldsymbol{\alpha}_j = \left\{\alpha_j^{(k)}\right\}_{k=1}^{v_j}$, $\alpha_j^{(k)} > 0$ for $k = 1,2,\ldots,v_j$ and $\pi_j^{(v_j)} = 1 - \sum_{k=1}^{v_j-1} \pi_j^{(k)}$ is constrained accordingly. Specification of the set of hyper-parameters $\boldsymbol{\alpha} = \left\{\boldsymbol{\alpha}_j\right\}_{j=q_b+1}^q$ might be based on the assessed reliability of the source of external information on the prior probability of each sire assignment.

We use [4a], [4b], and [6] as key expressions to determine the joint posterior density of all unknown parameters

$$p\left(\boldsymbol{\beta},\mathbf{a}_p,\mathbf{m}_p,\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})},\boldsymbol{\pi},\mathbf{G},\sigma_e^2 \mid \mathbf{y}\right) \propto$$

$$\times\prod_{i=1}^{n_p} p\left(y_{ij} \mid \boldsymbol{\beta},\mathbf{a}_p,\mathbf{m}_p,\sigma_e^2\right) \prod_{i=n_p+1}^{n} p\left(y_{ij} \mid \boldsymbol{\beta},\mathbf{a}_p,\mathbf{m}_p,\mathbf{s}_t^* = \mathbf{s}_t^{(\mathbf{k})},\sigma_a^2,\sigma_e^2\right)$$

$$\times p\left(\mathbf{a}_p,\mathbf{m}_p \mid \mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})},\mathbf{G}\right) p(\boldsymbol{\beta}) \mathrm{Prob}\left(\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})} \mid \boldsymbol{\pi}\right) p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) p(\mathbf{G}) p\left(\sigma_e^2\right) \qquad [8]$$

Here,

$$\mathrm{Prob}\left(\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})} \mid \boldsymbol{\pi}\right) = \prod_{j=q_b+1}^{q} \mathrm{Prob}\left(s_j^* = s_j^{(k)} \mid \boldsymbol{\pi}_j\right) = \prod_{j=q_b+1}^{q} \prod_{k=1}^{v_j} \left(\pi_j^{(k)}\right)^{I_j^{(k)}},$$

where $I_j^{(k)} = 1$ if $s_j^* = s_j^{(k)}$ and $I_j^{(k)} = 0$ otherwise. Furthermore,

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \prod_{j=q_b+1}^{q} p\left(\boldsymbol{\pi}_j \mid \boldsymbol{\alpha}_j\right) = \prod_{j=q_b+1}^{q} \prod_{k=1}^{v_j} \left(\pi_j^{(k)}\right)^{\alpha_j^{(k)}}.$$

The fully conditional distributions (FCD) of all unknown parameters/quantities or blocks thereof in [8] necessary to conduct MCMC inference with some details on the sampling strategy itself are derived in the Appendix to this chapter. A good exposition on MCMC implementations in hierarchical animal breeding models analogous to that presented in this chapter is provided by Wang et al. (1994b).

### 3. Simulation study

A simulation study was carried out to compare two models for the prediction of genetic merit allowing for uncertain paternity on some animals. The first model is the hierarchical model proposed in this chapter (Section 2), which infers upon this uncertainty using phenotypic data; the other model is based on the use of the Henderson's average numerator relationship.

Ten datasets were generated for each of two different types of traits. Trait 1 had medium direct heritability ($h_a^2 = 0.3$), medium maternal heritability ($h_m^2 = 0.2$) and a slightly negative direct-maternal correlation ($r_{am} = -0.2$) as, for example, would characterize weaning weight. Trait 2 had a high direct heritability ($h_a^2 = 0.5$), but null $h_m^2$ as would characterize post-weaning gain. The residual variance ($\sigma_e^2$) was 60 and 50, respectively for Traits 1 and 2.

Each population included 80 sires, 400 dams (480 parents) and 2 000 non-parent animals, all of which descended from 20 base sires and 100 base dams. From these base animals, five generations were created. Fifteen males and 75 females were randomly selected from each generation to be parents of the next generation. Furthermore, five sires and 25 dams from the previous generation's breedstock were retained, such that a total of 20 sires and 100 dams were used as the breeding group for each generation. That is, the population was structured to have overlapping generations. The probability of any offspring being assigned to an uncertain paternity situation was 0.3. If an animal had uncertain paternity, it was randomly assigned to one of six possible multiple-sire groups in each of the five generations. These groups had six different sizes: $v_j$ = 2, 3, 4, 6, 8 or 10 candidate sires. Once the group was chosen, one of the males in the group was selected to be the true sire with either equal ($1/v_j$) or unequal probability relative to the rest of the candidate sires (the actual probabilities used to assign progeny to sires in each group can be obtained from the author by request). The latter scenario was intended to represent the dominant male situation, common in beef cattle (DeNise, 1999). The five sires selected from the previous generation's breedstock had only certain progeny. An additional ten sires were used in group matings but also had certain progeny, whereas the remaining five sires had only uncertain progeny. One group of three sires in each population was formed with sires having only uncertain progeny with the purpose of comparing the performance of the two models in the case where sires have only their own record and pedigree as the only source of information for their genetic evaluation, other than uncertain progeny. All other mating groups had at least one sire that was known to be sires of other animals. We deliberately intended to mimic the situation observed in some ranches under genetic evaluation in Brazil. These ranches select their own young bulls to serve their herd by natural service (NS) and also collect semen from their own top bulls to be used in artificial insemination (AI). Moreover, they import external genetics especially through AI. In this scenario, the sires can be categorized in three different ways: 1) sires having only known progeny (i.e. imported AI bulls); 2) sires having both known and uncertain progeny assignments, such as top herd bulls that are used by AI or known NS mating but also by uncertain NS in multiple sire pastures during the breeding season and 3) sires having only uncertain progeny assignments.

Only one record was generated per each animal. For both traits, the overall mean was equal to 100 and a fixed effects factor with three levels, having values 25, -25 and 0, was randomly assigned to generate the individual records.

The ten replicates for each of the two traits were analyzed using three different models:

1) HIER: A hierarchical mixed effects model fully accounting for uncertainty on sire assignments as proposed in Section 2.

2) ANRM: A linear mixed effects model based on the average numerator relationship matrix (Henderson, 1988). Equal and fixed probabilities were assigned to each candidate sire of animals pertaining to uncertain paternity.

3) TRUE: A linear mixed effects model based on the true sire assignments, as if there was no uncertainty on assignments. This model was included to serve as a positive control for the other two models.

For all three models, a MCMC sampling chain of $G = 20\ 000$ cycles was run after a burn-in period of 4 000 cycles. In order to concentrate our attention on the relative performance of the models for breeding value prediction, variance components were considered to be known. Flat bounded priors were placed on each fixed effect. Naïve equal prior probabilities, i.e. inverse of the number of candidate sires within each group, were specified on each sire assignment to an animal. By setting $\alpha_j^{(k)} = \frac{1}{v_j}$ for $k = 1,2,\ldots,v_j$, we have that $\sum_{k=1}^{v_j} \alpha_j^{(k)} = 1$, and the same weight is statistically given to prior and data information in the sampling of sire assignments for the $j$th animal in the set of animals with uncertain paternity.

The parameters used to compare the methods studied were the mean squared error of prediction (MSEP), the mean bias of prediction (MBIAS) and Spearman rank correlations between estimated and true genetic values. The MSEP for each model was estimated as $\sum_{h=1}^{10}\sum_{j=1}^{q}\left(\hat{u}_{hj} - u_{hj}\right)^2 \Big/ q \Big/ 10$, where 10 denotes the number of replicates, $q$ is the total number of parent or non-parent animals with uncertain paternity per replicate, $\hat{u}_{hj}$ is the estimated genetic additive or maternal effect for animal $j$ in replicate $h$ and $u_{hj}$ is the true genetic additive or maternal effect for animal $j$ in replicate $h$. MBIAS was similarly estimated as $\sum_{h=1}^{10}\sum_{j=1}^{q}\left(\hat{u}_{hj}-u_{hj}\right)/q\Big/10$.

Variables describing uncertain paternity, specifically, $s_j^*$ and $\pi_j^{(k)}$, were analyzed separately for parents and non-parents, as parents were considered to have greater amounts of information on their genetic merit compared to non-parents. Sires had on average 23.6 progeny, while dams averaged 5.9 progeny. Within each group size category, animals with certain paternity and with uncertain paternity were considered separately. Pairwise comparisons based on genetic merits estimated under the three different models were performed using a $t$-test.

We also considered two model choice criteria: the *Pseudo Bayes Factor* (PBF) (Gelfand, 1996) and the *Deviance Information Criterion* (DIC) (Spiegelhalter et al., 2002). For comparing, say, models $M_1$ and $M_2$, the corresponding PBF was determined to be:

$$PBF_{1,2}=\prod_{i=1}^{n}\frac{p\left(y_{ij}\mid\mathbf{y}_{(-ij)},M_1\right)}{p\left(y_{ij}\mid\mathbf{y}_{(-ij)},M_2\right)},$$

where $p\left(y_{ij}\mid\mathbf{y}_{(-ij)},M_r\right)$ is the conditional predictive ordinate (CPO) for observation $y_{ij}$, intended to be a cross-validation density, which suggests what values of $y_{ij}$ are likely when Model $M_r$ is fit to all other observations $\mathbf{y}_{(-ij)}$ except $y_{ij}$. A MCMC approximation for the CPO of Model $M_r$ with parameters $\boldsymbol{\theta}$ is obtained by a harmonic mean of the $G$ MCMC cycles

$$p\left(y_{ij} \mid \mathbf{y}_{(-ij)}, M_r\right) \approx \frac{1}{\frac{1}{G}\sum_{l=1}^{G} p^{-1}\left(y_{ij} \mid \boldsymbol{\theta}^{(l)}, M_r\right)}.$$

The DIC is composed by a measure of global fit, posterior mean of the deviance, and a penalization for complexity of the model. The deviance for Model $M_r$ using the null standardization from Spiegelhalter et al. (2002) can be estimated by $\bar{D}_r = \frac{1}{G}\sum_{l=1}^{G} -2\log p\left(\mathbf{y} \mid \boldsymbol{\theta}^{(l)}, M_r\right)$. The 'complexity' of Model $M_r$ is determined as the effective number of parameters given by $p_{D(r)} = \bar{D}_r - D_r\left(\bar{\boldsymbol{\theta}}\right)$ where $D_r\left(\bar{\boldsymbol{\theta}}\right) = -2\log p\left(\mathbf{y} \mid \bar{\boldsymbol{\theta}}, M_r\right)$ with $\bar{\boldsymbol{\theta}}$ being the posterior mean of $\boldsymbol{\theta}$. That is, $p_{D(r)}$ represents the difference between the posterior mean of the deviance and the deviance based on the posterior mean of the parameters under Model $M_r$. The DIC for Model $M_r$ is then determined as:

$$DIC_r = \bar{D}_r + p_{D(r)}.$$

Smaller values of DIC are indicative of a better-fitting model.

### 4. Results

Since it was unclear to us whether the indicator variable $s_j^*$ or parameter $\pi_j^{(k)}$ should be used for inferring uncertainty with respect to assignment of sire $k$ to animal $j$, we considered both variables. Inference on the probabilities of the true sires for animals with uncertain paternity in the HIER model was based on determining the frequency of MCMC samples of $s_j^*$ that were equal to the true sire, designated as $\text{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$, and by determining $\text{E}\left(\pi_j^{(true)} \mid \mathbf{y}\right)$ the posterior mean of $\pi_j^{(true)}$, the probability parameter identified with $s_j^{(true)}$, the true sire of $j$. These summaries are presented separately for parent and non-parent animals with uncertain paternity in Table 1.1 for both Traits 1 and 2. The average posterior probabilities attributed to the true sire (i.e. based on $\text{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$) were between 1 and 10% larger than the respective priors ($1/v_j$ for a respective mating group of size $v_j$) for Trait 1 and between 4 and 13% larger than the priors for Trait 2. Inference on uncertain paternity using $\text{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$ had a slightly better general performance than inference based on $\text{E}\left(\pi_j^{(true)} \mid \mathbf{y}\right)$. The larger differences between average posterior and prior probabilities in Trait 2 may be a result of the higher heritability. These differences were generally statistically significant ($P<.05$), based on one-sample *t* tests.

The consistently higher probability attributed to $s_j^{(true)}$ by HIER indicates that this model tends to infer towards the correct sire; however, the small magnitude of these differences suggests that phenotypes may not be sufficiently informative to precisely infer upon paternity assignments under these two trait scenarios. The average $\text{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$ for mating groups of size $v_j$ = 3 and formed with sires with exclusively uncertain progeny were 0.348 for Trait 1 and 0.360 for Trait 2. These probabilities were consistent with those determined for other groups of size $v_j$ = 3 but including sires that had also certain progeny. That is, the HIER model performed similarly in terms of probabilities of assignments to sires whether or not sires have both certain and uncertain progeny or only uncertain progeny as source of information.

Table 1.1. Posterior means of probabilities of sires being true sires $\left(\mathrm{E}\left(\pi_j^{(true)} \mid \mathbf{y}\right)\right)$ and probability of sire assignments being equal to true sires $\left(\mathrm{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)\right)$ averaged across sires and replicates for Traits 1 and 2 by multiple-sire group size and parents versus non-parent animals.

| Parameter | Animal Category | Multiple-sire group size | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 6 | 8 | 10 |
| Trait 1 | | | | | | | |
| $\mathrm{E}\left(\pi_j^{(true)} \mid \mathbf{y}\right)$ | Parents | 0.513 | 0.341 | 0.259 | 0.175 | 0.126[a] | 0.105 |
| $\mathrm{E}\left(\pi_j^{(true)} \mid \mathbf{y}\right)$ | Non-parents | 0.509 | 0.339 | 0.259 | 0.172 | 0.130 | 0.103 |
| $\mathrm{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$ | Parents | 0.525 | 0.349 | 0.269 | 0.183 | 0.127 | 0.110 |
| $\mathrm{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$ | Non-parents | 0.517 | 0.345 | 0.268 | 0.178 | 0.134 | 0.105 |
| Trait 2 | | | | | | | |
| $\mathrm{E}\left(\pi_j^{(true)} \mid \mathbf{y}\right)$ | Parents | 0.510 | 0.343 | 0.265 | 0.177 | 0.132 | 0.105 |
| $\mathrm{E}\left(\pi_j^{(true)} \mid \mathbf{y}\right)$ | Non-parents | 0.520 | 0.346 | 0.270 | 0.179 | 0.134 | 0.106 |
| $\mathrm{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$ | Parents | 0.521 | 0.352 | 0.280 | 0.188 | 0.138 | 0.111 |
| $\mathrm{Prob}\left(s_j^* = s_j^{(true)} \mid \mathbf{y}\right)$ | Non-parents | 0.540 | 0.360 | 0.289 | 0.191 | 0.143 | 0.111 |

[a]Posterior probability is not statistically different from the prior of its group size at $\alpha$=.05

In terms of MBIAS, none of the three models were significantly different from any other under all situations analyzed, and the results are not presented here. The mean squared error of prediction (MSEP) and rank correlation on additive and maternal genetic effects of parents and non-parents, with uncertain paternity for Trait 1 (medium $h_a^2$ - additive and maternal effects) are presented in Figures 1.1 and 1.2, respectively. As

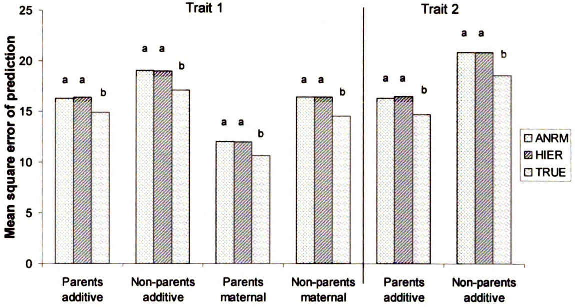

Figure 1.1. Mean squared error of prediction (MSEP) of posterior means of additive and maternal genetic effects of parent and non-parent animals with uncertain paternity for Traits 1 and 2 under three models, 1) HIER based on proposed hierarchical model, 2) ANRM based on Henderson's average numerator relationship matrix, and 3) TRUE based on knowledge of the true sire as a positive control. Within each group, bars sharing the same letter are not statistically different at $\alpha$=.05

expected, the MSEP was always smaller and rank correlation higher for TRUE compared to ANRM and HIER, showing that the use of multiple-sire matings adversely affects accuracy of genetic evaluations (Sullivan, 1995). Posterior means of additive and maternal genetic effects were very similar for HIER and ANRM with no significant difference in MSEP and rank correlations on these posterior means between these models. There was, however, a tendency for smaller MSEP and higher rank correlation under HIER for animals with uncertain paternity. There seems to be not enough information, at least in this simulated scenario, to discriminate between ANRM and HIER for MSEP and rank correlation of genetic evaluations using only phenotypic

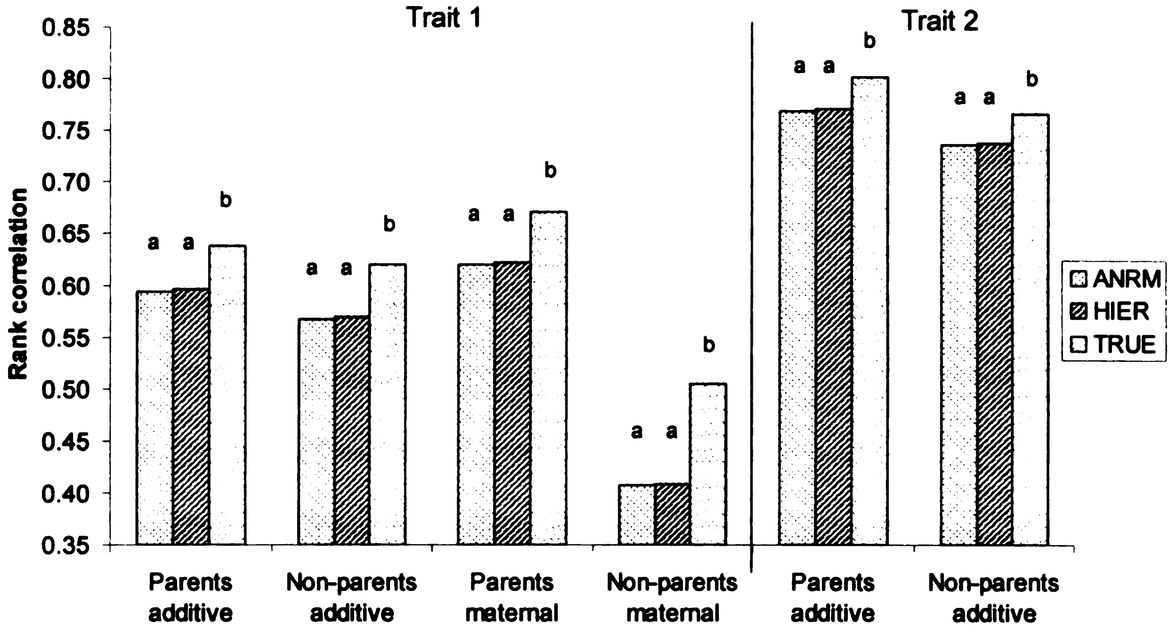

Figure 1.2. Rank correlation of additive and maternal genetic effects of parent and non-parent animals with uncertain paternity for Traits 1 and 2 under three models, 1) HIER based on proposed hierarchical model, 2) ANRM based on Henderson's average numerator relationship matrix, and 3) TRUE based on knowledge of the true sire as a positive control. Within each group, bars sharing the same letter are not statistically different at α=.05.

records. This result may be associated with the small differences between prior and posterior probabilities of sire assignments under HIER.

For Trait 2, the MSEP and rank correlation were also not statistically different between ANRM and HIER across the ten simulated datasets (Figures 1.1 and 1.2). Here, the differences in terms of rank correlation among models were somewhat smaller relative to Trait 1. This result may be due to the higher $h^2$, and therefore the decreased importance of pedigree information, i.e. sire assignments, relative to phenotypes for prediction of genetic effects.

We applied two model choice criteria, the PBF and DIC as previously described, to compare the statistical fit of the two models, ANRM and HIER. The PBF for all replicates were always favorable for HIER compared to ANRM, with magnitudes ranging from $2.1\times10^{2}$ to $2.4\times10^{7}$ for Trait 1, and from $6.3\times10^{7}$ to $2.6\times10^{24}$ for Trait 2. The calculated DIC were also always in favor of HIER compared to ANRM ranging from a differences of 9 to 41 for Trait 1 and from 33 to 115 for Trait 2. These results appear to be decisively in favor of the HIER model as Spiegelhalter et al. (2002) has suggested a DIC difference of 7 to be an important difference in model fit. For Trait 1, the average DIC over the ten replicates was 17 843 for HIER ($\bar{D}_{HIER}=17\ 135$ and $p_{D(HIER)}=709$) and 17 866 for ANRM ($\bar{D}_{ANRM}=17\ 164$ and $p_{D(ANRM)}=702$); and for Trait 2 we obtained an average DIC of 17 553 for HIER ($\bar{D}_{HIER}=16\ 605$ and $p_{D(HIER)}=949$) and of 17 630 for ANRM ($\bar{D}_{ANRM}=16\ 704$ and $p_{D(ANRM)}=926$). The primary reason for a smaller DIC for HIER compared to ANRM was the smaller mean deviance ($\bar{D}_r$) of HIER. The difference in terms of $\bar{D}_r$ was large enough to compensate the penalty for a larger effective number of parameters ($p_{D(r)}$) applied to HIER. These two model choice criteria (PBF and DIC) clearly indicate that the HIER model provides a better statistical fit than the ANRM model to the simulated data involving animals with uncertain paternity.

## 5. Discussion

We proposed in this study a fully Bayesian approach for prediction of genetic merit of animals having uncertain paternity. Similar to the empirical Bayes sire model method of Foulley et al. (1987), our procedure combines data and prior information to determine posterior probabilities of sire assignments. Nevertheless, our method represents an important extension since it uses more recently developed MCMC tools to provide small sample inference based on the animal model, the most common model for current genetic evaluations. Our method can be readily extended to multiple-trait or other quantitative genetic (e.g. random regression) models without great conceptual difficulty. It could also be easily generalized to the case of uncertain dams; however, this is not a typical scenario in livestock breeding.

The results obtained from our simulation study indicate that a model accounting for uncertainty on sire assignments provides a better fit to data characterized by uncertain paternity relative to a model based on the use of the average numerator relationship matrix (Henderson, 1988). The relative performance between the two models might be expected to increase with $h^2$ since the power of discriminating between candidate sires should intuitively increase. We previously have shown that when $h^2 = 0.10$, there was no significant difference between prior and posterior probabilities of sire assignments (Cardoso and Tempelman, 2001). However, the lower the $h^2$, the greater the importance of data on uncertain progeny in the prediction of a sire's genetic merit (Sullivan, 1995). The difference between the two models, nevertheless, does not necessarily increase with higher heritabilities, since the importance of pedigree information relative to phenotypic information decreases with respect to the prediction of genetic merit. Thus our work suggests the largest differences in performance between the two models exist for traits with intermediate $h^2$. Nonetheless, due to similarity in terms of rank correlation, and especially in the absence of prior information from e.g. genetic markers, the ANRM model may be preferable for genetic evaluation of large populations given the potential savings in computational time.

In the presence of prior information on sire assignments, the hierarchical model presented in this study represents an important alternative for genetic prediction. That is, in addition to the incorporation of prior probabilities on sire assignments, as also possible with ANRM, the HIER model allows for the integration of the uncertainty about these prior probabilities in the prediction of genetic merit. Genetic markers, for example, represent an important objective source of prior information. Moreover, the HIER model represents a general framework which could be extended to model the quality of genetic marker information contributing to sire assignment (Rosa et al., 2002).

The use of multiple-sire mating is common in large beef cattle populations raised in pastoral conditions. Currently, about 25-30% of the calves evaluated by the beef cattle improvement programs in Brazil derive from multiple-sire mating. Multiple sire matings are used to improve pregnancy rates, since the average size of breeding groups, as a function of paddock size, is too large to be sired by a single bull. Other examples of uncertain parentage include the use of AI followed by NS, accidental or unplanned breedings, and AI with pooled semen as is common in swine production. Multiple-sire matings are also commonly found in some sheep production systems.

The impact of modeling uncertain paternity, either through ANRM or HIER, is expected to be particularly important for large herds. These herds provide sizable gene pools for selection, thereby offering great potential for genetic improvement programs; however, the exclusive use of single matings is costly and generally impractical in these operations due to their size and labor commitments. Genetic evaluation systems that model uncertain paternity will aid genetic improvement of economically important traits in large populations raised in pastoral conditions and undergoing multiple-sire mating.

## Appendix

### *Specification of fully conditional distributions*

Let $\boldsymbol{\theta} = \left[\boldsymbol{\beta}', \mathbf{a}'_p, \mathbf{m}'_p\right]'$; $\mathbf{W}_t^{(\mathbf{k})} = \left[\mathbf{X} \quad \mathbf{Z}_1\big|_{\mathbf{s}_t^* = \mathbf{s}_t^{(\mathbf{k})}} \quad \mathbf{Z}_2\right]$, with $\mathbf{Z}_1\big|_{\mathbf{s}_t^* = \mathbf{s}_t^{(\mathbf{k})}}$ indicating the dependency of this design matrix on sire assignments $\mathbf{s}_t^* = \mathbf{s}_t^{(\mathbf{k})}$ for non-parent animals; and $\left(\boldsymbol{\Sigma}_p^{(\mathbf{k})}\right)^- = \begin{bmatrix} \mathbf{V}_{\boldsymbol{\beta\beta}}^{-1} & \mathbf{0}_{p \times 2q_p} \\ \mathbf{0}_{2q_p \times p} & \mathbf{G}^{-1} \otimes \mathbf{A}_{pp}^{-1}\big|_{\mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})}} \end{bmatrix}$, with $\mathbf{A}_{pp}^{-1}\big|_{\mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})}}$ indicating the dependence of parental relationships on sire assignments $\mathbf{s}_p^* = \mathbf{s}_p^{(\mathbf{k})}$ for parent animals, and $\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}$ being an $p \times p$ diagonal matrix consistent with a $N\left(\boldsymbol{\beta}_o, \mathbf{V}_{\boldsymbol{\beta\beta}}\right)$ prior assignment on $\boldsymbol{\beta}$. If $\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1} = \mathbf{0}_{p \times p}$, then $p(\boldsymbol{\beta}) \propto 1$. We, however, adopted a proper bounded uniform prior on $\boldsymbol{\beta}$, which is equivalent to specifying $\mathbf{V}_{\boldsymbol{\beta\beta}}^{-1} = \mathbf{0}_{p \times p}$ but with values of $\boldsymbol{\beta}$ constrained to be within the specified bounds. Then, it can be readily shown using results from Wang et al. (1994b) that the FCD of $\boldsymbol{\theta}$ is multivariate normal, that is,

$$\boldsymbol{\theta} \mid \mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}, \mathbf{G}, \sigma_e^2, \mathbf{y} \sim N(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{C}^{(\mathbf{k})}) \qquad \text{[A1]}$$

where

$$\hat{\boldsymbol{\theta}}^{(\mathbf{k})} = \mathbf{C}^{(\mathbf{k})}\left(\mathbf{W}_t^{(\mathbf{k})}{}'\left(\mathbf{R}_t^{(\mathbf{k})}\right)^{-1}\mathbf{y} + \begin{bmatrix} \mathbf{V}_{\boldsymbol{\beta\beta}}^{-1}\boldsymbol{\beta}_o \\ \mathbf{0}_{2q_p x1} \end{bmatrix}\right)$$

for $\mathbf{C}^{(\mathbf{k})}=\left(\mathbf{W}_{t}^{(\mathbf{k})\prime}\left(\mathbf{R}_{t}^{(\mathbf{k})}\right)^{-1}\mathbf{W}_{t}^{(\mathbf{k})}+\left(\boldsymbol{\Sigma}_{p}^{(\mathbf{k})}\right)^{-}\right)^{-1}$, with $\mathbf{R}_{t}^{(\mathbf{k})}=\mathbf{R}\big|_{\mathbf{s}_{t}^{*}=\mathbf{s}_{t}^{(\mathbf{k})}}$ indicating the dependency of **R** on sire assignments $\mathbf{s}_{t}^{*}=\mathbf{s}_{t}^{(\mathbf{k})}$ on non-parents.

The FCD of sire assignments in s$^{*}$ are considered separately for parents and non-parent animals. For parent animals, the FCD of the sire assignment on animal *j* is:

$$\operatorname{Prob}\left(s_{j}^{*}=s_{j}^{(k)}\mid\boldsymbol{\beta},\mathbf{a}_{p},\mathbf{m}_{p},\mathbf{s}_{-j}^{*}=\mathbf{s}_{-j}^{(\mathbf{k})},\boldsymbol{\pi},\mathbf{G},\sigma_{e}^{2},\mathbf{y}\right)$$

$$=\frac{\pi_{j}^{(k)}\left(\omega_{j}^{(k)}\right)^{-1}\exp\left(-.5\left(\omega_{j}^{(k)}\right)^{-1}\left(\left(\gamma_{j}^{(k)}\right)^{2}g^{11}+\left(\delta_{j}^{(k)}\right)^{2}g^{22}+2\left(\gamma_{j}^{(k)}\right)\left(\delta_{j}^{(k)}\right)g^{12}\right)\right)}{\sum_{k=1}^{v_{j}}\pi_{j}^{(k)}\left(\omega_{j}^{(k)}\right)^{-1}\exp\left(-.5\left(\omega_{j}^{(k)}\right)^{-1}\left(\left(\gamma_{j}^{(k)}\right)^{2}g^{11}+\left(\delta_{j}^{(k)}\right)^{2}g^{22}+2\left(\gamma_{j}^{(k)}\right)\left(\delta_{j}^{(k)}\right)g^{12}\right)\right)},$$

$$j=q_{b}+1,q_{b}+2,\ldots,q,\qquad\text{[A2]}$$

where $\mathbf{s}_{-j}^{*}=\mathbf{s}_{-j}^{(\mathbf{k})}$ is used to denote the conditioning on sire assignments for all animals other than *j*. For *non-parent* animals, the FCD of the sire assignment on animal *j* is:

$$\operatorname{Prob}\left(s_{j}^{*}=s_{j}^{(k)}\mid\boldsymbol{\beta},\mathbf{a}_{p},\mathbf{m}_{p},\mathbf{s}_{-j}^{*}=\mathbf{s}_{-j}^{(\mathbf{k})},\boldsymbol{\pi},\mathbf{G},\sigma_{e}^{2},\mathbf{y}\right)$$

$$=\frac{\pi_{j}^{(k)}\left(\sigma_{e}^{2}+\omega_{j}^{(k)}\sigma_{a}^{2}\right)^{-1/2}\exp\left(-.5\left(\sigma_{e}^{2}+\omega_{j}^{(k)}\sigma_{a}^{2}\right)^{-1}\left(e_{ij}^{(k)}\right)^{2}\right)}{\sum_{k=1}^{v_{j}}\pi_{j}^{(k)}\left(\sigma_{e}^{2}+\omega_{j}^{(k)}\sigma_{a}^{2}\right)^{-1/2}\exp\left(-.5\left(\sigma_{e}^{2}+\omega_{j}^{(k)}\sigma_{a}^{2}\right)^{-1}\left(e_{ij}^{(k)}\right)^{2}\right)},\qquad\text{[A3]}$$

where $e_{ij}^{(k)}=y_{ij}-\mathbf{x}_{ij}'\boldsymbol{\beta}-0.5a_{s_{j}^{(k)}}-0.5a_{d_{j}^{*}}-\mathbf{z}_{2ij}'\mathbf{m}_{p}$ and $j=q_{p}+1,q_{p}+2,\ldots,q$. Therefore, MCMC inference on sire assignments require random draws from generalized Bernoulli (i.e. single trial multinomial) distributions.

The FCD's for the probabilities of sire assignments are given by:

$$p\left(\boldsymbol{\pi}_{j}\mid\boldsymbol{\beta},\mathbf{a}_{p},\mathbf{m}_{p},\mathbf{s}^{*}=\mathbf{s}^{(\mathbf{k})},\mathbf{G},\sigma_{e}^{2},\mathbf{y}\right)\propto\prod_{k=1}^{v_{j}}\left(\pi_{j}^{(k)}\right)^{\alpha_{j}^{(k)}+I_{j}^{(k)}-1},\qquad\text{[A4]}$$

which corresponds to a series of Dirichlet distributions for $j = q_b+1, q_b+2,\ldots,q$.

The FCD's of each of $\sigma_e^2$ and $\mathbf{G}$ using the RAM specification do not have recognizable forms. Bink et al. (1998) suggested univariate Metropolis-Hastings sampling updates for various functions of variance components in their RAM-based specification. We alternatively base our MCMC algorithm on the method of composition using specifically Algorithm 2 of Chib and Carlin (1999) except that their data distribution is fully marginalized over the random effects whereas the RAM specification in [1] is only marginalized over the non-parent genetic effects. The joint posterior density of all parameters in a full animal model can be obtained from the reduced animal model as follows

$$p\left(\boldsymbol{\beta},\mathbf{a},\mathbf{m} \mid \mathbf{s}^* = \mathbf{s}^{(\mathbf{k})},\mathbf{G},\sigma_e^2,\mathbf{y}\right) = p\left(\boldsymbol{\beta},\mathbf{a}_p,\mathbf{m}_p \mid \mathbf{s}^* = \mathbf{s}^{(\mathbf{k})},\mathbf{G},\sigma_e^2,\mathbf{y}\right) \times p\left(\boldsymbol{\gamma}_t,\boldsymbol{\delta}_t \mid \mathbf{a}_p,\mathbf{m}_p,\mathbf{G},\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}\right) \quad . \qquad [A5]$$

That is, a random draw from [A5] is equivalent to a random draw from [A1] followed by a random draw from $p\left(\boldsymbol{\gamma}_t,\boldsymbol{\delta}_t \mid \mathbf{a}_p,\mathbf{m}_p,\mathbf{G},\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}\right)$ that can be readily derived as a sequence of univariate draws from the additive $\gamma_j^{(k)}$ and maternal $\delta_j^{(k)}$ Mendelian sampling term. Specifically, this involves sampling first from

$$\gamma_j^{(k)} \mid \boldsymbol{\beta},\mathbf{a}_p,\mathbf{m}_p,\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})},\mathbf{G},\sigma_e^2,\mathbf{y} \sim NID\left(\left(\frac{1}{\sigma_e^2}+\frac{\left(\omega_j^{(k)}\right)^{-1}}{\sigma_a^2}\right)^{-1}\frac{e_{ij}^{(k)}}{\sigma_e^2},\left(\frac{1}{\sigma_e^2}+\frac{\left(\omega_j^{(k)}\right)^{-1}}{\sigma_a^2}\right)^{-1}\right),$$

$$j = q_p+1, q_p+2,\ldots,q, \qquad [A6]$$

followed by

$$\delta_j^{(k)} \mid \boldsymbol{\gamma}_t, \boldsymbol{\beta}, \mathbf{a}_p, \mathbf{m}_p, \mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}, \mathbf{G}, \sigma_e^2, \mathbf{y} \sim NID\left( -\frac{g^{12}}{g^{22}} \gamma_j^{(k)}, \frac{\omega_j^{(k)}}{g^{22}} \right)$$

$$j = q_p{+}1, q_p{+}2, \ldots, q, \quad [A7]$$

Let $p(\mathbf{G})$ be a conjugate inverted Wishart prior density with parameters $v_g$ and $\mathbf{G}_o$ such that $E\left(\mathbf{G} \mid v_g, \mathbf{G}_o\right) = \frac{1}{v_g - 3} \mathbf{G}_o^{-1}$. The FCD of **G** given the augmentation of the RAM joint posterior density in [8] with $\boldsymbol{\gamma}_t$ and $\boldsymbol{\delta}_t$ is:

$$p\left(\mathbf{G} \mid \boldsymbol{\beta}, \mathbf{a}, \mathbf{m}, \mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}, \sigma_e^2, \mathbf{y}\right) \propto \left|\mathbf{G}\right|^{-\frac{q+v_g+3}{2}} \exp\left(-0.5 trace\left(\mathbf{G}^{-1}\left(\mathbf{S}_\mathbf{G} + \mathbf{G}_\mathbf{o}^{-1}\right)\right)\right), \quad [A8]$$

where

$$\mathbf{S}_\mathbf{G} = \begin{bmatrix} \mathbf{a}'\mathbf{A}^{-1}\mathbf{a} & \mathbf{a}'\mathbf{A}^{-1}\mathbf{m} \\ \mathbf{m}'\mathbf{A}^{-1}\mathbf{a} & \mathbf{m}'\mathbf{A}^{-1}\mathbf{m} \end{bmatrix}.$$

These components of $\mathbf{S}_\mathbf{G}$ can be readily computed without explicitly determining $\mathbf{a}_t$ and $\mathbf{m}_t$. For example, using results from Quaas (1988) and those in this chapter, $\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} = \mathbf{a}_p'\mathbf{A}_{pp}^{-1}\mathbf{a}_p + \boldsymbol{\gamma}_t'\boldsymbol{\Omega}_{tt}^{-1}\boldsymbol{\gamma}_t$, where $\boldsymbol{\gamma}_t'\boldsymbol{\Omega}_{tt}^{-1}\boldsymbol{\gamma}_t = \sum_{j=q_p+1}^{q} \frac{\gamma_j^2}{\omega_j}$.

Finally, let $p\left(\sigma_e^2\right)$ be an inverted-gamma density with parameters $\alpha_e$ and $\beta_e$. Then the FCD of $\sigma_e^2$ is also inverted-gamma and given by:

$$p\left(\sigma_e^2 \mid \boldsymbol{\beta}, \mathbf{a}, \mathbf{m}, \mathbf{s} = \mathbf{s}^{(k)}, \mathbf{G}, \mathbf{y}\right) \propto \left(\sigma_e^2\right)^{-(n/2+\alpha_e-1)} \exp\left(-\frac{1}{\sigma_e^2}\left(\frac{\mathbf{e}'\mathbf{e}}{2} + \beta_e\right)\right). \quad [A9]$$

The first $n_p$ elements of **e** are $\mathbf{e}_p = \left\{e_{ij}\right\}_{j=1}^{q_p}$ which are residuals due to records on parents.

The last $n_t$ elements of **e** are $\mathbf{e}_t^{(\mathbf{k})} = \left\{e_{ij}^{(k)} - \gamma_j^{(k)}\right\}_{j=q_p+1}^{q}$ with $\mathbf{e}_t^{(\mathbf{k})} = \mathbf{e}_t \big|_{\mathbf{s}_t^{*}=\mathbf{s}_t^{(\mathbf{k})}}$ indicating the dependence of $\mathbf{e}_t$ on sire assignments $\mathbf{s}_t^{*} = \mathbf{s}_t^{(\mathbf{k})}$ on non-parent animals.

The MCMC sampling scheme can thus be summarized as follows:

1) Draw samples of $\boldsymbol{\beta}$, $\mathbf{a}_p$, and $\mathbf{m}_p$ from [A1] using the proposition from the appendix of Wang et al.(1994b);
2) Draw samples of $\boldsymbol{\gamma}_t$ and $\boldsymbol{\delta}_t$ from [A6] and [A7];
3) Compute $\mathbf{S_G}$ using the samples of $\mathbf{a}_p$, $\mathbf{m}_p$, $\boldsymbol{\gamma}_t$, and $\boldsymbol{\delta}_t$ in order to sample **G** from a scaled inverted Wishart distribution [A8];
4) Determine $\mathbf{e}_t^{(\mathbf{k})}$ and combine with $\mathbf{e}_p$ to sample $\sigma_e^2$ from an inverted-gamma distribution [A9];
5) For each animal $j$ with uncertain paternity, independently draw a sire $s_j^{*}$ using as the probability of assignment either [A2] if the animal is parent or [A3] if the animal is non-parent.
6) For each animal $j$ with uncertain paternity, independently draw $\boldsymbol{\pi}_j$ from a Dirichlet distribution [A4].

# CHAPTER 2

## GENETIC EVALUATION OF BEEF CATTLE ACCOUNTING FOR UNCERTAIN PATERNITY

**ABSTRACT:** A hierarchical Bayes (HIER) model for the quantitative genetic analysis of performance data when some animals have uncertain paternity was compared to a model based on the use of Henderson's average numerator relationship matrix (ANRM). A simulation study consisted of ten datasets characterized by 30% of animals having uncertain paternity for each of two traits: one having moderate heritabilities for direct and maternal genetic effects on weaning weight (WWT) and another having high heritability for direct genetic effects on post-weaning gain (PWG). Posterior inference on the variance components was very similar between the two models. In an application to WWT and PWG data from Brazilian Herefords, posterior inference on variance components was also very similar between ANRM and HIER. Furthermore, rank correlations on posterior means for genetic effects between the two models exceeded 0.90. Nevertheless, large differences in posterior means between these two models were observed for some animals. Furthermore, animals with uncertain paternity had generally larger posterior standard deviations of genetic effects using the HIER model likely because the HIER model, unlike the ANRM model, infers upon sire assignment probabilities. Bayesian model choice criteria consistently favored the HIER model over the ANRM model in both simulation and Hereford data analysis studies.

**Key Words:** Bayesian inference, Beef cattle, Genetic evaluation, Multiple-sires, Uncertain paternity.

## 1. Introduction

The use of multiple-sire mating is common in extensive beef cattle production systems in countries such as Argentina, Australia, Brazil, and United States. Due to the size of herds and pasture paddocks, cows are generally exposed to more than one male within the same breeding season, thereby generating uncertainty on paternity assignments and adversely affecting accuracy of breeding value predictions.

Best linear unbiased prediction (BLUP) of genetic merit based on the average numerator relationship matrix (ANRM) (Henderson, 1988), has been the method of choice for genetic evaluation of animals with uncertain paternity. Furthermore, this procedure has been proven in simulation studies to increase selection response compared to the use of genetic groupings (Perez-Enciso and Fernando, 1992; Sullivan, 1995). The ANRM is based on knowledge of the true probabilities of each candidate male being the correct sire, and its inverse is readily computable (Famula, 1992). Alternative methods that infer upon uncertainty on paternity have been proposed (Foulley et al., 1987; Im, 1992) but have been restricted to sire model specifications.

We recently proposed a hierarchical animal model for inference on genetic merit of individuals with uncertain paternity and their sire assignments using Bayesian MCMC methods (Chapter 1). Based on a simulation study, the average posterior probabilities attributed to the true sire based on this model were between 1 and 13% larger than the respective priors (the inverse of the number of candidate sires) with differences depending upon heritabilities and multiple-sire group sizes. Posterior means of additive and maternal genetic effects obtained using the hierarchical model were very similar to those based on the ANRM; nevertheless, Bayesian model choice criteria consistently favored the proposed model over ANRM. In this simulation study, variance components were treated as known as typical of many current genetic evaluation systems.

The objectives of the current study were: 1) to validate the use of the model presented in Chapter 1 for quantitative genetic inference on variance components in situations characterized by uncertain paternity, 2) to demonstrate the utility of this model for the analysis of weaning weight and post-weaning gain data in Brazilian Herefords, and 3) to further compare the relative merit of this hierarchical model with a model based on the use of Henderson's ANRM.

## 2. Materials and methods

### 2.1. Data

*2.1.1. Simulation study*

The simulation study of Chapter 1 was revisited for the purpose of comparing their hierarchical model with one based on the use of the ANRM. This study consisted of ten simulated datasets or replicates for each of two different types of traits. Trait 1 had medium direct heritability ($h_a^2$ =0.3) and medium maternal heritability ($h_m^2$= 0.2) and a slightly negative direct-maternal genetic correlation ($r_{am}$= −0.2) as, for example, would characterize weaning weight. Trait 2 had a high direct heritability ($h_a^2$=0.5), but null $h_m^2$ as would characterize post-weaning gain. The residual variance ($\sigma_e^2$) was 60 and 50, respectively, for Traits 1 and 2.

Each simulated dataset involved 80 sires, 400 dams (480 parents) and 2000 non-parent animals, and the probability of any offspring deriving from an uncertain paternity situation (i.e. a multiple sire mating group) was 30%. Only one record was generated for each of the two traits per each animal. Our design was intended to mimic a likely situation in extensively managed beef cattle populations, where sires can be grouped into one of three different categories: 1) sires having only certain progeny (including AI bulls); 2) sires having certain and uncertain progeny, and 3) sires having only uncertain progeny. Additional details on the design of the simulation study can be found in Chapter 1.

### *2.1.2 Hereford performance data*

To demonstrate a comparison of the hierarchical model described in Chapter 1 with a model based on the use of the ANRM, MCMC procedures were used to analyze performance records of Hereford calves raised in a single Southern Brazilian herd from 1991 to 1999. These records are part of the data collected by the Brazilian Breeders Association and Gensys Associate Consultants within a large-scale genetic evaluation program called the "Delta G Connection".

Animals were raised on extensive sub-tropical pasture conditions. Traits analyzed were weaning weight (WWT) and post-weaning gain (PWG). The mean ± standard deviation for WWT was 172.6 ± 35.6 kg whereas that for PWG was 110.7 ± 30.4 kg. Ages of calves at weaning ranged from 100 to 293 days with mean 202 days, whereas mean post-weaning test periods ranged from 111 to 453 days with mean 218 days. This herd was characterized by extensive multiple-sire mating with excellent recording on identification of candidate bulls within each multiple-sire group. A total 5,399 records on WWT and 3,402 records on PWG were analyzed. Including base population animals within the pedigree file, there were a total of 6,905 animals genetically evaluated for WWT and 4,703 animals evaluated for PWG.

For WWT, there were 4,228 (61.2%) animals with certain paternity and 1,171 (17.0%) animals with uncertain paternity leaving the remaining 1,506 (21.8%) as base population animals. For PWG, there were 2,702 (57.4%) animals with certain paternity, 700 (14.9%) with uncertain paternity and 1,301 (27.7%) base population animals. The number of candidate males in the multiple-sire groups, with respective group sizes in parenthesis by trait, i.e. (WWT/PWG), were 2 (57/23), 3(2/2), 4 (234/167), 5 (207/117), 6 (272/176), 10 (135/43), 12 (16/9) and 17 (248/163).

### 2.2. Bayesian inference

*2.2.1. Analyses of Simulation Study*

For each of the ten replicates on each of the two traits (i.e. single trait analyses) in the simulation study, variance components were estimated using Reduced Animal Model (RAM) implementations (Quaas and Pollak, 1980) for each of three different models:

1) HIER: A hierarchical mixed effects model fully accounting for uncertainty on sire assignments as proposed in Chapter 1.
2) ANRM: A linear mixed effects model based on the average numerator relationship matrix (Henderson, 1988).
3) TRUE: A linear mixed effects model based on the true sire assignments, as if there were no uncertainty on assignments.

For the ANRM model, equal and fixed probabilities were assigned to each candidate sire for animals having uncertain paternity. The TRUE model was included to serve as a positive control for the other two models. In the HIER model, an equal prior probability, i.e., the inverse of the number of candidate sires within the particular mating group, was specified for each candidate sire of an animal with uncertain paternity using a Dirichlet prior density. The hyperparameters of this prior density were each equal to the inverse of the number of candidate sires in a group as in Chapter 1. Diffuse scaled inverted chi-square prior distributions were specified for each variance component, i.e. $\sigma^2 \sim \left(\nu\sigma_o^2\right)\chi_\nu^{-2}$ where $\sigma_o^2$ was the true value for the respective variance component with $\nu$ = 8 being the prior degrees of belief in all cases. Additionally, for the analysis of WWT, an inverted Wishart prior density $\mathbf{G} \sim IW_2\left(\nu\mathbf{G}_o,\nu\right)$ with $\mathrm{E}(\mathbf{G}) = \dfrac{\nu}{\nu-3}\mathbf{G}_o$ was placed on $\mathbf{G} = \begin{bmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{bmatrix}$ where $\mathbf{G}_0$ represented the true value of $\mathbf{G}$, again with $\nu$ = 8. Here $\sigma_a^2$ is the additive genetic variance, $\sigma_m^2$ is the maternal genetic variance and $\sigma_{am}$ is the direct maternal genetic covariance. For all three models, a MCMC sampling chain of 20,000 cycles was run after a burn-in period of 4,000 cycles on each dataset. Furthermore, the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and Pseudo Bayes Factors (PBF) (Gelfand, 1996) were computed from the MCMC output as model choice criteria. Additional details on DIC and PBF computations specific to the ANRM and HIER models are provided in Chapter 1. Smaller DIC values are indicative of better model fit whereas PBF ratios less than one for the ANRM relative to the HIER model favor the HIER model.

### *2.2.2. Analyses of Hereford performance data*

The trait PWG was modeled as a linear function of fixed effects (i.e. effects with flat or bounded uniform subjective priors) and random effects (i.e. effects with multivariate normal structural prior specifications). Fixed effects included gender, linear and quadratic effects of age of dam (2 to 12 years), linear and quadratic effects of age of calf at the end of the test period (305 to 658 days), and linear and quadratic effects of post-weaning test period (111 to 453 days). Random effects included multivariate normal additive genetic effects (1,932 parents), with correlation determined by the numerator relationship matrix (NRM), and normally, identically, and independently distributed (*niid*) effects of contemporary groups (year-season-management; 237 levels). The two competing models differed in their treatment of NRM; i.e. it was either inferred upon using the HIER model of Chapter 1 or it was determined using the ANRM of Henderson (1988). For both models, a Gibbs chain of 70,000 cycles was run, after 4,000 cycles of burn-in.

The statistical model for WWT included the fixed effects of gender, linear and quadratic effects of age of dam, linear and quadratic effects of age of calf at weaning (100 to 293 days), and the linear effect of birth weight (16 to 60 kg). Random effects included jointly multivariate normal additive and maternal genetic effects with correlation within and between these two sets of effects determined by a NRM (determined by HIER or by ANRM), $\sigma_a^2$, $\sigma_m^2$ and $r_{am}$ or, equivalently, the additive maternal genetic covariance $\sigma_{am} = r_{am}\sqrt{\sigma_a^2\sigma_m^2}$. Additional random effects included *niid* dam permanent environmental effects (2,281 levels) with variance $\sigma_{pe}^2$ and *niid* contemporary group effects (96 year-season-management subclasses) with variance $\sigma_{cg}^2$. MCMC chains of 180,000 and 120,000 cycles were run for HIER and ANRM, respectively, after 4,000 cycles of burn-in.

For each variance component, the initial monotone sequence approach (Geyer, 1992) was used to determine effective sample size (ESS), which is an estimate of the number of independent draws with information content equivalent to that contained within the dependent samples (Sorensen et al., 1995). The length of MCMC chain was determined such that all parameters had an ESS of at least 100. There was surprisingly slow "mixing" in the MCMC chain for WWT based on the RAM specification for the HIER model in Chapter 1. Therefore, we adopted a full animal model implementation for HIER and ANRM in the analyses of WWT, whereas the RAM implementation was maintained for the HIER and ANRM analyses of PWG. Priors on scalar variance components were scaled inverted chi-square distributions, that is, $\sigma^2 \sim \left(\nu\sigma_o^2\right)\chi_\nu^{-2}$. For PWG, we specified $\sigma_o^2 = 200$ for $\sigma_e^2$, $\sigma_o^2 = 100$ for $\sigma_a^2$, and $\sigma_o^2 = 800$ for $\sigma_{cg}^2$; whereas for WWT, we specified $\sigma_o^2 = 200$ for $\sigma_e^2$; $\sigma_o^2 = 100$ for $\sigma_{pe}^2$ and $\sigma_o^2 = 600$ for $\sigma_{cg}^2$. Moreover, the prior on the additive genetic variance-covariance matrix for WWT was specified to be $\mathbf{G} \sim IW_2\left(\nu\begin{bmatrix}100 & -5\\ -5 & 50\end{bmatrix},\nu\right)$. As with the simulation study, $\nu$ = 8 for all (co)variance components and a Dirichlet prior was specified on the candidate sire probabilities with the same hyperparameter specifications. Posterior means, medians, the 2.5th and 97.5th percentiles and standard deviations of the parameters were obtained from the marginal posterior densities. Posterior means and standard deviations of genetic effects were also determined using the ANRM and the HIER models. Finally, DIC and PBF were computed as measures of model choice.

## 3. Results

### 3.1. Posterior inference on variance components

Variance components were estimated for each of the ten replicates simulated for each trait. The posterior median, 2.5% and 97.5% posterior percentiles averaged over the ten replicates using all three models are presented in Table 2.1. The minimum and maximum posterior median estimated over the ten replicates is also presented.

The averaged posterior medians obtained by using the TRUE sire assignments (Table 2.1) were very close to the true variance component values which were always included within each replicate's 95% equal-tailed posterior probability interval (PPI), determined as the range between the 2.5th and 97.5th percentiles of the posterior densities for individual replicates on each trait. For both the ANRM and HIER models, the true variance component values were generally well included in the each replicate's 95% PPI and the posterior medians (Table 2.1) were generally close to their respective true values. The 95% PPI using both the HIER and ANRM models did not include the true value of the $\sigma_{am}$ and $\sigma_e^2$ for Trait 1 in one replicate, and the true $\sigma_a^2$ for Trait 2 in another replicate; nevertheless, collectively, these results were well within probabilistic (i.e. 95% coverage) expectation. There was naturally more uncertainty being modeled using ANRM and slightly more so using HIER relative to the TRUE model which, as expected, translates into reduced precision on variance component inference. Furthermore, as we compare the average, minimum and maximum posterior medians across the ten replicates

Table 2.1. Posterior median (PMED), 95% posterior probability intervals (PPI), and effective sample size (ESS) of variance components averaged across ten replicated datasets on Traits 1 and 2, obtained by the hierarchical model (HIER), a model based on the average numerator relationship matrix (ANRM), and a model based on known sire assignments (TRUE). Maximum (MAX) and minimum (MIN) posterior medians across the ten replicates are also reported.

| Method | Parameter[a] | PMED | PPI | MIN | MAX | ESS |
|---|---|---|---|---|---|---|
| *Trait 1* | | | | | | |
| ANRM | $\sigma_a^2=30$ | 29.79 | (19.01, 44.36) | 20.42 | 39.37 | 140 |
| | $\sigma_{am}=-5$ | -5.17 | (-14.46, 2.22) | -14.55 | 1.93 | 111 |
| | $\sigma_m^2=20$ | 21.40 | (14.39, 30.54) | 13.33 | 25.31 | 138 |
| | $\sigma_e^2=60$ | 59.55 | (50.64, 67.62) | 50.41 | 66.11 | 183 |
| HIER | $\sigma_a^2=30$ | 34.31 | (21.41, 52.20) | 21.06 | 49.75 | 114 |
| | $\sigma_{am}=-5$ | -7.62 | (-17.94, 0.41) | -19.70 | 0.34 | 119 |
| | $\sigma_m^2=20$ | 22.71 | (15.33, 32.02) | 14.29 | 27.61 | 184 |
| | $\sigma_e^2=60$ | 56.24 | (45.51, 65.23) | 42.74 | 65.68 | 139 |
| TRUE | $\sigma_a^2=30$ | 29.45 | (19.16, 43.54) | 21.95 | 37.19 | 152 |
| | $\sigma_{am}=-5$ | -4.74 | (-13.78, 2.51) | -10.95 | 2.25 | 117 |
| | $\sigma_m^2=20$ | 21.15 | (14.31, 30.19) | 13.14 | 24.26 | 143 |
| | $\sigma_e^2=60$ | 59.86 | (51.76, 67.30) | 54.54 | 65.15 | 210 |
| *Trait 2* | | | | | | |
| ANRM | $\sigma_a^2=50$ | 48.91 | (39.17, 60.31) | 40.94 | 62.11 | 527 |
| | $\sigma_e^2=50$ | 49.74 | (43.14, 56.30) | 44.31 | 55.02 | 644 |
| HIER | $\sigma_a^2=50$ | 50.07 | (39.91, 61.57) | 42.18 | 66.13 | 508 |
| | $\sigma_e^2=50$ | 47.52 | (40.85, 54.37) | 40.80 | 54.08 | 575 |
| TRUE | $\sigma_a^2=50$ | 48.67 | (38.86, 60.33) | 41.74 | 63.01 | 528 |
| | $\sigma_e^2=50$ | 50.09 | (43.79, 56.32) | 45.49 | 54.50 | 671 |

[a] $\sigma_a^2$= additive variance, $\sigma_m^2$= maternal variance, $\sigma_{am}$= additive-maternal covariance, and $\sigma_e^2$= residual variance with true values specified.

for each trait as obtained by HIER (Table 2.1) to their true values, there appears to be a tendency for the posterior median to underestimate $\sigma_e^2$ for both traits. In the RAM implementation of HIER, candidate sires with sampled genetic effects that lead to smaller residual deviances should have a greater probability of being sampled as the sire of individuals with uncertain paternity, given the form of the full conditional density used to sample sire assignments of non-parent individuals (See Equation [A3] in Chapter 1). Similarly larger posterior medians of $\sigma_a^2$ tended to be associated with the HIER model.

Model choice criteria (DIC and PBF) were always in favor of the HIER model versus the ANRM model for both traits in all replicates (i.e. low PBF ratios for ANRM/HIER). The PBF, obtained as the *n*-ary product of observation-specific conditional predictive ordinate (CPO) ratios for ANRM/HIER across all records (Cardoso and Tempelman, 2003), varied from $4.85\times10^{-2}$ to $2.78\times10^{-22}$ for Trait 1 and from $1.49\times10^{-7}$ to $8.12\times10^{-28}$ for Trait 2. The DIC for both traits and models averaged across the 10 replicates are presented in Table 2.2. This criterion consists of a deviance component $\bar{\mathrm{D}}$, which was always smaller for HIER compared to ANRM, and a penalty ($p_D$) for effective number of parameters, which was always larger for HIER compared to ANRM. However, $p_D$ was not large enough to overcome the smaller $\bar{\mathrm{D}}$ such that DIC = $\bar{\mathrm{D}} + p_D$ was always smaller for HIER in all replicates and for both traits. Both criteria therefore favored the HIER model as a better fit to simulated data, when variance components were unknown, being consistent with our previous conclusions where variance components were treated as known (Chapter 1).

Table 2.2. The mean deviance $\bar{D}$, penalty for effective number of parameters ($p_D$) and the Deviance Information Criterion (DIC = $\bar{D}$ + $p_D$) averaged across the analysis of 10 simulated replicates for each of Traits 1 and 2 based on the ANRM and HIER models.

| Method | $\bar{D}$ | $p_D$ | DIC |
|---|---|---|---|
| *Trait 1* | | | |
| ANRM | 17,153 | 701 | 17,853 |
| HIER | 16,976 | 788 | 17,764 |
| *Trait 2* | | | |
| ANRM | 16,714 | 913 | 17,627 |
| HIER | 16,590 | 949 | 17,539 |

### 3.2. Inference on Brazilian Hereford data

#### *3.2.1 Post-weaning gain*

DIC and its components for each of the ANRM and HIER model analyses of PWG are presented in Table 2.3. Similar to results obtained from the simulation study, $\bar{D}$ for HIER was smaller than that for ANRM. Although $p_D$ was larger for HIER compared to ANRM, the resulting DIC favored HIER over ANRM, as the advantage attained in $\bar{D}$ was greater than the smaller $p_D$ associated with ANRM. The HIER model appears to fit the data better since a DIC difference between two models exceeding seven is considered to be somewhat decisive (Spiegelhalter et al., 2002). Moreover, the PBF for ANRM/HIER obtained for PWG was $9.695 \times 10^{-2}$, indicating that the HIER model fitted the data approximately 10 times better than ANRM using this measure.

Both models provided similar inference for variance components and associated genetic parameters (Table 2.4). Our heritability determination did not include $\sigma^2_{cg}$ as part of the phenotypic variance in order to facilitate comparisons with literature (de Mattos et al., 2000; Koots et al., 1994; Meyer, 1992), where contemporary group effects are

Table 2.3. The deviance ($\bar{D}$), penalty for effective number of parameters ($p_D$) and the Deviance Information Criterion (DIC) for ANRM and HIER models used for the analyses of post-weaning gain (PWG) and weaning weight (WWT) on Brazilian Herefords.

| Method | $\bar{D}$ | $p_D$ | DIC |
|---|---|---|---|
| *PWG* | | | |
| ANRM | 28,379 | 829 | 29,208 |
| HIER | 28,329 | 860 | 29,189 |
| *WWT* | | | |
| ANRM | 44,072 | 2,136 | 46,208 |
| HIER | 44,043 | 2,148 | 46,190 |

generally considered as fixed. The posterior median additive heritability ($h_a^2$) was 0.23 and 0.24, for ANRM and HIER, respectively. These point estimates, although smaller than the average value of 0.31 based on an extensive literature review (Koots et al., 1994), are well within the anticipated range for extensively managed production environments and considerably larger than point heritability estimates for yearling weights of Hereford cattle raised in Australia (Meyer, 1992). The 95% PPI for each variance component overlapped widely between the two models, suggesting no practical difference between models for variance component and genetic parameter inference on PWG.

The rank correlation between posterior means of additive genetic effects obtained by the two different models was greater than 0.99, whether estimated for all animals or stratified by base population animals (i.e., no pedigree information), for animals with certain paternity, or for animals with uncertain paternity. For animals ranked in the top 10% and top 5% for posterior mean additive genetic values by HIER, the rank correlation with ANRM was still greater than 0.99 for base animals and animals with certain

Table 2.4. Posterior median, 95% posterior probability intervals (PPI), and effective sample sizes (ESS) of variance components (in kg$^2$) and genetic parameters for post-weaning gain in Brazilian Herefords, obtained by ANRM and HIER models.

| Parameter[a] | Posterior median | PPI | ESS |
|---|---|---|---|
| *ANRM* | | | |
| $h_a^2$ | 0.231 | (0.153, 0.316) | 254 |
| $\sigma_a^2$ | 73.8 | (48.0, 103.6) | 254 |
| $\sigma_e^2$ | 246.5 | (221.5, 271.2) | 368 |
| $\sigma_{cg}^2$ | 404.5 | (334.3, 494.0) | 34,871 |
| *HIER* | | | |
| $h_a^2$ | 0.244 | (0.162, 0.336) | 337 |
| $\sigma_a^2$ | 78.2 | (51.1, 111.2) | 334 |
| $\sigma_e^2$ | 242.9 | (216.5, 268.2) | 487 |
| $\sigma_{cg}^2$ | 404.5 | (333.9, 493.8) | 35,691 |

[a] $h_a^2$= additive heritability; $\sigma_a^2$= additive variance; $\sigma_e^2$= residual variance; and $\sigma_{cg}^2$ =contemporary group variance.

paternity. However, for animals with uncertain paternity, the rank correlation dropped to 0.98 and 0.94, for animals in the top 10% and 5%, respectively. Despite these slightly lower rank correlations, there was general agreement between the two models in terms of ranks with any substantial differences naturally relating to the modeling of uncertainty on paternity assignments. In particular, the differences in posterior means on direct genetic effects between ANRM and HIER ranged from $-1.22$ to 2.94 kg which is not necessarily trivial relative to the genetic standard deviation $\sqrt{\sigma_a^2}$ .

Both HIER and ANRM models lead to similar posterior standard deviation of additive genetic values for base population animals and for animals with known sires (data not shown); however, for individuals with uncertain paternity, the HIER model generally had larger posterior standard deviations of additive genetic values, particularly when these standard deviations were high (Figure 2.1). The posterior standard deviations using the ANRM model were generally less than those of the HIER model above a posterior standard deviation of 7.13 kg as based on the intersection of the two lines provided in Figure 2.1. The first line is the line of best fit between the two sets of posterior standard deviations whereas the other line is a unitary line (i.e. of slope 1 and intercept 0). The estimated slope (0.68) from the line of best fit indicates that beyond a posterior standard deviation of 7.13 kg, which includes almost all animals as observed in Figure 2.1, the posterior standard deviation for a ANRM based genetic evaluation increases only 0.68 kg for every kg increase in the HIER posterior standard deviation. These posterior standard deviations can be readily interpreted as being analogous to standard errors of prediction predominantly used by industry to determine reliabilities of estimated breeding values based on BLUP. Therefore, potentially upward biases in reported reliabilities for estimated breeding values using ANRM may occur for animals having uncertain paternity. This result is anticipated since the ANRM model treats sire assignments probabilities as known whereas the HIER model infers upon these probabilities. Admittedly, the slightly larger additive genetic variance for HIER may also contribute to the slightly larger standard deviations observed in Figure 2.1. It is curious to note from Figure 2.1 that the two animals having the smallest posterior standard deviations for additive genetic merit (as pointed out with black arrows) have their corresponding ANRM versus HIER posterior standard deviations fall on a unitary line

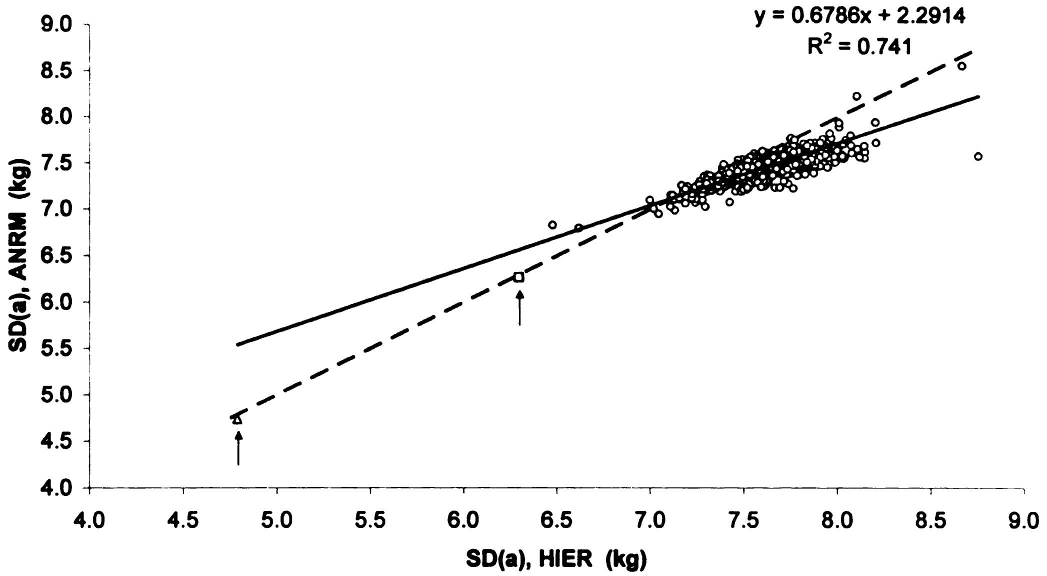

Figure 2.1. Scatter plot of standard deviation (SD) of additive genetic effects of post-weaning gain, in kg, of Brazilian Herefords with uncertain paternity, obtained by ANRM vs. HIER. Solid line represents the least-squares fit represented by the reported regression equation presented in the graph whereas the dashed line has slope one and null intercept. $R^2$ is the coefficient of determination for least-squares fit. Black arrows point to sires that have substantial numbers (9 and 50) of progeny of their own.

(i.e. with slope one and null intercept). Both individuals are sires having a fairly large number of known progeny in the dataset. The sire with smallest standard deviation (represented by a triangle symbol in Figure 2.1) had 50 known progeny whereas the other sire (represented by a square symbol in Figure 2.1) had 9 known progeny with records on PWG. Subsequently, the estimated reliabilities of their genetic evaluations for PWG are almost identical between the two models, as would be expected in such a scenario.

### *3.2.2. Weaning weight*

As with PWG, we observed a smaller $\bar{\mathrm{D}}$ and DIC for the HIER model compared to ANRM model for the analysis of WWT (Table 2.3). However, the $p_D$ difference between the HIER and ANRM models was smaller for WWT compared to PWG. This may be due to the lower total (i.e. additive and maternal) heritability (Willham, 1972) for WWT as estimated by a posterior median of 0.13 in this population. As for PWG, WWT heritabilities were calculated omitting the contemporary group variance in the phenotypic variance. Low heritabilities lead to less powerful discrimination between candidate sires based only on phenotypes (Chapter 1). For example, we have previously observed that there was no difference between prior and posterior probabilities of sire assignments when $h_a^2$=0.10 (Cardoso and Tempelman, 2001). Given these circumstances, the ANRM and HIER models might be expected to fit equally well to WWT. Nevertheless, there was an appreciable DIC difference between the two models in favor of the HIER model (Table 2.3). Moreover, the PBF of ANRM to HIER obtained for WWT was $1.654 \times 10^{-1}$; i.e. HIER is estimated to fit the data approximately six times better than ANRM.

Posterior inference on variance components and genetic parameters using ANRM and HIER were very similar as presented in Tables 2.5 and 2.6. The 95% PPI for each variance component and genetic parameter widely overlapped with posterior medians being very similar between the HIER and ANRM models (Table 2.5). Despite a substantially negative posterior median, the 95% PPI for $\sigma_{am}$ included zero under both models.

Table 2.5. Posterior median, 95% posterior probability intervals (PPI), and effective sample sizes (ESS) of variance components (in kg$^2$) for weaning weight in Brazilian Herefords, obtained using ANRM and HIER models.

| Parameter[a] | Posterior median | PPI | ESS |
|---|---|---|---|
| *ANRM* | | | |
| $\sigma_a^2$ | 58.1 | (33.8, 95.9) | 153 |
| $\sigma_{am}$ | -21.1 | (-58.3, 4.1) | 131 |
| $\sigma_m^2$ | 80.4 | (40.7, 131.5) | 130 |
| $\sigma_{pe}^2$ | 194.5 | (157.4, 231.6) | 331 |
| $\sigma_e^2$ | 206.3 | (183.1, 225.2) | 219 |
| $\sigma_{cg}^2$ | 295.9 | (223.2, 401.1) | 34,234 |
| *HIER* | | | |
| $\sigma_a^2$ | 58.0 | (34.1, 108.3) | 167 |
| $\sigma_{am}$ | -23.4 | (-62.6, 1.2) | 127 |
| $\sigma_m^2$ | 82.7 | (45.5, 135.6) | 171 |
| $\sigma_{pe}^2$ | 195.0 | (157.1, 230.7) | 394 |
| $\sigma_e^2$ | 205.9 | (177.2, 224.9) | 221 |
| $\sigma_{cg}^2$ | 296.4 | (223.9, 401.4) | 48,448 |

[a] $\sigma_a^2$= additive variance; $\sigma_{am}$= additive-maternal covariance; $\sigma_m^2$= maternal variance; $\sigma_{pe}^2$= permanent maternal environment variance; $\sigma_e^2$= residual variance; and $\sigma_{cg}^2$ =contemporary group variance.

The posterior mean for $h_a^2$ was 0.12 under both models (Table 2.6) and appears low compared to results from a study on Angus cattle raised and controlled under similar conditions where the corresponding point estimate was 0.26 (Cardoso et al., 2001) and compared to results of several Hereford studies in the literature (Koots et al., 1994; Meyer, 1992). The posterior mean for $h_m^2$ was 0.18 for ANRM and 0.19 for HIER, being similar to point estimates derived from studies on Hereford populations of Australia

Table 2.6. Posterior medians, 95% posterior probability intervals (PPI), and effective sample sizes (ESS) of genetic parameters for weaning weight in Brazilian Herefords, obtained using ANRM and HIER models.

| Parameter[a] | Posterior median | PPI | ESS |
|---|---|---|---|
| *ANRM* | | | |
| $h_a^2$ | 0.117 | (0.067, 0.201) | 151 |
| $r_{am}$ | -0.320 | (-0.617, 0.076) | 142 |
| $h_m^2$ | 0.183 | (0.089, 0.323) | 129 |
| $h_t^2$ | 0.132 | (0.083, 0.194) | 248 |
| $p^2$ | 0.393 | (0.317, 0.461) | 290 |
| *HIER* | | | |
| $h_a^2$ | 0.117 | (0.068, 0.229) | 156 |
| $r_{am}$ | -0.350 | (-0.632, 0.021) | 164 |
| $h_m^2$ | 0.190 | (0.101, 0.344) | 157 |
| $h_t^2$ | 0.129 | (0.079, 0.196) | 363 |
| $p^2$ | 0.396 | (0.319, 0.463) | 337 |

[1] $h_a^2$= additive heritability; $r_{am}$ = additive-maternal correlation; $h_m^2$= maternal heritability; $h_t^2$= total heritability; and $p^2$= permanent maternal environment fraction of the phenotypic variance.

(Meyer, 1992), the United States of America, Canada and Uruguay (de Mattos et al., 2000). The posterior mean of $\sigma_{pe}^2$ was surprisingly high, accounting for almost 40% of the phenotypic variance (excluding $\sigma_{cg}^2$) under both models (Tables 2.5 and 2.6). Previous investigators determined that maternal environments account for 15 to 23% of the phenotypic variance for WWT in Herefords (de Mattos et al., 2000; Meyer, 1992). Relatively higher estimates of $\sigma_{pe}^2$ tend to be found in Herefords than for other breeds (Meyer, 1992, 1993) indicating that perhaps milk production is a critical discriminating factor with Hereford dams. Moreover, maternal environment effects have been consistently found to be more important than maternal genetic effects for Hereford herds (Meyer, 1992). Furthermore, estimates of $\sigma_{pe}^2$ may be biased upwards by the presence of non-additive genetic variation.

As for the PWG analysis, there was general agreement between the two models for posterior means for genetic effects for WWT. The rank correlation between posterior means of additive genetic effects was greater than 0.99 for all animals with certain and uncertain paternity and 0.97 for all base animals. For animals ranked in the top 10% and top 5% by posterior means of additive genetic effects by HIER, the rank correlation between the two models was greater than 0.97 for animals with certain paternity. However, for base animals and animals with uncertain paternity, the rank correlations dropped to between 0.90 and 0.94. For maternal genetic effects, the corresponding correlation was always 0.99 or greater, regardless of stratification. As for PWG, any difference from a perfect rank correlation in estimated breeding values between the two models, beyond Monte Carlo variability, would be due to the modeling of uncertainty on paternity assignments. Differences in posterior means of additive genetic effects for WWT between ANRM and HIER models ranged from -2.17 to 1.57 kg for animals with uncertain paternity.

The two models had similar posterior standard deviations of additive and maternal genetic effects for base animals and animals with certain paternity (data not shown); however individuals with uncertain paternity tended to have larger standard deviations of additive (Figure 2.2) and maternal (data not shown) genetic values under the HIER model compared to the ANRM model. As with PWG, this also suggests a potential upward bias

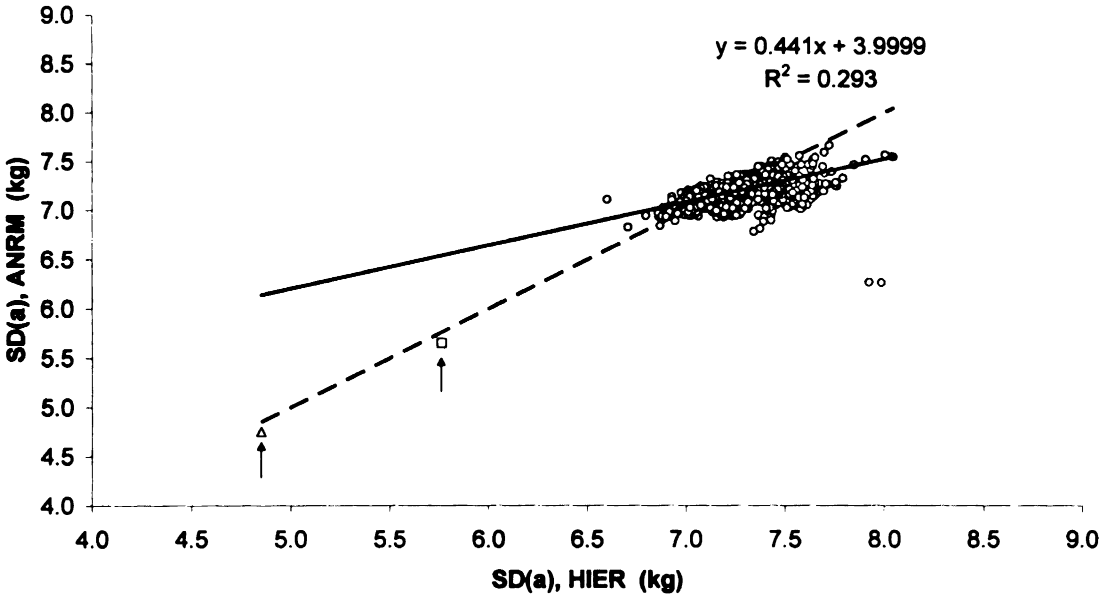

Figure 2.2. Scatter plot of standard deviation (SD) of additive (a) genetic effects of weaning weight, in kg, of Brazilian Herefords with uncertain paternity, obtained by ANRM vs. HIER. Solid line represents the least-squares fit represented by the reported regression equation presented in the graph whereas the dashed line has slope one and null intercept. $R^2$ is the coefficient of determination for least-squares fit. Black arrows point to sires that have substantial numbers (17 and 87) of progeny of their own.

in reported reliabilities for estimated breeding values using ANRM BLUP for WWT. We note that for WWT, unlike that for PWG, the posterior density of the additive genetic variance was virtually identical between the two models, thereby eliminating this as a possible cause for general differences in posterior standard deviations between the two models for individual additive genetic effects. The same two animals having additive genetic effects with the smallest posterior standard deviations for PWG in Figure 2.1, were also observed in Figure 2.2 (pointed by black arrows) as having the smallest posterior standard deviations for WWT. The respective number of known progeny with WWT records for the two sires were 87 (sire represented by a triangle symbol in Figure 2.2) and 17 (sire represented by a square symbol in Figure 2.2). The corresponding standard deviations for the two sires under the two models fall close to a unitary line in Figure 2.2 indicating that, as with PWG, the uncertainty on genetic merit for sires with a substantial number of known progeny is virtually identical between the two models.

### 4. Discussion

We validated by simulation study that the hierarchical Bayes model (HIER) of Chapter 1 is able to provide reliable inference on variance components when paternity is uncertain for some animals. Furthermore, Bayesian model choice criteria indicated this model to be a better fit to such data relative to a model based on the average numerator relationship matrix (ANRM) (Henderson, 1988).

We have also presented an application to PWG and WWT records on Brazilian Herefords. We found that the HIER model provided a better data fit compared to ANRM for the analysis of PWG and WWT. The difference between the two models is heritability ($h^2$) dependent: as $h^2$ decreases, the power for discriminating between candidate sires decreases such that the models become less distinctive from each other. On the other hand, the lower the $h^2$, the greater is the importance of pedigree relative to individual phenotypic information and hence of modeling uncertain paternity for the prediction of genetic merit (Famula, 1993; Sullivan, 1995).

The main advantage of the more complex HIER model compared to the ANRM model is in terms of properly accounting for reduced precision on genetic merit inference due to uncertainty on sire assignments. For genetic evaluations, there may be pragmatically little difference between models for rankings of estimated breeding values; however the estimated reliabilities associated with breeding values of animals with uncertain paternity would tend to be appreciably lower using the HIER model compared to the ANRM model, since the latter assumes that the true probabilities of paternity are known. Furthermore, there were some appreciable differences in posterior means of additive genetic merit for some animals. In a simulation study, Kerr et al. (1994a) found that ANRM tends to overestimate the accuracy of prediction and the selection response compared to the true additive relationship matrix, as consequence of a wrong gametic model assumption (i.e. genes cannot be transmitted from multiple-sires).

The computational time required to complete each MCMC cycle was actually very similar between the ANRM and HIER models. It appears that the time required to sample sire assignments and probabilities in HIER was balanced against the extra elements that need to be added to the mixed model equations using ANRM, since even the inverse of the ANRM is quite dense due to the incorporation of average relationships (Famula, 1992). For large scale genetic evaluations on a national or breed association level, it appears that a computationally tractable empirical Bayes or "plug-in" strategy may be advisable and may perhaps lead to potentially very little or no difference in estimated breeding values and standard errors of prediction relative to MCMC based inference. This conclusion was drawn by Kizilkaya et al. (2002) in the context of threshold mixed model analyses of calving ease scores. Current genetic evaluation plug-in methods such as BLUP based on fixed values for variance components and other hyperparameters may produce reasonable predictors provided the variance components and other hyperparameters can be well estimated using marginal likelihood based methods or reasonably specified (Natarajan and Kass, 2000). This issue may deserve further investigation if implementation of HIER model for genetic evaluation systems is considered.

The HIER model represents an important alternative for genetic prediction on beef cattle populations undergoing multiple-sire mating. These populations are generally raised in extensive pastoral conditions with multiple-sire matings used to increase the probability of pregnancy, since the size of cow breeding groups, as a function of paddock size, is generally too large to be sired by a single bull. Other causes of uncertain parentage include the use of artificial insemination followed by natural breeding or accidental/unplanned breeding. The paternal contribution to the breeding value of animals with uncertain paternity using the HIER model would be a function of the posterior probabilities of each male in the group being the correct sire of the individual given the prior and data information. Uncertain parentage issues pervade other livestock production systems as well (Van-Arendonk et al., 1998) such that our proposed model may be suitably adapted in those instances. Genetic markers additionally provide a useful source of objective prior information to be incorporated in the analysis.

Large herds provide a great potential for selection and genetic progress, but very often the exclusive use of single mating is very costly, maybe impractical, due to size of the operations and labor required. Some genetic evaluation programs, however, exclude animals with uncertain paternity from inclusion. Nonetheless, simulation results have shown that the loss in selection response due to multiple-sire mating compared to single mating is less than 10%, particularly when dams are recorded (Kerr et al., 1994b; Sullivan, 1995), while the exclusion of performance records on animals with uncertain paternity from the evaluation can represent a reduction on expected selection response up to 33% compared with the use of Henderson's ANRM and up to 24% compared to genetic grouping (Sullivan, 1995), depending on $h^2$ and percentage of animals with uncertain paternity. The HIER model confers additional modeling advantages over the ANRM model, provided that the trait of interest does not have a very low heritability.

It is unlikely that identification or pedigrees of all candidate sires would be recorded for all individuals with uncertain paternity in some populations such that a hybrid genetic evaluation model might be used, for example, one that models sire assignment uncertainty (as in HIER) but uses phantom or genetic grouping (Westell et al., 1988) for animals where candidate sire identifications are not available. One natural phantom group would be the mating groups themselves. However, inference on the average genetic merit of the males in the group may be highly confounded with other fixed effects (i.e. contemporary groups), especially if small groups are created (Quaas, 1988). Treating contemporary groups and/or genetic groups, as random effects with mild levels of connectedness might mitigate this issue somewhat.

# CHAPTER 3

# HIERARCHICAL BAYES MULTIPLE-BREED INFERENCE WITH AN APPLICATION TO GENETIC EVALUATION OF A NELORE-HEREFORD POPULATION

**ABSTRACT:** The primary objective of this study was to develop and apply a hierarchical multiple-breed animal model (MBAM) to estimate genotypic effects, breed-specific additive genetic variances and variance due to the segregation between breeds. Phenotypic records were modeled as function of additive (A), dominance and A × A genetic fixed effects and random animal additive genetic effects using appropriate multiple-breed additive variance-covariance specifications. We validated the MBAM on five simulated datasets derived from a population based on crosses from two breeds having a two-fold difference in genetic variance. Posterior means of all variance components obtained by MBAM, in each of the five populations were seemingly unbiased with 95% posterior probability intervals (PPI) having expected coverage. We also analyzed a dataset of 22,717 post weaning gain (PWG) records of a Nelore-Hereford population (40,082 animals in the pedigree). MBAM inference on Nelore and Hereford additive genetic variances (in kg$^2$) differed substantially. Herefords had a posterior mean genetic variance of 85.2 with a 95% PPI of (63.2, 108.5) whereas the corresponding values for the Nelores were 23.8 and (13.0, 39.5), respectively. The posterior mean variance due to the segregation between these breeds was 8.4 with a 95% PPI of (2.3, 24.8). The posterior mean of the genetic variance obtained by a conventional animal model (AM) was 60.5, presumed common for the two breeds, with a 95% PPI of (44.3, 77.7). The homogeneous residual variance posterior mean was 339.4 with a 95% PPI of (324.4, 354.0) using MBAM; corresponding values using AM were 346.0 and (331.0, 360.8), respectively. The Pseudo Bayes Factor heavily favored the MBAM over the AM for both simulated and PWG data, thereby having important implications for improved precision on genetic merit predictions. The main advantage of MBAM is the flexibility in modeling heteroskedastic genetic variances of the breed composition groups, hence improving genetic predictions.

**Key Words:** Bayesian inference, Beef cattle, Crossbreeding, Genetic Evaluation, Multiple-breed, Post-weaning gain.

## 1. Introduction

Crossbreeding increases efficiency of livestock production by exploiting heterosis and complementarity between breeds (Gregory et al., 1999). Hence, an increasing proportion of the livestock populations are crossbred animals. Crossbreeding is synergistic with selection as factors to improve beef production. Furthermore, genetic trend is proportional to the accuracy of selection (Falconer and Mackay, 1996). For crossbred animals, this accuracy depends on properly specified genetic covariances between relatives (Fernando, 1999).

The genetic merit of an animal is comprised of the mean of its breed composition group plus its individually specific deviation from the group (Arnold et al., 1992; Elzo, 1994; Klei et al., 1996). Other than residual effects, individual deviations are due to random additive and non-additive genetic effects which can be estimated using phenotypic records on the individual and its relatives. In order to most efficiently use this data, it is imperative to properly model genetic covariances between relatives as specified by Lo et al. (1993) for an additive genetic deviations model. A model including fixed additive and non-additive genotypic effects with random additive individual deviations may be satisfactorily parsimonious for genetic evaluation of multiple-breed populations. A hierarchical Bayes model effectively combines data and prior information and provides a particularly useful framework for inference on additive breed and segregation variances.

The objectives of this study were to: 1) propose a hierarchical Bayes construction of the multiple-breed animal model from Lo et al. (1993) to estimate genotypic effects and individual additive deviations when breed and segregation variance components are unknown; 2) validate the model using simulated data, and to 3) apply the proposed model to a dataset of post-weaning gains on purebred and crossbred animals derived from Hereford and Nelore cattle and raised in diverse environments.

## 2. Material and Methods

### 2.1. Crossbreeding Model

*Genotypic effects.* Let $g$ denote a particular genotype (i.e. breed composition) composed of $B$ breeds. Also, let $f_b$ be the proportion of alleles from the $b$th breed ($b$=1,2,...,$B$), and $f_{bb'}$ be the probability that for a randomly chosen locus from an individual in $g$, one allele is derived from Breed $b$ and the other allele is derived from Breed $b'$, allowing for the possibility that $b=b'$ ($b'$=1,2,...,$B$). A general model is assumed for the genotypic effect; that is, the deviation $\delta_g$ of $g$ from the overall population mean $\mu$, based on the two-loci theory and absence of inbreeding was presented by Wolf et al. (1995):

$$\delta_g = \sum_{b=1}^{B} \gamma_{A_b} f_b + \sum_{b=1}^{B} \sum_{b'=b}^{B} \gamma_{D_{bb'}} f_{bb'} + \sum_{b=1}^{B} \sum_{b'=1}^{B} \gamma_{AA_{bb'}} f_b \, f_{b'}$$
$$+ \sum_{b=1}^{B} \sum_{b'=1}^{B} \sum_{b''=b'}^{B} \gamma_{AD_{b(b'b'')}} f_b \, f_{b'b''} + \sum_{b=1}^{B} \sum_{b'=b}^{B} \sum_{b''=1}^{B} \sum_{b'''=b''}^{B} \gamma_{DD_{(bb')(b''b''')}} f_{bb'} f_{b''b'''} \, , \qquad [1]$$

where $\gamma_{A_b}$ is the additive effect of Breed $b$, $\gamma_{D_{bb'}}$ is the dominance effect involving Breeds $b$ and $b'$, $\gamma_{AA_{bb'}}$ is the additive × additive effect involving Breeds $b$ and $b'$, $\gamma_{AD_{b(b'b'')}}$ is the additive × dominance effect involving the interaction of Breed $b$ with the dominance effect of Breeds $b'$ and $b''$ and $\gamma_{DD_{(bb')(b''b''')}}$ is the dominance × dominance effect involving the interaction between the dominance effect of Breeds $b$ and $b'$ and the dominance effect of Breeds $b''$ and $b'''$. The extension of [1] to include other effects (e.g., maternal breed effects) can naturally involve analogous terms. The coefficients required in [1] can be obtained from the parental generation for the animal as follows: $f_b = 0.5\left(f_b^s + f_b^d\right)$; $f_{bb} = f_b^s f_b^d$; $f_{bb'} = f_b^s f_{b'}^d + f_{b'}^s f_b^d$, for $b = 1,\ldots, B$; $b' = 1,\ldots, B$; $b < b'$. Here $s$ and $d$ denote paternal and maternal group, respectively.

In order ensure identifiability of the parameters in [1], restrictions must be invoked. Restrictions on the breed proportion coefficients are natural, namely $\sum_{b=1}^{B} f_b = 1$, $\sum_{b \le b'} f_{bb'} = 1$ and $f_b = f_{bb} + 0.5 \sum_{b'} f_{bb'}$. For the case of crosses involving only two breeds, we suggest the following parameter restrictions: $\gamma_{A_2} = 0$, $\gamma_{D_{11}} = \gamma_{D_{22}} = 0$, and $\gamma_{AA_{11}} = \gamma_{AA_{22}} = 0$, such that $\gamma_{AD_{1(11)}} = \gamma_{AD_{1(22)}} = \gamma_{AD_{2(11)}} = \gamma_{AD_{2(12)}} = \gamma_{AD_{2(22)}} = 0$, and $\gamma_{DD_{(11)(11)}} = \gamma_{DD_{(12)(22)}} = \gamma_{DD_{(11)(22)}} = \gamma_{D_{(12)(22)}} = \gamma_{DD_{(22)(22)}} = 0$. Then [1] simplifies to:

$$\delta_g = \gamma_{A_1} f_1 + \gamma_{D_{12}} f_{12} + \gamma_{AA_{12}} 2 f_1 \, f_2 + \gamma_{AD_{1(12)}} f_1 \, f_{12} + \gamma_{DD_{(12)(12)}} f_{12}^2, \qquad [2]$$

having five parameters and therefore requiring at least six breed composition groups to allow their estimability jointly with $\mu$. Here, [2] is simply a reparameterization of the model proposed by Hill (1982). Conceivably, genetic effects could interact with other non-genetic effects, such as age of dam, gender and region (Arthur et al., 1999; Klei et al., 1996), yielding a straightforward extension of Equation [1]. On the other hand, simpler models can be obtained by setting some effects of the general model in [1] equal to zero. For instance, a recombination loss model based on additive × additive epistasis (Dickerson, 1973; Kinghorn, 1980) is attained by letting all $\gamma_{AD_{b(b'b'')}}$ and $\gamma_{DD_{(bb')(b''b''')}}$ effects be equal to zero.

## 2.2. Hierarchical Bayes model construction

*First stage.* The first stage of the model specifies the conditional sampling density of the $n \times 1$ data vector $\mathbf{y} = \{y_j\}$. The component of this density for a record on individual $j$, is

$$y_j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{a}, \sigma_e^2 \sim N\left(\mathbf{x}'_{1j}\boldsymbol{\beta} + \mathbf{x}'_{2j}\boldsymbol{\gamma} + \mathbf{z}'_j\mathbf{a}, \sigma_e^2\right), \quad j \in S, \qquad [3]$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of non-genetic effects (e.g. gender, age of dam, contemporary groups, etc.), $\boldsymbol{\gamma}$ is a $t \times 1$ vector of genetic fixed effects with breed-specific components for $\gamma_{A_b}, \gamma_{D_{bb'}}, \gamma_{AA_{bb'}}, \gamma_{AD_{b(b'b'')}}$, and $\gamma_{DD_{(bb')(b''b''')}}$ as specified in Equation [1]; $\mathbf{a}$ is a $q \times 1$ vector of animal additive genetic effects; and $\mathbf{x}'_{1j}$, $\mathbf{x}'_{2j}$, and $\mathbf{z}'_j$ are known row incidence vectors, with the elements of $\mathbf{x}'_{2j}$ determined by the coefficients for genetic effects ($f_b$'s and $f_{bb'}$'s) as defined in Equation [1]. Moreover, $S$ represents the sample of size $n$ of animals having records; typically $n < q$ since $\mathbf{a}$ includes effects for ancestor animals without records. Finally $\sigma_e^2$ represents the residual variance, assumed to be homogeneous across breed groups.

*Second stage.* Prior densities are assumed for location parameters:

$$\boldsymbol{\beta} \mid \boldsymbol{\beta}_o, \mathbf{V}_\beta \sim N\left(\boldsymbol{\beta}_o, \mathbf{V}_{\boldsymbol{\beta}}\right), \qquad [4]$$

$$\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_o, \mathbf{V}_{\boldsymbol{\gamma}} \sim N\left(\boldsymbol{\gamma}_o, \mathbf{V}_{\boldsymbol{\gamma}}\right), \qquad [5]$$

and

$$\mathbf{a} \mid \boldsymbol{\varphi} \sim N\left(\mathbf{0}, \mathbf{G}\left(\boldsymbol{\varphi}\right)\right), \qquad [6]$$

where $\boldsymbol{\beta}_o$, $\gamma_o$, $\mathbf{V}_{\boldsymbol{\beta}}$ and $\mathbf{V}_{\boldsymbol{\gamma}}$ are specified. The additive genetic variance-covariance matrix $\mathbf{G}\left(\boldsymbol{\varphi}\right)$ is a function of more than one dispersion parameter in $\boldsymbol{\varphi}$ for crossbred populations as defined by Lo et al. (1993). Elements of $\mathbf{G}\left(\boldsymbol{\varphi}\right)$ can be computed by the tabular method having as diagonal elements:

$$\operatorname{Var}\left(a_j\right) = \sum_{b=1}^{B} f_b^j \sigma_{A_b}^2 + \sum_{b=1}^{B-1} \sum_{b'>b}^{B} 2\left(f_b^s f_{b'}^s + f_b^d f_{b'}^d\right) \sigma_{S_{bb'}}^2 + .5 \operatorname{cov}\left(a_j^s, a_j^d\right) \qquad [7]$$

for $j = 1,2,\ldots,q$. Here $a_j^s$ and $a_j^d$ represent, respectively, the additive genetic effect of the sire and the dam of $j$; $\sigma_{A_b}^2$ is the additive variance of breed $b$; and $\sigma_{S_{bb'}}^2$ is variance due to the segregation between breed $b$ and $b'$ or the additional variance observed in the $F_2$ generation over the $F_1$, such that $\boldsymbol{\varphi} = \left[\left\{\sigma_{A_b}^2\right\}_{b=1}^{B}, \left\{\sigma_{S_{bb'}}^2\right\}_{b=1, b'>b}^{B-1,\ B}\right]$ defines the components of genetic variance. Following Quaas (1988), Lo et al. (1993) showed that the inverse of $\mathbf{G}\left(\boldsymbol{\varphi}\right)$ can be computed using:

$$(\mathbf{G}(\boldsymbol{\varphi}))^{-1} = (\mathbf{I}-\mathbf{P})'(\boldsymbol{\Omega}(\boldsymbol{\varphi}))^{-1}(\mathbf{I}-\mathbf{P}),$$

where **I** is an identity matrix, **P** is a matrix relating progeny to parents and $\boldsymbol{\Omega}(\boldsymbol{\varphi})$ is a diagonal matrix with the *j*th diagonal element is defined as:

$$\omega_j = \mathrm{Var}\left(a_j\right) - .25\left(\mathrm{Var}\left(a_j^s\right) + \mathrm{Var}\left(a_j^d\right)\right) - .5\,\mathrm{cov}\left(a_j^s, a_j^d\right), \qquad [8]$$

which is a linear function of elements of $\boldsymbol{\varphi}$. For the case of non-inbred populations $\mathrm{cov}\left(a_j^s, a_j^d\right) = 0$ and [8] can be written as:

$$\begin{aligned}\omega_j =& \sum_{b=1}^{B} f_b^j \sigma_{A_b}^2 + \sum_{b=1}^{B-1}\sum_{b'>b}^{B} 2\left(f_b^s f_{b'}^s + f_b^d f_{b'}^d\right)\sigma_{S_{bb'}}^2 \\ &-0.25\left(\sum_{b=1}^{B} f_b^s \sigma_{A_b}^2 + \sum_{b=1}^{B-1}\sum_{b'>b}^{B} 2\left(f_b^{pgs} f_{b'}^{pgs} + f_b^{pgd} f_{b'}^{pgd}\right)\sigma_{S_{bb'}}^2\right) \\ &-0.25\left(\sum_{b=1}^{B} f_b^d \sigma_{A_b}^2 + \sum_{b=1}^{B-1}\sum_{b'>b}^{B} 2\left(f_b^{mgs} f_{b'}^{mgs} + f_b^{mgd} f_{b'}^{mgd}\right)\sigma_{S_{bb'}}^2\right),\end{aligned}$$

where *pgs*, *pgd*, *mgs*, and *mgd* represent, respectively, the paternal grandsire, paternal granddam, maternal grandsire and maternal granddam of individual *j*. In this case $(\mathbf{G}(\boldsymbol{\varphi}))^{-1}$ can be directly determined without computing $\mathbf{G}(\boldsymbol{\varphi})$.

*Third stage:* Scaled inverted chi-square prior densities are specified on the variance components as follows:

$$p\left(\sigma_e^2 \middle| \nu_e, s_e^2\right) \propto \left(\sigma_e^2\right)^{-\left(\frac{\nu_e+2}{2}\right)} \exp\left(-\frac{\nu_e s_e^2}{2\sigma_e^2}\right), \qquad [9]$$

$$p\left(\sigma_{A_b}^2 \middle| \nu_{A_b}, s_{A_b}^2\right) \propto \left(\sigma_{A_b}^2\right)^{-\left(\frac{\nu_{A_b}+2}{2}\right)} \exp\left(-\frac{\nu_{A_b} s_{A_b}^2}{2\sigma_{A_b}^2}\right), \; b=1,\ldots, B; \qquad [10]$$

$$p\left(\sigma_{S_{bb'}}^2 \middle| \nu_{S_{bb'}}, s_{S_{bb'}}^2\right) \propto \left(\sigma_{S_{bb'}}^2\right)^{-\left(\frac{\nu_{S_{bb'}}+2}{2}\right)} \exp\left(-\frac{\nu_{S_{bb'}} s_{S_{bb'}}^2}{2\sigma_{S_{bb'}}^2}\right), \; b=1,\ldots, B\text{-}1; b'=b+1,\ldots, B. [11]$$

Here, $\nu_e$, $s_e^2$, $\nu_{A_b}$, $s_{A_b}^2$, $b = 1,\ldots,B$, and $\nu_{S_{bb'}}$, $s_{S_{bb'}}^2$, $b = 1,\ldots,B\text{-}1$, $b' = b+1,\ldots,B$ are specified hyperparameters.

*Joint posterior density:* The joint posterior density is the product of [3], [4], [5], [6], [9], [10], and [11] and is given by:

$$p\left(\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{a},\boldsymbol{\varphi},\sigma_e^2 \mid \boldsymbol{\zeta},\mathbf{y}\right)=\prod_{j\in S} p\left(y_j \mid \boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{a},\sigma_e^2\right) p\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}_o,\mathbf{V}_{\boldsymbol{\beta}}\right) p\left(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_o,\mathbf{V}_{\boldsymbol{\gamma}}\right)$$
$$\times p(\mathbf{a} \mid \boldsymbol{\varphi})\, p\left(\sigma_e^2 \mid \nu_e,s_e^2\right)\prod_{b=1}^{B} p\left(\sigma_{A_b}^2 \mid \nu_{A_b},s_{A_b}^2\right)\prod_{b=1}^{B-1}\prod_{b'>b}^{B} p\left(\sigma_{S_{bb'}}^2 \mid \nu_{S_{bb'}},s_{S_{bb'}}^2\right),$$

where $\boldsymbol{\zeta}=\left(\boldsymbol{\beta}_o,\mathbf{V}_\beta,\boldsymbol{\gamma}_o,\mathbf{V}_\gamma,\nu_e,s_e^2,\nu_{A_1},\ldots,\nu_{A_B},s_{A_1}^2,\ldots,s_{A_B}^2,\nu_{S_{12}},\ldots,\nu_{S_{B-1B}},s_{A_{12}}^2,\ldots,s_{S_{B-1B}}^2\right)$ denotes the vector of hyperparameters.

*Full conditional densities:* The full conditional densities (FCD) of all unknown parameters/quantities or blocks thereof necessary to conduct MCMC inference are derived in the Appendix to this chapter.

### 2.3. Simulation Study

A crossbreeding system simulation study was conducted. Five datasets were generated, each based on two base purebred populations consisting of 240 animals per purebred. For each dataset, six sires and 20 dams from each purebred population were then randomly selected and mated *inter se* to produce purebred offspring, and to animals of the other breed to produce the $F_1$ generation. The $F_1$ animals were then randomly selected (12 sires and 40 dams) and mated *inter se* and to the base populations to produce $F_2$ and backcross animals, respectively. In a final stage, six sires and 20 dams from each of the backcross groups and 12 sires and 40 dams from the $F_2$'s were randomly selected and mated to produce an advanced generation of intercross animals with several different breed compositions. The total number of animals was set to 4,000, but the number of animals of each genotype was random and unbalanced within each dataset. There were approximately 300 animals for each purebred, 370 $F_1$'s, 450 $F_2$'s, 490 animals for each backcross, and 1,600 advanced intercross animals. Inbreeding was avoided in all matings. One record per animal was generated based on an arbitrarily chosen overall mean and fixed genotypic effect, a random additive genetic effect and a random normally, identically, and independently distributed residual effect. The additive genetic effect of an animal *j* was generated using [8] as $a_j = 0.5a_j^s + 0.5a_j^d + z_j\sqrt{\omega_j}$, where $z_j \sim N(0,1)$, $j$=1,2,...,$q$. The parental contributions from $a_j^s$ and $a_j^d$ are null when parents were not identified, as in the case of base population animals, such that $\omega_j$ is then the corresponding breed-specific additive genetic variance.

The genetic variance for Breed 1 ($\sigma_{A_1}^2$) was set to 100.0, for Breed 2 ($\sigma_{A_2}^2$) to 50.0, and for the segregation variance between Breeds 1 and 2 ($\sigma_{S_{12}}^2$) to 20.0. The residual variance ($\sigma_e^2$) was set to 100.0. To validate our Bayesian model choice criterion (see later), we also simulated five populations where $\sigma_{A_1}^2 = \sigma_{A_2}^2 = 75.0$, $\sigma_{S_{12}}^2 = 0.0$, and $\sigma_e^2 = 100.0$, this situation being equivalent to the conventional animal model assumption of no influence of crossbreeding on genetic variances.

Inference was based on MCMC and two different models: the Multiple-breed Animal Model (MBAM) described in this study and a conventional Animal Model (AM) that assumed equal breed genetic variances with no between-breed segregation variance. Uniform bounded priors were utilized for the variance components. The length of chain was $G$ = 60,000 cycles after a burn-in period comprising 10,000 cycles. For each dispersion parameter, the initial monotone sequence approach (Geyer, 1992) was used to calculate effective sample sizes (ESS), which estimates the number of independent samples with information content equivalent to that contained within the 60,000 dependent samples (Sorensen et al., 1995).

Various measures were used to compare MBAM and AM in terms of genetic merit prediction including the mean squared error of prediction (MSEP), the mean bias of prediction (MBIAS) and Spearman rank correlations between estimated and true genetic values. The MSEP for each model was determined as $\sum_{h=1}^{5}\sum_{j=1}^{q}\left(\hat{a}_{hj}-a_{hj}\right)^2 \Big/ q \Big/ 5$, where 5 denotes the number of replicates, $q$ = 4000 is the total number of animals per replicate, $\hat{a}_{hj}$ is the estimated direct additive genetic effect for animal $j$ in the replicate $h$ and $a_{hj}$ is the true value of $\hat{a}_{hj}$. MBIAS was similarly determined as $\sum_{h=1}^{5}\sum_{j=1}^{q}\left(\hat{a}_{hj}-a_{hj}\right) \Big/ q \Big/ 5$.

*Model choice criterion.* We considered the *Pseudo Bayes Factor* (PBF) (Gelfand, 1996) as the basis for model choice. The PBF involves the evaluation of the first stage density in [3] on each MCMC sample. For comparing MBAM and AM, the corresponding PBF is determined to be:

$$PBF_{MBAM,AM} = \prod_{j \in S} \frac{p\left(y_j \mid \mathbf{y}_{(-j)}\right)_{MBAM}}{p\left(y_j \mid \mathbf{y}_{(-j)}\right)_{AM}},$$

where $p\left(y_j \mid \mathbf{y}_{(-j)}\right)_{MBAM}$ and $p\left(y_j \mid \mathbf{y}_{(-j)}\right)_{AM}$ are model specific conditional predictive ordinates (CPO) for observation $y_j$. A MCMC approximation for the CPO of model $M$ is obtained by a harmonic mean:

$$p\left(y_j \mid \mathbf{y}_{(-j)}\right)_M \approx \frac{1}{\frac{1}{G}\sum_{l=1}^{G} p^{-1}\left(y_j \mid \boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}^{(l)}, \mathbf{a}^{(l)}, \sigma_e^{2(l)}\right)_M},$$

where $\boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}^{(l)}, \mathbf{a}^{(l)}$ and $\sigma_e^{2(l)}$ are the post burn-in MCMC samples for $\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{a}$ and $\sigma_e^2$, respectively, $l = 1,2,\ldots,G$. An overall PBF across the five replicates was obtained by $\overline{PBF}_{MBAM,AM} = \exp\left(\sum_{h=1}^{5} \log\left(PBF_{MBAM,AM}\right)_h / 5\right)$.

### 2.4. Application to field data

We applied the model described in this study (MBAM) to the analysis of post-weaning gain (PWG) records of a beef cattle population under genetic evaluation in Brazil and consisting of Hereford and Nelore cattle and their various crosses. We also applied a conventional AM to compare results with those from the proposed MBAM. As in the simulation study, MCMC was used for inference in both models.

The records were collected between 1974 and 2000 by the Brazilian National Breeders Association and Gensys Associated Consultants within a large-scale genetic evaluation program called "Delta G Connection". After deleting contemporary groups with less than ten animals and sires with less than five offspring, there were 22,717 records of PWG from 15 different herds with a total of 40,082 animals in the pedigree file. Animals were raised on extensive pasture conditions in three different regions, of which two were in the tropical area and one in the sub-tropical area of the Country. Region 1 comprised of two farms located between 14°S and 16°S latitude with 5,410 records (23.8%), Region 2 had three farms located between 21°S and 23°S with 3,110 records (13.7%), and Region 3 had ten farms located between 30°S and 32°S with 14,197 records (62.5%). The average PWG was 98.2 kg ± 41.2. The age of calves at weaning ranged from 114 to 300 days (208 days on average) and the average post-weaning test period was 280 days, ranging from 106 to 483 days.

Animals with records had breed compositions that ranged from purebred Hereford to 7/8 Nelore; however, purebred Herefords and $F_1$'s provided about 90% of the records (Table 3.1). There were no purebred Nelore animals with records. Dams were mostly represented by purebreds of the two breeds (Table 3.1). Most animals with records in the tropical part of the Country (Regions 1 and 2) were $F_1$'s. This structure is due to the fact that these herds belonged to Hereford breeders from southern Brazil and Nelore breeders from central Brazil cooperating on the production of crossbred Braford animals.

Non-genetic effects included in both models (elements of $\beta$) were the main effects of region, gender, length of the post-weaning test period, and linear and quadratic age of dam effects, the latter being included to model possible compensatory growth due to age of dam. The elements of $\gamma$ were specified based on the 'epistatic loss' model (Kinghorn, 1980) for a two-breed scenario; i.e. Equation [2] with $\gamma_{AD_{1(12)}} = \gamma_{DD_{(12)(12)}} = 0$. The fixed effects portion of the model was further augmented to allow for interaction between gender, length, and age of dam polynomials with breed proportion. Region by breed proportion interaction was not modeled due to multicollinearity problems since nearly all animals with records in Regions 1 and 2 were $F_1$'s (Table 3.1). Contemporary groups (herd, year, season and management subclasses) were modeled as uncorrelated random effects. Due to lack of objective prior information on this population, we adopt bounded uniform priors on $\beta$ and $\gamma$, and conjugate relatively noninformative specifications on variance components, specifically, $\nu_{A_1} = \nu_{A_2} = \nu_{S_{12}} = \nu_e = 5$; $s_{A_1}^2 = s_{A_2}^2 = 80$; $s_{S_{12}}^2 = 10$,

Table 3.1. Distribution of post-weaning gain records per region according to individual and maternal breed composition

| Breed composition | Individual |  |  | Maternal |  |  |
|---|---|---|---|---|---|---|
|  | Region 1 | Region 2 | Region 3 | Region 1 | Region 2 | Region 3 |
| Nelore | 0 | 0 | 0 | 5,347 | 3,001 | 0 |
| $BC_{(Nelore)}$ | 0 | 0 | 91 | 0 | 0 | 2 |
| $F_1$(H×N) | 5,346 | 2,997 | 0 | 63 | 108 | 0 |
| $F_1$(N×H) | 0 | 0 | 375 | 0 | 0 | 994 |
| $F_2$ | 0 | 18 | 35 | 0 | 0 | 0 |
| 3/8Nelore | 0 | 79 | 1,006 | 0 | 0 | 405 |
| $BC_{(Hereford)}$ | 8 | 12 | 560 | 0 | 0 | 26 |
| Hereford | 0 | 0 | 11,660 | 0 | 0 | 12,317 |
| Others | 56 | 4 | 470 | 0 | 1 | 453 |

and $s_e^2 = 350$.

The length of the MCMC chain for PWG was 200,000 cycles after burn-in for both MBAM and AM. Posterior means, modes, key percentiles and standard deviations of the parameters were obtained from their marginal posterior densities.

## 3. Results and Discussion

### 3.1. Simulation study

Posterior means, modes, standard deviations, and 95% confidence sets for variance components averaged over the five simulated populations as obtained by MCMC using MBAM or AM are presented in Table 3.2. The average posterior mean and mode of all variance components obtained by MBAM appeared to indicate these point estimates as being essentially unbiased. The individual 95% confidence sets included the true parameter value in 19 out of 20 cases (based on four variance components estimated

Table 3.2. True value, posterior mean (PMEAN), posterior standard deviation (PSD), posterior mode (PMODE), 95% posterior probability intervals (PPI), and effective sample size (ESS) for variance components (VC) averaged over the five simulated populations, obtained by a conventional animal model and by a multiple-breed animal model

| VC[a] | True | PMEAN | PSD | PMODE | PPI | ESS |
|---|---|---|---|---|---|---|
| *Conventional animal model* | | | | | | |
| $\sigma_A^2$ | - | 84.33 | 9.71 | 83.16 | (66.59, 104.46) | 856 |
| $\sigma_e^2$ | 100.00 | 105.11 | 5.94 | 105.58 | (93.18, 116.47) | 1,108 |
| *Multiple-breed animal model* | | | | | | |
| $\sigma_{A_1}^2$ | 100.00 | 106.42 | 14.11 | 105.53 | (80.04, 135.25) | 1,245 |
| $\sigma_{A_2}^2$ | 50.00 | 48.67 | 10.43 | 46.95 | (30.17, 70.90) | 1,046 |
| $\sigma_{S_{12}}^2$ | 20.00 | 21.34 | 8.24 | 19.63 | (6.86, 38.56) | 1,205 |
| $\sigma_e^2$ | 100.00 | 100.20 | 6.09 | 99.98 | (88.04, 111.93) | 1,003 |

[a] $\sigma_A^2$ = additive genetic variance, $\sigma_e^2$ = residual variance, $\sigma_{A_1}^2$ = additive genetic variance for breed 1, $\sigma_{A_2}^2$ = additive genetic variance for breed 2, $\sigma_{S_{12}}^2$ = additive genetic variance due to segregation between breeds 1 and 2.

within each of five populations), thereby falling within probabilistic expectation. The single additive genetic variance estimated using AM was slightly greater than the average of the two true values for the breed-specific genetic variances, this deviation being partly due to the non-zero between-breed segregation variance. Considering the complexity of the multiple-breed population structure and the sample sizes in this study, the results presented in Table 3.2 indicate that MBAM based on MCMC provides a reliable procedure to estimate additive genetic variance components on populations consisting of crossbred animals.

The PBF was used to compare the MBAM versus AM for fit of each simulated dataset. The PBF for MBAM/AM varied from $9.785 \times 10^2$ to $3.337 \times 10^6$ in the five replicates of the two-breed population, being always in favor of MBAM and with an overall value of $2.543 \times 10^4$. However, when the five data sets with a purebreeding assumption on genetic variances were analyzed, the PBFs were either in favor of AM or inconclusive; the range of values was from $1.163 \times 10^{-5}$ to 5.472 and the overall PBF was $6.073 \times 10^{-2}$, being in favor of AM. Hence, the PBF was either able to correctly choose the right model with certainty or, at the very least, did not ever decisively choose the wrong model.

In addition, we compared the models in terms of prediction of additive genetic effects. The overall mean squared error of prediction (MSEP), averaged across all animals and the five simulated populations, was $34.92 \pm 0.45$ for the AM and $34.61 \pm 0.45$ for the MBAM. Also, the overall mean bias (MBIAS) was $0.23 \pm 0.28$ and $0.24 \pm 0.27$, respectively for AM and MBAM and the overall Spearman rank correlation (RANK) between predicted and true genetic values was $0.75 \pm 0.01$ for AM and $0.75 \pm 0.01$ for MBAM. These same comparisons were also considered within breed composition groups (data not shown) with results not clearly pointing out advantages for either model.

The main advantage of MBAM over AM appears to be in terms of accounting for genotype differences in genetic variability. To further illustrate this point, we present in Table 3.3 the true and the estimated additive genetic variance ($\sigma_A^2$) for different breed composition groups. With MBAM, the genetic variance of each breed composition group is a function of breed specific variances and the segregation variance; for example, the

Table 3.3. Empirical averages and standard errors (SE) of posterior mean (PMEAN) additive genetic variances, obtained by a multiple-breed animal model for different breed compositions in the simulation study. For all breed compositions, the corresponding posterior mean and standard deviation using the conventional animal model (AM) was 84.3 ± 3.3

| Breed composition | True value | PMEAN ± SE[a] |
|---|---|---|
| $P_1$ | 100.0 | 106.4 ± 5.4 |
| $BC_1$ | 97.5 | 102.7 ± 2.8 |
| $F_1$ | 75.0 | 77.5 ± 3.1 |
| $F_2$ | 95.0 | 98.9 ± 4.1 |
| $BC_2$ | 72.5 | 73.8 ± 4.1 |
| $P_2$ | 50.0 | 48.7 ± 4.8 |

[a] Based on five simulated two-breed populations.

genetic variance of the $F_2$ groups $\sigma^2_{A_{F_2}}$ is obtained by $0.5\sigma^2_{A_1} + 0.5\sigma^2_{A_2} + \sigma^2_{S_{12}}$ and, in general, for genotype $g$ by $\sigma^2_{A_g} = \sum_{b=1}^{B} f_b^g \sigma^2_{A_b} + \sum_{b=1}^{B-1}\sum_{b'>b}^{B} 2\left(f_b^s f_{b'}^s + f_b^d f_{b'}^d\right)\sigma^2_{S_{bb'}}$ (Lo et al., 1993), whereas a common genetic variance is attributed to all breed compositions in AM. The genetic variances estimated by MBAM were always closer to the true value compared to AM, for all considered genotypes. For example, the parental groups ($P_1$ and $P_2$) posterior means for genetic variances, $\sigma^2_{A_1}$ and $\sigma^2_{A_2}$, were respectively 9.3% and 66.0% closer to the true values when estimated by MBAM compared to AM (Table 3.3). It is clear then that MBAM adequately characterizes the heterogeneous genetic variability due to different breed compositions in the simulated crossbred populations.

### 3.2. Post-weaning gain analysis

*Model choice.* For the analysis of PWG in the Nelore-Hereford cross dataset, the PBF comparing MBAM/AM was $1.152\times10^4$. This implies that the marginal probability of data $p(\mathbf{y}\mid M_r)$ obtained by MBAM was more than 10,000 times larger than that obtained by AM, suggesting then that the MBAM is a decisively better fit.

*Genotypic effects.* Posterior means and standard deviations of genetic fixed effects on PWG obtained by MBAM and AM using Kinghorn's epistatic loss parameterization (Kinghorn, 1980) are presented in Table 3.4. Inferences obtained using both models were similar, and no significant differences could be detected between MBAM and AM in terms of posterior means of genetic effects (Table 3.4). Mean PWG decreased as the Nelore proportion increased. As expected, dominance favorably affected PWG while additive × additive interaction, i.e. recombination loss, adversely affected PWG. We also attempted to fit the two-loci model (Hill, 1982; Wolf et al., 1995); however, this fit was not successful due to extremely high correlations between coefficients of genetic effects: ranging from 0.92 between additive × additive $(2f_1f_2)$ and dominance × dominance $(f_{12}^2)$ to a maximum of 0.99 between $f_{12}^2$ and additive × dominance $(f_1\,f_{12})$ coefficients. A similar problem was observed by Birchmeier et al. (2002), who eventually decided upon a model with only additive and dominance effects but no epistatic effects. Nonadditive genetic fixed effects are generally difficult to estimate using field data because of confounding and multicollinearity (Birchmeier et al., 2002; Klei et al., 1996), particularly between various epistatic effects. Consequently, genetic fixed effects in proposed multiple-breed genetic evaluation systems have only generally included

Table 3.4. Posterior means (PMEAN) ± posterior standard deviations (PSD) in kg of genetic fixed effects obtained by a multiple-breed animal model (MBAM) and an animal model (AM) for post weaning gain in Nelore-Hereford crosses

| Effect (Parameter) | MBAM | AM |
|---|---|---|
| | PMEAN ± PSD | PMEAN ± PSD |
| Additive $\left(\gamma_{A_1}\right)$ | -28.30 ± 9.95 | -30.95 ± 10.25 |
| Dominance $\left(\gamma_{D_{12}}\right)$ | 37.10 ± 5.00 | 35.77± 4.91 |
| Additive × additive $\left(\gamma_{AA_{12}}\right)$ | -31.56 ± 9.05 | -29.05 ± 8.72 |

additive and dominance effects (Cunningham, 1987; Klei et al., 1996; Miller and Wilton, 1999; Sullivan et al., 1999). Prior information on genetic fixed effects, as it is available from the literature, might be useful for analyses of poorly structured datasets as is common for crossbred beef cattle (Quaas and Pollak, 1999). The use of prior density specifications on elements of $\gamma$ further mitigates the effects of multicollinearity amongst genetic effects coefficients. Reliable estimates of dominance effects are available from the literature (e.g. Gregory et al., 1999); however, reliable estimates of epistatic effects are lacking (Arthur et al., 1999; Koch et al., 1985). Simple recombination loss specifications for $\gamma$ (Dickerson, 1973; Kinghorn, 1980; Kinghorn, 1987) provide a useful compromise between the two-loci model (Hill, 1982; Wolf et al., 1995) and currently used additive/dominance models in that epistatic effects are modeled with only one parameter in a two breed population as we have done.

The posterior means and 95% PPI for breed group effects as obtained by MBAM for the most frequent genotypes with records (Hereford, Backcrosses, $F_1$, $F_2$ and 3/8 Nelore) in Region 3 (between 30°S and 32°S) are presented in Figure 3.1. These means were very similar under MBAM and AM, varying from a minimum difference of 0.10 kg for

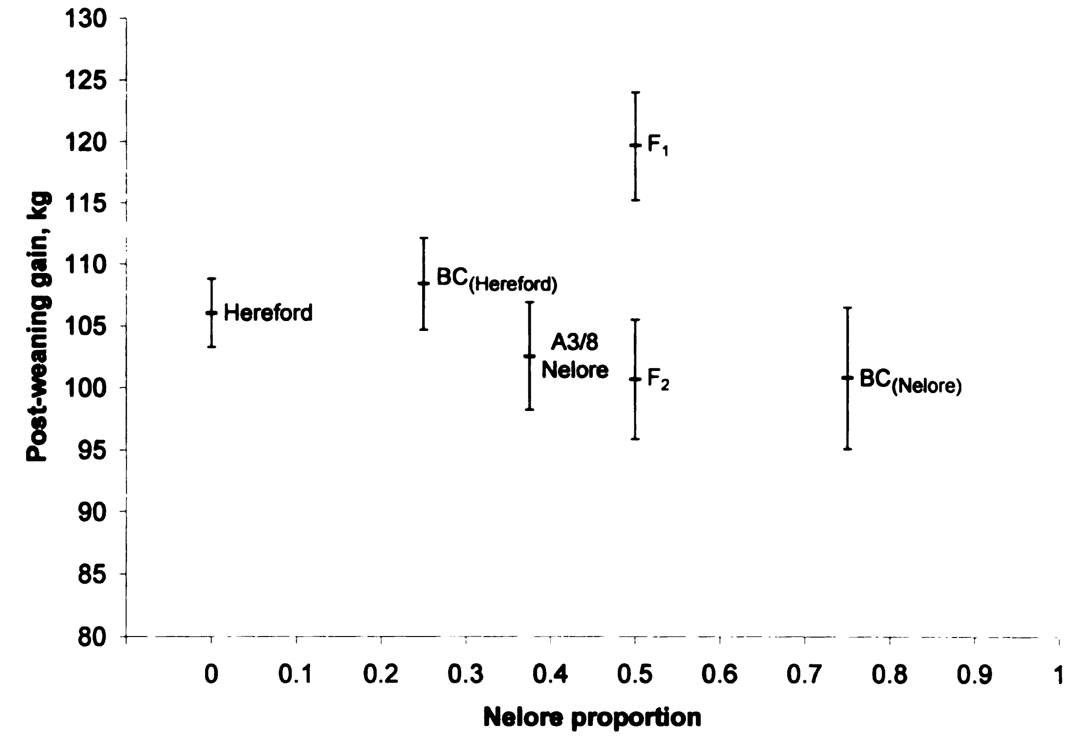

Figure 3.1. Posterior means (intermediate tick mark) and 95% posterior probability intervals (end tick marks) of post-weaning gain of Hereford, Hereford backcross ($BC_{(Hereford)}$), advanced 3/8 (A3/8) Nelore 5/8 Hereford, $F_1$, and $F_2$ and Nelore backcross ($BC_{(Nelore)}$) calves obtained by a Multiple-Breed Animal Model (MBAM)

Herefords to a maximum difference of 1.00 kg observed on $F_1$'s. The same mean differences would be observed in the other two regions since interaction between region and breed composition effects were not modeled.

*Additive genetic variances and heritabilities.* Variance components for post-weaning gain (PWG) estimated by MBAM and AM are presented in Table 3.5. The genetic variances obtained for the Nelore and Hereford breeds by MBAM differed substantially in magnitude. Herefords had a posterior mean genetic variance that was almost fourfold that obtained for the Nelore breed with no apparent overlap between the 95% posterior

Table 3.5. Posterior mean (PMEAN), posterior standard deviation (PSD), posterior mode (PMODE), 95% posterior probability intervals (PPI), and effective sample size (ESS) of variance components (VC) estimated for post-weaning gain in Nelore-Hereford crosses, obtained by a conventional animal model and by a multiple-breed animal model

| VC[a] | PMEAN | PSD | PMODE | PPI | ESS |
|---|---|---|---|---|---|
| *Conventional animal model* | | | | | |
| $\sigma_A^2$ | 60.46 | 8.65 | 61.01 | (44.31, 77.69) | 225 |
| $\sigma_{cg}^2$ | 900.47 | 43.40 | 906.53 | (818.11, 988.72) | 15,389 |
| $\sigma_e^2$ | 345.99 | 7.60 | 346.80 | (331.03, 360.79) | 291 |
| *Multiple-breed animal model* | | | | | |
| $\sigma_{A_1}^2$ | 23.80 | 6.86 | 22.44 | (13.02, 39.52) | 171 |
| $\sigma_{A_2}^2$ | 85.17 | 11.28 | 84.11 | (63.17, 108.45) | 223 |
| $\sigma_{A_{12}}^2$ | 8.42 | 5.84 | 4.71 | (2.32, 24.75) | 2,436 |
| $\sigma_{cg}^2$ | 897.98 | 43.31 | 888.59 | (816.69, 986.13) | 13,685 |
| $\sigma_e^2$ | 339.39 | 7.51 | 339.59 | (324.35, 353.98) | 259 |

[a] $\sigma_A^2$ = additive genetic variance, $\sigma_{cg}^2$ = variance due to contemporary groups, $\sigma_e^2$ = residual variance, $\sigma_{A_1}^2$ = additive genetic variance for Nelore breed, $\sigma_{A_2}^2$ = additive genetic variance for Hereford breed, $\sigma_{A_{12}}^2$ = additive genetic variance due to segregation between breeds.

probability intervals (PPI) obtained for these parameters (Table 3.5). Using the conventional AM, a posterior mean intermediate to the Nelore and Hereford posterior mean genetic variances in the MBAM was obtained, as shown in the posterior densities of these variances (Figure 3.2). The posterior mean of the variance due to the segregation between the Hereford and Nelore breeds had a magnitude of about 35.4% of the Nelore genetic variance posterior mean, but represented only 9.9% of Hereford genetic variance posterior mean (Table 3.5). These percentages were larger than those found for birth and weaning weight of crosses of Angus and Brahman in Florida, ranging from 1.4 to 3.1%

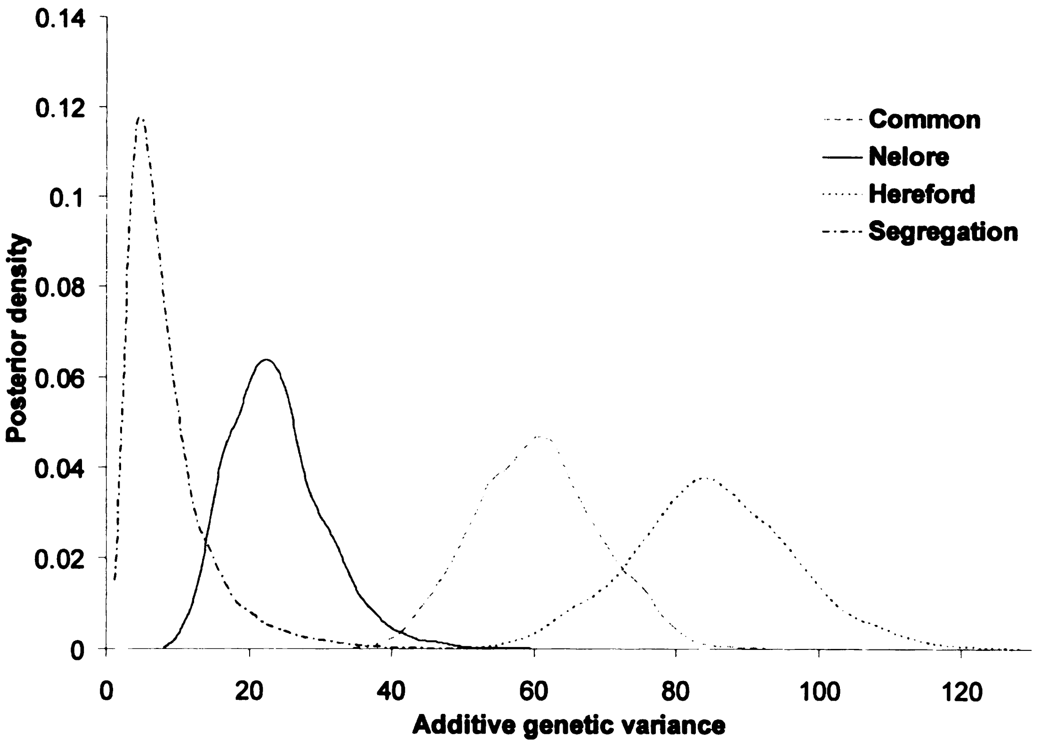

Figure 3.2. Posterior densities of the additive genetic breed and segregation variances obtained by a Multiple-Breed Animal Model (MBAM) and of the homogeneous additive genetic variance (assumed common to both breeds) obtained by a conventional Animal Model (AM) for post-weaning gain in Nelore-Hereford crosses

(Elzo and Wakeman, 1998). The magnitude of the segregation variance relative to the Hereford genetic variance (9.9%) was, however, somewhat smaller compared to results obtained for birth weight of Hereford-Nelore crosses in Argentina (16.5%) (Birchmeier et al., 2002). The Nelore variance was 73.5% of the magnitude of the Hereford variance in birth weight in their study, whereas Nelores had a genetic variance for PWG that was only 27.9% that for Herefords (Table 3.5) in our study. Heritabilities for the most prevalent genotypes in the studied population are presented in Table 3.6. Heritabilities were determined by not including the contemporary group variance as part of the

Table 3.6. Posterior means (PMEAN), posterior standard deviations (PSD), posterior modes (PMODE), 95% posterior probability intervals (PPI), and effective sample size (ESS) of direct additive heritability of post-weaning gain (PWG) for different Nelore-Hereford genotypes, obtained by a conventional animal model and by a multiple-breed animal model

| Breed composition | PMEAN | PSD | PMODE | PPI | ESS |
|---|---|---|---|---|---|
| *Conventional animal model* | | | | | |
| Overall | 0.15 | 0.02 | 0.15 | (0.11, 0.19) | 225 |
| *Multiple-breed animal model* | | | | | |
| Nelore | 0.07 | 0.02 | 0.06 | (0.04, 0.11) | 168 |
| $BC_{(Nelore)}$ | 0.11 | 0.02 | 0.11 | (0.08, 0.15) | 188 |
| $F_1$ | 0.14 | 0.02 | 0.14 | (0.10, 0.18) | 179 |
| $F_2$ | 0.16 | 0.02 | 0.15 | (0.12, 0.20) | 251 |
| A3/8-Nelore[a] | 0.17 | 0.02 | 0.17 | (0.13, 0.22) | 244 |
| $BC_{(Hereford)}$ | 0.18 | 0.02 | 0.18 | (0.14, 0.22) | 214 |
| Hereford | 0.20 | 0.02 | 0.20 | (0.15, 0.25) | 219 |

[a] Advanced generation of 3/8 Nelore and 5/8 Hereford composition.

phenotypic variance; this was intended to make our estimates comparable with other within-herd heritability estimates from the literature where contemporary groups are often fitted as fixed effects. (Eler et al., 1995; Koots et al., 1994; Meyer, 1992). The posterior mean of the additive heritability ($h_A^2$) obtained by MBAM were between a minimum of 0.07 for purebred Nelores and a maximum of 0.20 for purebred Herefords, with other genotypes having intermediate values. As expected, the posterior mean $h_A^2$ under the AM had an intermediate value (0.15). These estimates are considerably smaller the average value 0.31 determined for PWG as based on 177 studies (Koots et al., 1994). The Hereford $h_A^2$ estimate is, however, within the expected range for the extensive production environments in Brazil and is larger than the heritability estimate of 0.16 for yearling weight of Hereford cattle raised on pasture production systems in Australia (Meyer, 1992). The posterior mean obtained for $h_A^2$ for Nelores was very low and was less than the corresponding estimate of 0.16 observed for yearling weight in another Brazilian Nelore purebred dataset (Eler et al., 1995). Since purebred Nelores were only represented by parents without records in our population, it may be difficult to compare these two different sets of results due to, for example, different management systems.

Despite the availability of methodology to estimate multiple-breed additive genetic variances (Birchmeier et al., 2002; Elzo, 1994), several recently proposed models (Klei et al., 1996; Miller and Wilton, 1999; Quaas and Pollak, 1999; Sullivan et al., 1999) assume that all breeds have the same additive genetic variance and that there is no genetic variance due to segregation between breeds in advanced crosses. One advantage of the Bayesian MBAM over the specifications of Elzo (1994) and Birchmeier et al. (2002) is the ability to incorporate prior information on breed specific and segregation genetic variance components. Even though existing prior information for segregation variances is limited (Birchmeier et al., 2002; Elzo and Wakeman, 1998), there is extensive information available on breed specific variances (e.g. Koots et al., 1994; Meyer, 1992) that could be incorporated using [10].

*Animal additive genetic effects.* Ranking animals for genetic merit and eventual selection is a chief objective in breeding programs. The Spearman rank correlation between additive genetic effects obtained by MBAM and AM overall genotypes and for the most frequent genotypes in the dataset, are presented in Table 3.7. Rank correlations,

Table 3.7. Spearman rank correlation between posterior mean of additive genetic effects obtained by a multiple-breed animal model and by a conventional animal model for all animals and for different percentile MBAM groupings of animals within the most frequent breed compositions.

| Breed composition | N | Spearman Rank Correlation | | |
|---|---|---|---|---|
| | | 100% | Top 10% | Top 5% |
| Overall | 40,082 | 0.977 | 0.884 | 0.835 |
| Nelore | 7,445 | 0.995 | 0.970 | 0.972 |
| $BC_{(Nelore)}$ | 91 | 0.984 | 0.933 | 0.800 |
| $F_1(H\times N)$ | 8,343 | 0.987 | 0.894 | 0.891 |
| $F_1(N\times H)$ | 375 | 0.975 | 0.951 | 0.946 |
| $F_2$ | 53 | 0.968 | 0.900 | 1.000 |
| A3/8- Nelore[a] | 337 | 0.996 | 0.985 | 0.975 |
| $BC_{(Hereford)}$ | 580 | 0.995 | 0.916 | 0.906 |
| Hereford | 19,976 | 0.994 | 0.948 | 0.941 |

[a] Advanced generation of 3/8 Nelore and 5/8 Hereford composition.

across all animals and within each genotype were very high, being always greater than 0.96. However, if consideration is limited to the top 10% and top 5% animals ranked by MBAM, the rank correlation for genetic merit across all genotypes decreased, respectively, to 0.88 and to 0.84. Hence, substantial differences exist between MBAM and AM models when selecting top animals as breedstock. The observed decrease in rank correlations within genotype from using all animals to the top 10% and the top 5% was generally not as sharp as when we considered all groups together (Table 3.7). It appears that ranks are less affected within genotypes than across all animals being likely due to the different genetic variances (and consequently dispersion of genetic effects) that are specified for each breed composition group under MBAM as opposed to the conventional

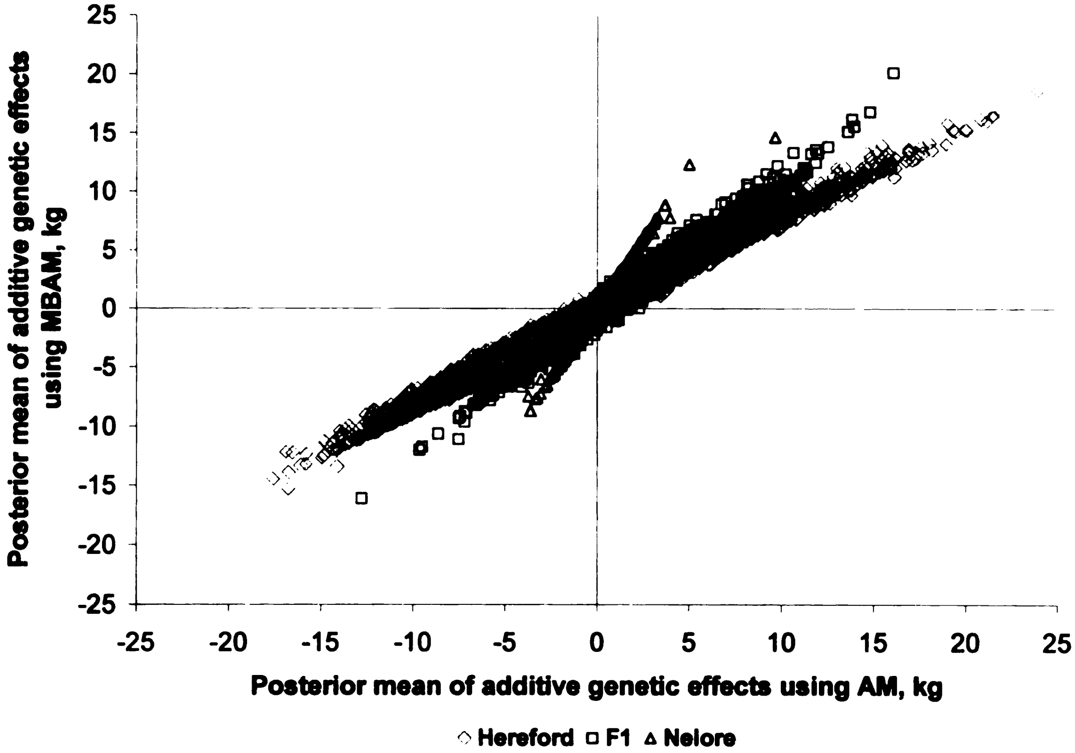

Figure 3.3. Scatter plot of posterior means of additive genetic effects of Herefords, Nelores and $F_1$'s obtained by a Multiple-Breed Animal Model (MBAM) versus those obtained by a conventional Animal Model (AM)

AM. The scatter plots in Figure 3.3 of the posterior means of additive genetic effects for MBAM versus AM for Hereford, Nelore and $F_1$'s provide additional evidence for the difference between models in terms of accommodating the breed composition specific variability of genetic effects; that is, the MBAM is much more flexible in accommodating the different variability observed on diverse genotypes. There was also some difference between models in terms of posterior standard deviations of animal genetic values as can be observed for Hereford, Nelore and $F_1$'s in Figure 3.4. Posterior standard deviations would be used for computing accuracy of the genetic evaluations in Figure 3.3.

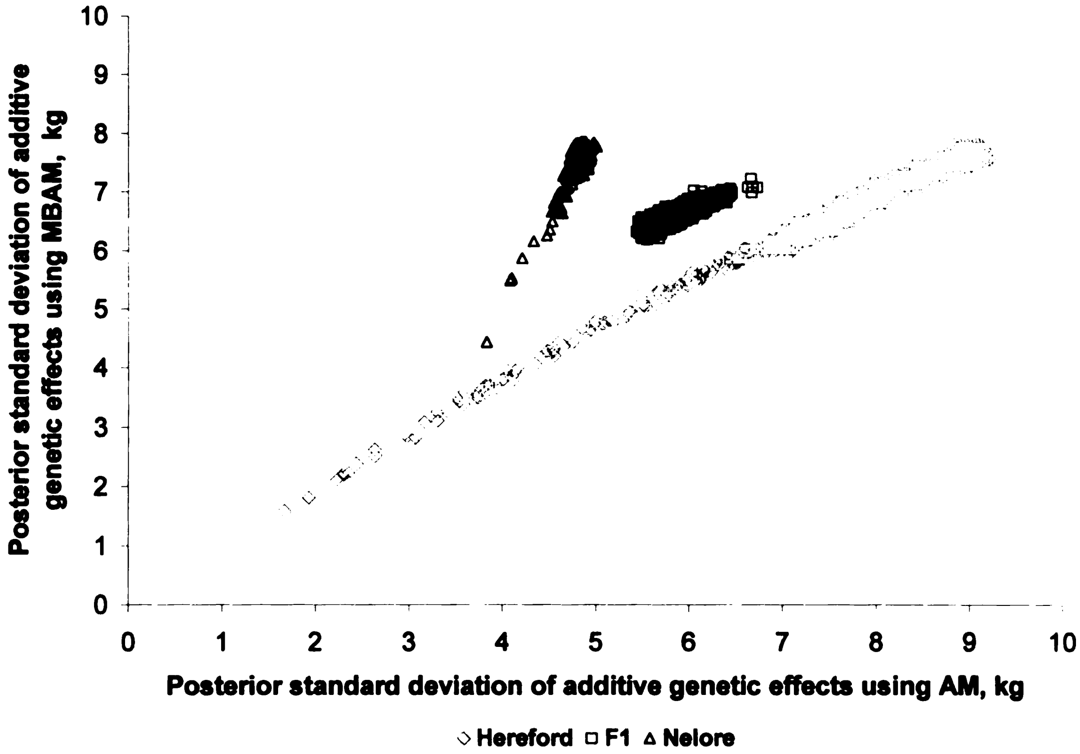

Figure 3.4. Scatter plot of posterior standard deviations of additive genetic effects of Herefords, Nelores and F$_1$'s obtained by a Multiple-Breed Animal Model (MBAM) versus those obtained by a conventional Animal Model (AM)

Expected progeny differences (EPD) in a multiple-breed scenario are a function of fixed and random genetic effects (Arnold et al., 1992; Elzo, 1994; Klei et al., 1996; Sullivan et al., 1999). The coefficients for the fixed genetic effects (additive, dominance, etc.) will depend on the mate's genotype and, therefore, comparison between candidates for selection should be made for specific breed compositions of the mates. The additive genetic effect corresponds to the general combining ability of the individual and does not depend on the genotype of the mates. The determination of specific combining abilities requires the estimation of non-additive genetic variances. Even though theory for a full additive and dominance two-breed genetic model is available (Lo et al., 1995), this model requires a much larger number of variance components to be estimated (up 25 when inbreeding is present) and thus may be cumbersome for practical applications.

Generalization of the model applied in this study for the case of multiple-traits or additive-maternal genetic effects could be attained by using multiple-breed variance-covariance genetic matrices (Cantet and Fernando, 1995) and a Wishart proposal density in the Metropolis-Hastings algorithm.

Due the computational limitations of MCMC, MBAM could be implemented for genetic evaluation of large beef cattle populations using an empirical Bayes approach. In this case, variance components could be previously estimated for a subset of the data on the population of interest using MCMC. The mixed model equations as from [A1] in the Appendix can then be used to provide empirical best linear unbiased estimates (BLUE) for elements of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and BLUP of $\mathbf{a}$.

### 4. Implications

The multiple-breed animal model specifies the additive genetic variance of each breed composition group as function of breed-specific and segregation variances, thereby sufficiently characterizing the genetic heteroskedasticity of these groups in crossbred populations. In contrast, the standard animal model assumes constant genetic variances across groups and no segregation variance. Accordingly, the proposed hierarchical model enhances flexibility for modeling the dispersion of genetic merit within breed groups, thereby having important implications for improved precision on prediction of genetic merit. This advantage increases with increasing differences in breed-specific variances and with the magnitude of the segregation variances.

## Appendix

*Fully Conditional Densities (FCD)*

In what follows the FCD are present using the notation "*ELSE*" to denote the data vector **y** and all other parameters treated as known in the FCD in question.

*Fixed and random location parameters.* Let

$$\mathbf{C}=\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1+\mathbf{V}_{\boldsymbol{\beta}}^{-1}\sigma_e^2 & \mathbf{X}_1'\mathbf{X}_2 & \mathbf{X}_1'\mathbf{Z}\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2+\mathbf{V}_{\boldsymbol{\gamma}}^{-1}\sigma_e^2 & \mathbf{X}_2'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X}_1 & \mathbf{Z}'\mathbf{X}_2 & \mathbf{Z}'\mathbf{Z}+\left(\mathbf{G}(\boldsymbol{\varphi})\right)^{-1}\sigma_e^2\end{bmatrix}^{-1},\text{ and }\mathbf{r}=\begin{bmatrix}\mathbf{X}_1'\mathbf{y}+\mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_o\sigma_e^2\\ \mathbf{X}_2'\mathbf{y}+\mathbf{V}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\gamma}_o\sigma_e^2\\ \mathbf{Z}'\mathbf{y}\end{bmatrix},$$

where $\mathbf{X}_1=\{\mathbf{x}_{1j}'\}$, $\mathbf{X}_2=\{\mathbf{x}_{2j}'\}$, and $\mathbf{Z}=\{\mathbf{z}_j'\}$, $j\in S$. Following Wang et al. (1994b), it can be shown the location parameters have the following multivariate normal distribution,

$$[\boldsymbol{\beta}' \quad \boldsymbol{\gamma}' \quad \mathbf{a}']' \mid ELSE \sim N\left(\left[\hat{\boldsymbol{\beta}}' \quad \hat{\boldsymbol{\gamma}}' \quad \hat{\mathbf{a}}'\right]', \mathbf{C}\right), \qquad [A1]$$

where $\left[\hat{\boldsymbol{\beta}}' \quad \hat{\boldsymbol{\gamma}}' \quad \hat{\mathbf{a}}'\right]' = \mathbf{Cr}$.

*Residual variance.* The FCD of error variance is scaled inverted chi-square:

$$\begin{aligned}&p\left(\sigma_e^2 \mid ELSE\right)\propto\left(\sigma_e^2\right)^{-\left(\frac{n+\nu_e+2}{2}\right)}\\ &\times\exp\left(-\frac{1}{2\sigma_e^2}\left((\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma}-\mathbf{Za})'(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}-\mathbf{X}_2\boldsymbol{\gamma}-\mathbf{Za})+\nu_e s_e^2\right)\right)\end{aligned} \qquad [A2]$$

*Additive genetic variances.* The FCD of genetic variances are not of standard forms:

$$p\left(\sigma^2_{A_b} \mid ELSE\right) \propto \left|\mathbf{G}(\boldsymbol{\varphi})\right|^{-\frac{1}{2}} \left(\sigma^2_{A_b}\right)^{-\left(\frac{\nu_{A_b}+2}{2}\right)} \exp\left(-\frac{\mathbf{a}'\left(\mathbf{G}(\boldsymbol{\varphi})\right)^{-1}\mathbf{a}}{2} - \frac{\nu_{A_b} s^2_{A_b}}{2\sigma^2_{A_b}}\right), b=1,\dots, B;\text{[A3]}$$

$$p\left(\sigma^2_{S_{bb'}} \mid ELSE\right) \propto \left|\mathbf{G}(\boldsymbol{\varphi})\right|^{-\frac{1}{2}} \left(\sigma^2_{S_{bb'}}\right)^{-\left(\frac{\nu_{S_{bb'}}+2}{2}\right)} \exp\left(-\frac{\mathbf{a}'\left(\mathbf{G}(\boldsymbol{\varphi})\right)^{-1}\mathbf{a}}{2} - \frac{\nu_{S_{bb'}} s^2_{S_{bb'}}}{2\sigma^2_{S_{bb'}}}\right),$$

$$b=1,\dots, B\text{-}1;\ b'=b+1,\dots, B.\ \text{[A4]}$$

A Metropolis-Hastings (MH) algorithm, can be used to sample from [A3] and [A4]. The MH implementation was based on a random walk specification (Chib and Greenberg, 1995) with scaled inverted chi-square proposal density. A proposal value $\sigma_c^{2*}$ is generated from a scaled inverted chi-square distribution with scaling factor equal to the value of the parameter in the previous cycle $\sigma_c^{2[t-1]}$ times the degrees of freedom $\nu_c^*$, i.e.:

$$q\left(\sigma_c^{2[t-1]}, \sigma_c^{2*}\right) \propto \left(\sigma_c^{2*}\right)^{-\left(\frac{\nu_c^*+2}{2}\right)} \left(\nu_c^*\sigma_c^{2[t-1]}\right)^{\left(\frac{\nu_c^*}{2}\right)} \exp\left(-\frac{\nu_c^*\sigma_c^{2[t-1]}}{2\sigma_c^{2*}}\right).$$

To improve mixing, the degrees of freedom $\nu_c^*$ were tuned during the burning period such that the acceptance of proposal values was about 40% (Chib and Greenberg, 1995). The acceptance rate is given by:

$$\alpha_c\left(\sigma_c^{2[t-1]}, \sigma_c^{2*}\right) = \begin{cases} \min\left(\dfrac{p\left(\sigma_c^{2*} \mid \boldsymbol{\theta}_{\left(-\sigma_c^2\right)}, \mathbf{y}\right) q\left(\sigma_c^{2*}, \sigma_c^{2[t-1]}\right)}{p\left(\sigma_c^{2[t-1]} \mid \boldsymbol{\theta}_{\left(-\sigma_c^2\right)}, \mathbf{y}\right) q\left(\sigma_c^{2[t-1]}, \sigma_c^{2*}\right)}, 1\right), & \text{if } p\left(\sigma_c^{2[t-1]} \mid \boldsymbol{\theta}_{\left(-\sigma_c^2\right)}, \mathbf{y}\right) q\left(\sigma_c^{2[t-1]}, \sigma_c^{2*}\right) > 0 \\ 1, & \text{otherwise} \end{cases}$$

where $\boldsymbol{\theta}_{\left(-\sigma_c^2\right)}$ is the vector of all parameters but $\sigma_c^2$, $c$ = $A_b$, $S_{bb'}$. Finally, to facilitate the computations, we use the fact that $\left|\mathbf{G}(\boldsymbol{\varphi})\right| = \left|\boldsymbol{\Omega}(\boldsymbol{\varphi})\right| = \prod_{j=1}^{q} \omega_j$ and that $\mathbf{a}'\left(\mathbf{G}(\boldsymbol{\varphi})\right)^{-1}\mathbf{a}$ can be computed as $\sum_{j=1}^{q_b} \omega_j^{-1} a_j^2 + \sum_{j=q_b+1}^{q} \omega_j^{-1}\left(a_j - .5a_j^s - .5a_j^d\right)^2$, where $q_b$ denotes the number of *base* animals (animals with both parents unknown).

# CHAPTER 4

# ROBUST QUANTITATIVE GENETIC INFERENCE ON POST-WEANING GAIN OF HEREFORD-NELORE CATTLE USING A MULTIPLE-BREED AND STRUCTURAL RESIDUAL VARIANCES ANIMAL MODEL

**ABSTRACT**: The objective of this study was to propose and apply hierarchical Bayes models with structural residual variances, combining heteroskedastic and heavy-tailed densities, for the prediction of genetic merit in multiple-breed populations. Data comprised 22,717 post weaning gain (PWG) records of a Nelore-Hereford population (40,082 animals in the pedigree). A 3 × 2 factorial specification for the residual variances based on distributional (Gaussian versus Student *t* versus Slash) and variability (homoskedastic versus heteroskedastic) assumptions was evaluated: Gaussian homoskedastic (G-HO); Student *t* homoskedastic (T-HO); Slash homoskedastic (S-HO); Gaussian heteroskedastic (G-HE); Student *t* heteroskedastic (T-HE); and Slash heteroskedastic (S-HE). Based on Pseudo Bayes Factors, the T-HE provided the best fit to PWG with G-HO performing the worst. For the T-HE model, the posterior mean (PMEAN) ± posterior standard deviation (PSD) of the degrees of freedom parameter ($\nu$) was 7.33 ± 0.48 therefore, reflecting evidence for a residual distribution much heavier tailed than Gaussian for PWG. An illustration of the use of robust models to investigate outliers is also presented. Amongst various fixed factors (breed proportion, breed heterozygosity and sex), only breed heterozygosity may be important (P<0.10) as a cause for residual heteroskedasticity. Contemporary group (CG) effects were an important random source of residual heteroskedasticity with the ratio between the largest and smallest CG residual variances being about 20. In the comparison of homoskedastic heavy-tailed versus heteroskedastic heavy-tailed models, there was some evidence that homoskedastic error models may misinterpret records in high variance subclasses as outliers. Inference on genetic variance components changed considerably depending on the structural specification for the residual variance. The Herefords had a larger PMEAN genetic variance compared to the Nelores in the G-HO and T-HO models, whereas the converse was true in the G-HE and T-HE models. The between-breed segregation variance was the least affected among genetic components by the different model specifications. Inferences based on the conventional G-HO model led to remarkable rerankings of animal genetic effects for selection compared to the better fitting T-HE specification. Therefore, the use of normal homoskedastic residual specifications in current genetic evaluation models may be impeding genetic progress.

**Key Words:** Bayesian inference, Heteroskedasticity, Heterogeneity of variance, Robust models, Structural models.

## 1. Introduction

Crossbreeding and selection are two of the most important tools available to increase the efficiency of livestock production through genetic means, utilizing heterosis and complementarity between breeds (Gregory et al., 1999). However, prediction of genetic merit (genetic evaluation) and selection on multiple-breed populations are complicated by different genetic backgrounds and degrees of crossing present in these populations. The response to selection is proportional to the accuracy of genetic merit predictions

(Falconer and Mackay, 1996), which naturally depends on correct specification of the genetic evaluation model.

Livestock performance is generally measured across diverse production systems and environments, with data quality often compromised by the occurrence of recording error, preferential treatment and the effect of injury or disease. Hierarchical Bayes models provide a general framework to address problems arising from poorly structured data (Sorensen and Gianola, 2002). A variety of hierarchical model constructions have been used to address heteroskedasticity (Foulley et al., 1992; Foulley and Quaas, 1995; Gianola et al., 1992; SanCristobal et al., 1993) and robustness to outliers (Stranden and Gianola, 1998, 1999). These two issues have been tackled separately; however, there is no conceptual difficulty in considering them jointly (Kizilkaya and Tempelman, 2003).

Residual heteroskedasticity has been reported in beef cattle for growth performance (Garrick et al., 1989; Nunez-Dominguez et al., 1995; Rodriguez-Almeida et al., 1995) and carcass scan traits (Reverter et al., 1997) with region, herd, level of production, herd size, year, sex and class of age of dam being identified as contributors. Breed composition as a source of heteroskedasticity could also be considered in crossbred populations (Arnold et al., 1992; Garrick et al., 1989; Rodriguez-Almeida et al., 1995).

Several statistical approaches have been considered for modeling heteroskedasticity, the more notable approaches being that due to Gianola et al. (1992) and SanCristobal et al. (1993). The method of Gianola et al. (1992) is based on regarding herd residual variances as random variables from a scaled inverted chi-square distribution such that the estimates obtained represent a compromise between a data based statistic (REML) and parameters of the distribution of variances (hyperparameters) based on a borrowing of information across subclasses. In SanCristobal et al. (1993), the structural linear method for log variances of Foulley et al. (1992) was extended to genetic and residual effects. The method uses a log link to model variances as a linear function of unknown dispersion parameters.

Beyond heteroskedasticity, the accuracy of estimated breeding values nevertheless depends upon the quality of the performance and pedigree data provided. The presence of observations influenced by factors not accounted for in the statistical analysis and potentially having extreme influence (i.e. outliers), can severely bias parameter estimates and genetic evaluations, since most linear mixed models used in animal breeding assume that residual and random effects follow a (light-tailed) normal distribution. Preferential treatment, disease, inappropriate contemporary group formation, record errors and animal misidentification are possible reasons for outliers in beef cattle populations. Data editing generally involves deleting observations that are considered extremely far from the phenotypic mean of its class (usually three or more standard deviations) or the ratio record/mean of its class fall outside the range of 60% to 140% (Bertrand and Wiggans, 1998).

An alternative to the deletion of plausible, albeit extreme, observations is the use of symmetric heavy-tailed densities, such as the Student *t*, slash and contaminated normal, for specifying residual distributions (Lange and Sinsheimer, 1993). These densities are examples of Normal/independent distribution families that can better accommodate extreme observations due to their heavy-tailed features (Lange and Sinsheimer, 1993; Rogers and Tukey, 1972). In animal breeding, Stranden and Gianola recently introduced a hierarchical Bayes model that specifies residuals to have Student *t* rather than normal densities. The use of Student $t$ residuals have been shown to better accommodate data characterized by preferential treatment, compared to a normal residual density (Stranden and Gianola, 1998).

The objectives of this study were: 1) to propose a hierarchical Bayes model combining residual heteroskedasticity and heavy-tailed residual densities for the prediction of genetic merit in multiple-breed populations, and 2) to apply the proposed model to a dataset on post-weaning gains from purebred and crossbred animals derived from Nelores and Herefords raised in diverse environments, in order to identify sources of residual heteroskedasticity and to assess the need for outlier-robust genetic evaluations.

## 2. Material and Methods

### 2.1. Nelore-Hereford data

Data analyzed in this study consisted of post-weaning gain (PWG) records of a beef cattle population comprising of Herefords, Nelores and their crosses under genetic evaluation in Brazil. The records were collected between 1974 and 2000 by the Brazilian National Breeders Association and Gensys Associated Consultants within a large-scale genetic evaluation program called “Delta G Connection”. After deleting contemporary groups with less than ten animals and sires with less than five offspring, there were 22,717 records of PWG from 15 different herds with a total of 40,082 animals in the pedigree file. The animals were grown on extensive pasture conditions in three different regions, of which two were in the tropical area and one in the sub-tropical area of the Country. Region 1 comprised of two farms located between 14°S and 16°S latitude with

5,410 records (23.8%), Region 2 had three farms located between 21°S and 23°S with 3,110 records (13.7%), and Region 3 had ten farms located between 30°S and 32°S with 14,197 records (62.5%). The average PWG ± standard deviation was 98.2 ± 41.2 kg.

Breed composition groups were highly unbalanced across the various regions. Animals with records had breed proportions ranging from purebred Hereford to 7/8 Nelore (there were no purebred Nelore animals with records); however, purebred Herefords and $F_1$'s represented about 90% of the records. Dams were mostly represented by purebreds of the two breeds and most of the animals with records in the tropical part of the country (Regions 1 and 2) were $F_1$s. More details on the structure of this data set are presented in Chapter 3.

### 2.2. Hierarchical Bayes model

*2.2.1. Multiple-breed animal model with structural residual variances*

The first stage of the model specifies the conditional sampling density of the $n \times 1$ data vector $\mathbf{y} = \left\{y_j\right\}_{j \in S}$. The component of this density for a record on individual $j$, is

$$y_j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{a}, \sigma^2_{e_j} \sim N\left(\mathbf{x}'_{1j}\boldsymbol{\beta} + \mathbf{x}'_{2j}\boldsymbol{\gamma} + \mathbf{z}'_j\mathbf{a}, \sigma^2_{e_j}\right),\ j \in S, \qquad [1]$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of non-genetic effects, $\boldsymbol{\gamma}$ is a $t \times 1$ vector of "fixed" genetic effects, $\mathbf{a}$ is a $q \times 1$ vector of animal additive genetic effects; and $\mathbf{x}'_{1j}$, $\mathbf{x}'_{2j}$, and $\mathbf{z}'_j$ are known row incidence vectors; more information on these location parameters can be found in Chapter 3. Moreover, $S$ represents the sample of size $n$ of animals having records; typically $n < q$ since $\mathbf{a}$ includes effects for ancestor animals without records. Finally, $\sigma^2_{e_j}$ represents the residual variance, specific for animal $j$. Following Foulley et al. (1992), a linear mixed model is assumed for the log residual variance: $\log\sigma^2_{e_j} = \mu_e + \mathbf{p}'_j\boldsymbol{\tau} + \mathbf{q}'_j\boldsymbol{\upsilon} + r_j$, where $\boldsymbol{\tau} = \{\tau_l\}_{k=1}^{K}$ and $\boldsymbol{\upsilon} = \{\upsilon_l\}_{l=1}^{L}$ are, respectively, vectors of "fixed" and random dispersion parameters, and $\mathbf{p}'_j$ and $\mathbf{q}'_j$ are known incidence vectors relating the elements on $\boldsymbol{\tau}$ and $\boldsymbol{\upsilon}$ to the residual variance of animal *j*. Moreover, $r_j$ is an "error" term for the residual variance associated with record *j*. This specification translates to a multiplicative model on $\sigma^2_{e_j}$:

$$\sigma^2_{e_j} = \exp\left(\mu_e + \mathbf{p}'_j\boldsymbol{\tau} + \mathbf{q}'_j\boldsymbol{\upsilon} + r_j\right) = e^{\mu_e} \times \left[\prod\nolimits_{k=1}^{K} e^{\tau_k p_{jk}}\right] \times \left[\prod\nolimits_{l=1}^{L} e^{\upsilon_l q_{jl}}\right] \times e^{r_j},$$

which can be expressed as:

$$\sigma^2_{e_j} = \frac{\bar{\sigma}^2_e \times \left[\prod\nolimits_{k=1}^{K} \lambda_k^{p_{jk}}\right] \times \left[\prod\nolimits_{l=1}^{L} \xi_l^{q_{jl}}\right]}{w_j}, \tag*{[2]}$$

where $\bar{\sigma}^2_e = e^{\mu_e}$ represents a "reference" residual variance; $\lambda_k = e^{\tau_k}$ and $\xi_l = e^{\upsilon_l}$ are, respectively, "fixed" and random positive multiplicative dispersion parameters or scaling factors; $p_{jk}$ and $q_{jl}$ are known incidence quantities, corresponding to the *k*th and *l*th elements of $\mathbf{p}'_j$ and $\mathbf{q}'_j$, respectively; and $w_j = e^{-r_j}$ is an independent weight variable. Note that for identifiability purposes, restrictions must be applied to the parameters in $\boldsymbol{\lambda} = \{\lambda_k\}_{k=1}^{K}$ in the same manner to that required for classical fixed effects (Searle, 1971).

In the second stage we specify our prior assumptions on all unknowns defined in Equations [1] and [2]. For the location parameters we adopt the following ordinary assumptions:

$$\boldsymbol{\beta} \mid \boldsymbol{\beta}_o, \mathbf{V}_\beta \sim N\left(\boldsymbol{\beta}_o, \mathbf{V}_{\boldsymbol{\beta}}\right), \tag*{[3]}$$

$$\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_o, \mathbf{V}_\gamma \sim N\left(\boldsymbol{\gamma}_o, \mathbf{V}_\gamma\right), \qquad [4]$$

and

$$\mathbf{a} \mid \boldsymbol{\varphi} \sim N\left(\mathbf{0}, \mathbf{G}(\boldsymbol{\varphi})\right), \qquad [5]$$

where $\boldsymbol{\beta}_o$ and $\boldsymbol{\gamma}_o$ are prior means, and $\mathbf{V}_\beta$, $\mathbf{V}_\gamma$ and $\mathbf{G}(\boldsymbol{\varphi})$ are prior variance-covariance matrices. The additive genetic variance-covariance matrix $\mathbf{G}(\boldsymbol{\varphi})$ is a function of more than one dispersion parameter in $\boldsymbol{\varphi}$ for crossbred populations, as defined by Lo et al. (1993), i.e. $\boldsymbol{\varphi} = \left[\left\{\sigma^2_{A_b}\right\}_{b=1}^{B}, \left\{\sigma^2_{S_{bb'}}\right\}_{b=1,b'>b}^{B-1,\ B}\right]$. Here, $\sigma^2_{A_b}$ is the additive variance of breed $b$ and $\sigma^2_{S_{bb'}}$ is variance due to the segregation between breed $b$ and $b'$ or the additional variance observed in the $F_2$ generation over the $F_1$. Additional details are provided in Chapter 3.

For dispersion parameters or scaling factors determining record specific residual variances, prior distributions are specified based on the nature of the factor. For the "fixed" effects factor, we adopt inverted gamma distributions, specified as follows:

$$p\left(\bar{\sigma}^2_e \middle| \bar{\alpha}_e, \bar{\beta}_e\right) \propto \left(\bar{\sigma}^2_e\right)^{-\left(\bar{\alpha}_e+1\right)} \exp\left(-\frac{\bar{\beta}_e}{\bar{\sigma}^2_e}\right); \qquad [6]$$

$$p\left(\lambda_k \middle| \alpha_{(\lambda)k}, \beta_{(\lambda)k}\right) \propto \left(\lambda_k\right)^{-\left(\alpha_{(\lambda)k}+1\right)} \exp\left(-\frac{\beta_{(\lambda)k}}{\lambda_k}\right),\ k=1,\ldots,K. \qquad [7]$$

Bounded uniform priors (between zero and an arbitrary positive value) may alternatively be specified on $\lambda_k$.

Furthermore, a structural prior is used to model the random multiplicative scaling factors included on $\boldsymbol{\xi} = \{\xi_l\}_{l=1}^{L}$. We conveniently choose this structural prior to be an inverted gamma distribution with parameters $\eta$ and $\eta-1$ (Kizilkaya and Tempelman, 2002),

$$p(\xi_l \mid \eta) = \frac{(\eta-1)^{\eta}}{\Gamma(\eta)}(\xi_l)^{-(\eta+1)}\exp\left(-\frac{\eta-1}{\xi_l}\right), \quad l = 1, 2, \ldots, L;\ \eta > 1. \qquad [8]$$

Here $\mathrm{E}(\xi_l) = 1$ and $\mathrm{Var}(\xi_l) = \frac{1}{\eta-2}$ (defined for $\eta > 2$), such that as $\eta \to \infty$, the random effect factor's influence on residual heteroskedasticity diminishes. Note that with the specification in [8], there is a borrowing of information across the $L$ levels of the random factor, just as there is for classical random effects modeling of location parameters.

Finally to conclude this stage, there are several possible distributional assumptions on the weights $\mathbf{w} = \left\{w_j\right\}_{j\in S}$ that yield a Normal/independent specification on records (e.g. Contaminated Normal, Student *t*, Slash, or Double Exponential distributions). We specifically consider two heavy-tailed alternatives to the Normal distribution: the Student *t* and the Slash distributions.

A Student *t* distribution on $y_j$'s is obtained by letting $p\left(w_j \mid \nu\right)$ be Gamma($\nu/2$, $\nu/2$), which has the following density function:

$$p\left(w_j \mid \nu\right) = \frac{(\nu/2)^{\nu/2} w_j^{\nu/2-1}\exp\left(-w_j \nu/2\right)}{\Gamma(\nu/2)}, \quad j \in S,\ \nu > 0,\ w_j > 0. \qquad [9]$$

Alternatively $p\left(w_j \mid \nu\right)$ may be specified as having the following distributional form:

$$p\left(w_j \mid \nu\right) = \nu w_j^{\nu-1}, \quad j \in S,\ \nu > 0,\ 0 < w_j \leq 1, \qquad [10]$$

in which case we have a Slash distribution on $y_j$'s.

The third stage of the model corresponds to inverted gamma prior distributions on the genetic variance components in $\boldsymbol{\varphi}$, defined as follows:

$$p\left(\sigma_{A_b}^2 \middle| \alpha_{A_b}, \beta_{A_b}\right) \propto \left(\sigma_{A_b}^2\right)^{-\left(\alpha_{A_b}+1\right)} \exp\left(-\frac{\beta_{A_b}}{\sigma_{A_b}^2}\right), \ b=1,\ldots,B; \qquad [11]$$

$$p\left(\sigma_{S_{bb'}}^2 \middle| \alpha_{S_{bb'}}, \beta_{S_{bb'}}\right) \propto \left(\sigma_{S_{bb'}}^2\right)^{-\left(\alpha_{S_{bb'}}+1\right)} \exp\left(-\frac{\beta_{S_{bb'}}}{\sigma_{S_{bb'}}^2}\right), b=1,\ldots, B;$$

$$b'=b+1,\ldots, B. \quad [12]$$

Finally, arbitrary priors on $\eta$ and on $\nu$ can be specified as $p(\eta)$ and $p(\nu)$, respectively.

The joint posterior density is the product of [1], [3], [4], [5], [6], [7], [8], [9] or [10], [11], [12], $p(\eta)$ and $p(\nu)$, given by:

$$\begin{aligned} p\left(\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{a},\boldsymbol{\varphi},\bar{\sigma}_e^2,\boldsymbol{\lambda},\boldsymbol{\xi},\mathbf{w},\eta,\nu \mid \boldsymbol{\zeta},\mathbf{y}\right) &= \prod_{j\in S} p\left(y_j \mid \boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{a},\sigma_{e_j}^2\right) p\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}_o,\mathbf{V}_{\boldsymbol{\beta}}\right) p\left(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}_o,\mathbf{V}_{\boldsymbol{\gamma}}\right) \\ &\times p(\mathbf{a} \mid \boldsymbol{\varphi})\, p\left(\bar{\sigma}_e^2 \middle| \bar{\alpha}_e,\bar{\beta}_e\right) \prod_{k=1}^{K} p\left(\lambda_k \middle| \alpha_{(\lambda)k},\beta_{(\lambda)k}\right) \prod_{l=1}^{L} p(\xi_l \mid \eta) \prod_{j\in S} p\left(w_j \mid \nu\right) \qquad , \\ &\times \prod_{b=1}^{B} p\left(\sigma_{A_b}^2 \middle| \alpha_{A_b},\beta_{A_b}\right) \prod_{b=1}^{B-1} \prod_{b'>b}^{B} p\left(\sigma_{S_{bb'}}^2 \middle| \alpha_{S_{bb'}},\beta_{S_{bb'}}\right) p(\eta) p(\nu) \end{aligned}$$

where

$$\boldsymbol{\zeta}=\left(\boldsymbol{\beta}_o,\mathbf{V}_{\boldsymbol{\beta}},\boldsymbol{\gamma}_o,\mathbf{V}_{\boldsymbol{\gamma}},\bar{\alpha}_e,\bar{\beta}_e,\alpha_{(\lambda)1},\ldots,\alpha_{(\lambda)K},\beta_{(\lambda)1},\ldots,\beta_{(\lambda)K},\alpha_{A_1},\ldots,\alpha_{A_B},\beta_{A_1},\ldots,\beta_{A_B},\alpha_{S_{12}},\ldots,\alpha_{S_{B-1B}},\beta_{S_{12}},\ldots,\beta_{S_{B-1B}}\right)$$

is the vector of hyperparameters.

The full conditional densities (FCD) of all unknown parameters/quantities or blocks thereof necessary to conduct MCMC inference are derived in the Appendix to this chapter.

### *2.2.2. Model specification for Nelore-Hereford data*

The following linear model was used to explain a PWG record ($y_{jlrs}$) on animal $j$; pertaining to contemporary group (GC) $l$ ($l$=1,2,...,940), Region $r$ ($r$=1,2,3), and Sex $s$ ($s$=1,2):

$$y_{jlrs} = \mu + \beta_1 PWP_j + \beta_2 AoD_{d_j} + \beta_3 AoD_{d_j}^2 + CG_l + R_r + S_s + \gamma_{A_1} f_1^j + \gamma_{D_{12}} f_{12}^j + \gamma_{AA_{12}} 2 f_1^j f_2^j$$
$$+ \gamma_{A_1 \times PWP} f_1^j PWP_j + \gamma_{A_1 \times AoD} f_1^{d_j} AoD_{d_j} + \gamma_{A_1 \times AoD^2} f_1^{d_j} AoD_{d_j}^2 + \gamma_{A_1 \times S} f_1^j S_s + a_j + e_j$$

Here, $\mu$ is the overall mean; $\beta_1$, $\beta_2$, and $\beta_3$ are unknown elements of **β** associated, respectively, with the length of the post-weaning test period ($PWP_j$), linear and quadratic age of dam ($AoD_{d_j}$) effects, where the subscript $d_j$ refers to the dam of animal $j$. Moreover, $CG_l$ represents the effect of the $l$th CG (herd, year, season and management subclasses), assumed to be an uncorrelated random effect; $R_r$ represents the effect of the $r$th region; and $S_s$ is the gender effect, which are also non-genetic effects pertaining to **β**. The effects corresponding to elements of γ were $\gamma_{A_1}$, the additive effect of the Nelore breed; $\gamma_{D_{12}}$ the dominance effect involving the Nelore and Hereford breeds; $\gamma_{AA_{12}}$, the additive × additive effect involving the Nelore and Hereford breeds; and interactions of the individual Nelore fraction of alleles ($f_1^j$) with $PWP$ and sex, respectively denoted by $\gamma_{A_1 \times PWP}$ and $\gamma_{A_1 \times S}$, and of the maternal Nelore fraction of alleles ($f_1^{d_j}$) with $AoD$ and $AoD^2$, respectively denoted by $\gamma_{A_1 \times AoD}$ and $\gamma_{A_1 \times AoD^2}$. Additionally, $f_{12}^j$ is the heterozygosity coefficient, which represents the probability that for a randomly chosen locus from individual $j$, one allele is derived from the Nelore breed and the other allele is derived from the Hereford breed. Furthermore, $a_j$ is the random additive genetic effect of animal $j$ and $e_j$ is residual term associated with the record on animal $j$.

A 3 × 2 factorial representation was used in the specification of six different models for residual variances with one factor being the distributional specification (Gaussian, Student *t* or Slash) and the other factor defining the nature of the variability (homoskedastic versus heteroskedastic). All six model versions were applied and compared in terms of fit to the PWG data, thereby having the same location parameters and genetic dispersion parameters specification, but differing in their residual variance structure as follows:

1) Gaussian homoskedastic (G-HO): This was the model used in Chapter 3. Residuals and records were assumed to have Normal distribution with a homogeneous residual variance, that is $w_j = 1$ and $\sigma^2_{e_j} = \bar{\sigma}^2_e$ for all $j$.

2) Student *t* homoskedastic (T-HO): in this model, conditioned on $w_j$'s, residuals were assumed to have Normal distribution with a common residual variance; however the $w_j$'s are assume to have the distribution specified on [9] and consequently residuals have a Student *t* distribution.

3) Slash homoskedastic (S-HO): in this model, conditioned on $w_j$'s, residuals were again assumed to have Normal distribution with a common residual variance, whereas the $w_j$'s are assume to have the distribution specified on [10] and consequently records have a Slash distribution.

4) Gaussian heteroskedastic (G-HE): in this case residuals and records were assumed to have Normal distribution with heterogeneous residual variance, that is $w_j = 1$, for all

*j*, and $\sigma^2_{e_j} = \bar{\sigma}^2_e \lambda_1^{p_{j1}} \lambda_2^{p_{j2}} \lambda_3^{p_{j3}} \xi_l$, such that the residual variance for animal *j* on CG *l*, is a multiplicative function of Sex ($\lambda_1$), proportion of Nelore breed ($\lambda_2$) and heterozygosity coefficient between Nelore-Hereford breeds ($\lambda_3$) "fixed" effects and of a CG random effect ($\xi_l$). Here, $p_{j1}$ is equal to one if the animal is a male and zero if female, $p_{j2} = f_1^j$ and $p_{j3} = f_{12}^j$.

5) Student *t* heteroskedastic (T-HE): this model combines the properties of models 2) and 4), such that $\sigma^2_{e_j} = \dfrac{\bar{\sigma}^2_e \lambda_1^{p_{j1}} \lambda_2^{p_{j2}} \lambda_3^{p_{j3}} \xi_l}{w_j}$, for all *j*, where the $w_j$'s are assume to have the distribution specified in [9].

6) Slash heteroskedastic (S-HE): this model combines the properties of models 3) and 4), such that $\sigma^2_{e_j} = \dfrac{\bar{\sigma}^2_e \lambda_1^{p_{j1}} \lambda_2^{p_{j2}} \lambda_3^{p_{j3}} \xi_l}{w_j}$, for all *j*, where the $w_j$'s are assume to have the distribution specified in [10].

The length of the MCMC chain for PWG was 200,000 cycles after 15,000 cycles of burn-in for all six models. Means, modes, key percentiles and standard deviations of the parameters were obtained from their respective marginal posterior densities. For each dispersion parameter, the initial monotone sequence approach (Geyer, 1992) was used to calculate effective sample sizes (ESS), which estimates the number of independent samples with information content equivalent to that contained within the 200,000 dependent samples (Sorensen et al., 1995).

*2.2.3. Model choice criterion*

We considered the *Pseudo Bayes Factor* (PBF) (Gelfand, 1996) as a means of model choice. The PBF involves the evaluation of the first stage density in [1] for each MCMC cycle; let $p\left(y_j \mid \mathbf{y}_{(-j)}, M_m\right)$ be the conditional predictive ordinate (CPO) for observation $y_j$, intended to be a cross-validation density, suggesting what values of $y_j$ are likely when Model $M_m$ is fit to all other observations $\mathbf{y}_{(-j)}$ except $y_j$. Letting $\boldsymbol{\theta}' = \left[\boldsymbol{\beta}' \quad \mathbf{g}' \quad \mathbf{a}' \quad \bar{\sigma}_e^2 \quad \boldsymbol{\lambda}' \quad \boldsymbol{\xi}'\right]$, a MCMC approximation for the CPO is obtained by a harmonic mean $p\left(y_j \mid \mathbf{y}_{(-j)}, M_m\right) \approx \left(\frac{1}{G}\sum_{l=1}^{G} p^{-1}\left(y_j \mid \boldsymbol{\theta}^{(l)}, M_m\right)\right)^{-1}$, where $\theta^{(l)}$ is the post burn-in MCMC sample for $\boldsymbol{\theta}$, $l = 1,2,\ldots,G$.

An approximate log marginal likelihood (LML) overall observations can be obtained by: $LML_m = \sum_{j \in S} \log p\left(y_j \mid \mathbf{y}_{(-j)}, M_m\right) \approx \sum_{j \in S} \left(\frac{1}{G}\sum_{l=1}^{G} p^{-1}\left(y_j \mid \boldsymbol{\theta}^{(l)}, M_m\right)\right)^{-1}$. Finally, for comparing, say, models $M_1$ and $M_2$, the corresponding PBF is determined to be $PBF_{1,2} = \exp\left(LML_1 - LML_2\right)$.

*2.2.4. Prior specifications*

Due to lack of objective prior information on this population, we adopt bounded uniform priors on $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$. Conjugate specifications were adopted on genetic variance components, specifically, $\alpha_{A_1} = \alpha_{A_2} = \alpha_{S_{12}} = 2.5$; $\beta_{A_1} = \beta_{A_2} = 2.5 \times 80$ and $\beta_{S_{12}} = 2.5 \times 10$, such that the prior guesses and these components were $\sigma_{A_1}^2 = \sigma_{A_2}^2 = 80$ and $\sigma_{S_{12}}^2 = 10$. For the "reference" residual variance $\bar{\sigma}_e^2$, we had $\bar{\alpha}_e = 2.5$, and $\bar{\beta}_e$ was specified such that the prior guess on the marginal residual variance with respect to **w**, denoted by $\sigma_E^2$, was the same for all models. This prior was based on a REML estimate of $\bar{\sigma}_e^2$ for this PWG data being $\sigma_E^2 = 350$. Accordingly, $\bar{\beta}_e = 2.5 \times 350$ for the Gaussian models (G-HO and G-HE); $\bar{\beta}_e = 2.5 \times \frac{\nu - 2}{\nu} 350$ for the Student *t* models (T-HO and T-HE); and $\bar{\beta}_e = 2.5 \times \frac{\nu - 1}{\nu} 350$ for the Slash models (S-HO and S-HE). Moreover, we specified $p\left(\eta | \alpha_{(\eta)} = 0.03, \beta_{(\eta)} = 0.01\right) \propto \eta^{\alpha(\eta)^{-1}} \exp\left(-\beta_{(\eta)}\eta\right)$, which is based on a prior guess on the heterogeneity parameter of $\eta = 3$, but with very large variance (300). Furthermore, $p\left(\nu | \alpha_{(\nu)}, \beta_{(\nu)}\right) \propto \nu^{\alpha(\nu)^{-1}} \exp\left(-\beta_{(\nu)}\nu\right)$, with $\alpha_{(\nu)} = 0.04$ and $\beta_{(\nu)} = 0.01$ for Student *t* models, and $\alpha_{(\nu)} = 0.015$ and $\beta_{(\nu)} = 0.01$ for the Slash models. That is, our guess of the prior mean was $\nu = 4$ for T-HO and T-HE and $\nu = 1.5$ for S-HO and S-HE.

### *2.2.5. Genetic parameters*

The additive genetic variance of genotype *g*, denoted $\sigma_{A_g}^2$, was obtained by $\sigma_{A_g}^2 = f_1^g \sigma_{A_1}^2 + f_2^g \sigma_{A_2}^2 + 2\left(f_1^s f_2^s + f_1^d f_2^d\right)\sigma_{S_{12}}^2$, where $f_b^i$ indicates the proportion of breed *b* with *b*=1 being Nelore and *b*=2 being Hereford. The superscript *i* refers the genotype of the group itself (*g*), paternal (*s*) and maternal (*d*). The marginal residual variance of genotype *g* ($\sigma_{E_g}^2$) was calculated by $\sigma_{E_g}^2 = \bar{\sigma}_e^2 \lambda_1^{0.3197} \lambda_2^{f_1^g} \lambda_3^{f_{12}^g}$ for Gaussian models; by $\sigma_{E_g}^2 = \frac{\nu}{\nu - 2} \bar{\sigma}_e^2 \lambda_1^{0.3197} \lambda_2^{f_1^g} \lambda_3^{f_{12}^g}$ for Student *t* models; and by

$\sigma_{E_g}^2 = \frac{\nu}{\nu - 1}\bar{\sigma}_e^2 \lambda_1^{0.3197} \lambda_2^{f_1^g} \lambda_3^{f_{12}^g}$ for Slash models. Here 0.3197 represents $\bar{p}_{.1} = \frac{1}{n}\sum_{j \in S} \bar{p}_{j1}$ or the proportion of calves that were male and $f_{12}^g$ is the heterozygosity coefficient of genotype $g$. For homoskedastic models $\lambda_1=\lambda_2=\lambda_3=1$. The phenotypic variance of $g$ is then $\sigma_{P_g}^2 = \sigma_{E_g}^2 + \sigma_{A_g}^2$ such that its additive heritability is $h_{A_g}^2 = \sigma_{A_g}^2 / \sigma_{P_g}^2$. We deliberately omit the contemporary group component in the phenotypic variance to make our heritability estimates comparable with most results found in the animal breeding literature (Koots et al., 1994), where contemporary groups are often considered as fixed factors.

## 3. Results and Discussion

### 3.1. Model choice

Among the six different models used to analyze the PWG data, the T-HE model provided the best fit. This model had the largest LML (Table 4.1) and had a PBF of $9.328 \times 10^{19}$, when compared to the S-HE model, which provided the next best fit. It is clear that the conventional G-HO model is not an appropriate choice for this data set as PBF's approach zero when this model is compared to all other five models fit (Table 4.1).

In terms of accounting for outliers, the Student *t* specification performed better than the Slash for both assumptions about the residual variance, homoskedasticity and heteroskedasticity (Table 4.1). Therefore, we concentrate our inferences throughout this chapter on the Student *t* models, just presenting some key comparisons with the Slash models in terms of outlier detection in Section 3.2. Other researchers analyzing birth weights of rats, however, found better data fits for the Slash distribution compared to the

Table 4.1. Log Marginal Likelihood (LML) in the diagonal and log Pseudo Bayes Factor (difference between the LML of the models represented in the corresponding row and column of the table) for six different models used in the analyses of post-weaning gain of a Nelore-Hereford population

| Model[a] | G-HO | T-HO | S-HO | G-HE | T-HE | S-HE |
|---|---|---|---|---|---|---|
| G-HO | -100,563 | -940 | -852 | -1,108 | -1,511 | -1,465 |
| T-HO | | -99,623 | 88 | -168 | -571 | -525 |
| S-HO | | | -99,711 | -256 | -659 | -613 |
| G-HE | | | | -99,455 | -403 | -357 |
| T-HE | | | | | -99,052 | 46 |
| S-HE | | | | | | -99,098 |

[a]G: Gaussian; T: Student *t*; S: Slash; HO: homoskedastic; HE: heteroskedastic.

Student *t*, Contaminated Normal and Gaussian distributions, nevertheless the advantage over the Student *t* distribution was minimal (Rosa et al., 2003). From Table 4.1, it appears that heteroskedasticity was more important than outlier robustness, as the G-HE model fitted the data much better than either robust homoskedastic models, T-HO and S-HO.

## 3.2. Robustness and detection of outliers

### *3.2.1. Residual analysis of the Gaussian homoskedastic model*

Our primary motivation for adopting robust models derived from the analysis of the fitting to PWG provided by the Chapter 3 model. As can be seen in Figure 4.1, there were several standardized residuals lying outside the range of ± 4.0 standard deviations. Moreover, the kurtosis of the standardized residual distribution was 2.72, indicating that this distribution is leptokurtic. Clearly, the normality assumption on the residuals of this PWG data set is not met, explaining the substantial advantage in fit observed (Table 4.1)

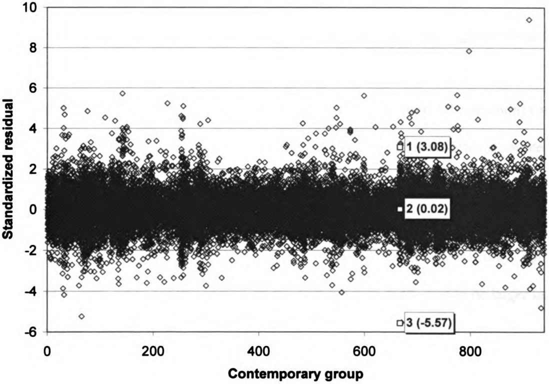

Figure 4.1. Scatter plot of standardized residuals of post-weaning gain on contemporary group id using the Gaussian homoskedastic model. Three residuals from the same contemporary group are highlighted for further inference to be presented later: 1. Represents a mild outlier, being about three standard deviations (SD) from zero; 2. Represents a null residual (perfect fit); and 3. Consists of an extreme outlier, -5.57 SD from zero.

of robust models when compared to their normal counterparts. The skewness of the estimated residuals was 0.34, being of moderate magnitude. Procedures dealing with skewness are, however, available (Fernandez and Steel, 1998) and can be added to the hierarchy of the models without great difficulty.

The posterior distributions of the robustness parameters ($\nu$) were fairly symmetric for all robust models used (data not shown). The posterior mean ± standard deviation of $\nu$, was $7.33 \pm 0.48$ and $2.20 \pm 0.09$, respectively for the T-HE and S-HE models.

Corresponding estimates for the T-HO and S-HO models were $4.79 \pm 0.21$ and $1.66 \pm 0.05$, respectively. The effective number of samples was always greater than 500 for these parameters.

The smaller the values of $\nu$ (i.e. the fatter tails of the heavy-tailed distribution) found when residual homoskedasticity as opposed to heteroskedasticity is modeled illustrate the interdependence between heteroskedasticity and detection of outliers. This was verified for both Student *t* and Slash distributions. It is reasonable to assert then that some observations appearing to be outliers under a homoskedastic model may not be considered as such when the variance of its resident subclass is allowed to be larger using heteroskedastic models.

Beyond attenuating the effect of extreme observations on parameter estimates, robust models can be used to better detect the presence of outliers (Rosa et al., 2003). The posterior distributions of the weight variables $w_j$'s provide valuable information to classify observations as outliers or not outliers. To illustrate this point, we deliberately chose three observations of the same contemporary group and save all samples of their corresponding weight variables: 1- represents a mild outlier, being 3.08 standard deviations (SD) above zero; 2- represents a near zero residual or a perfect fit (0.02 SD); and 3- consists of an extreme outlier, being -5.57 SD from zero (Figure 4.1). Graphs of the posterior distribution of these observations obtained form T-HE and S-HE models are presented in Figure 4.2. One essential difference between posterior distribution of $w_j$'s from Student *t* and Slash models is that in the former model $w_j$'s are defined on the positive real line whereas $w_j$'s are only defined between 0 and 1 in the latter model. This

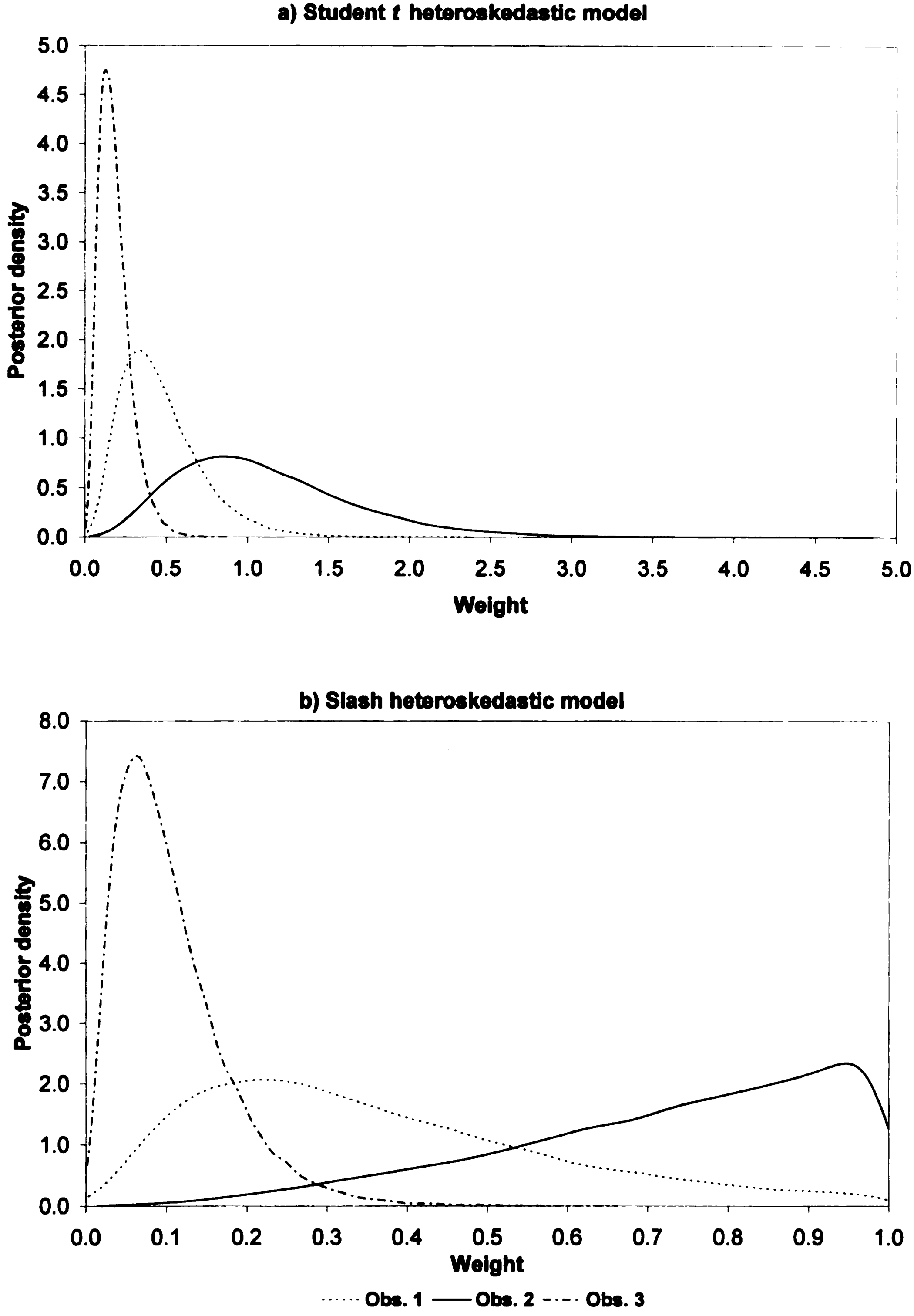

Figure 4.2. Posterior distribution of weight variables corresponding to observation 1 (Obs. 1 - a mild outlier); observation 2 (Obs. 2 – a nearly perfect model fit) and observation 3 (Obs 3. – an extreme outlier) under two robust models: a) Student *t* heteroskedastic model and b) Slash heteroskedastic model.

distinction results in fairly different shape of posterior distributions between the T-HE and S-HE models for $w_1$, $w_2$ and $w_3$, weights corresponding to observations 1, 2 and 3, respectively. Nevertheless, both models successfully identify observation 3 as an extreme outlier. The posterior mean and 95% posterior probability interval (PPI) for $w_3$, were 0.18 and (0.05, 0.42) for the T-HE model. Corresponding values for the S-HE model were 0.10 and (0.02, 0.28). These sharp distributions concentrated around low values for $w_3$ qualify observation 3 as an outlier and attenuate its effect on other parameters estimates. On the other hand, $w_2$ has a relatively flat distribution widely spread throughout the corresponding parameter spaces, for the T-HE and S-HE models. The posterior mean and 95% PPI for $w_2$ corresponding to these models were 1.11 and (0.31, 2.39), and 0.72 and (0.25, 0.99), respectively. These results clearly indicate that observation 2 is not an outlier. The case of observation 2, serves as an illustration to the interdependence between robustness and heteroskedasticity mentioned above. The standardized residual of this observation from G-HE model was 2.53, which is less than 3.07, which was obtained using G-HO. This is, at least partially, due to that observation's contemporary group (CG) having inflated residual variance (the posterior mean ± standard deviation of its corresponding CG random scaling factor was $1.63 \pm 0.23$). The T-HE model does not conclusively declare this observation as an outlier, because the 95% PPI for $w_1$ under this model (0.12, 1.07) includes 1.00, which corresponds to the neutral weight value. Yet, the posterior mean of $w_1$, being 0.46, indicates that observation 1 is down-weighted for inferences.

Under the Slash distribution the inference on $w_1$ is subtler, because the parameter space is constrained to be between 0 and 1 and we cannot use an objective criterion as in the case of the Student *t* distribution (i.e. if 1.00 is include or not in the confidence set for the desired level of precision). The posterior mean and 95% PPI for $w_1$ were, respectively, 0.36 and (0.06, 0.89) for the S-HE model. Despite the relatively low value of the posterior mean of $w_1$, the wide range of the confidence set and the shape of its weight variable distribution in Figure 4.2a make difficult to assert this observation as an outlier.

In addition, weight variables could be used for data collection quality control. For instance, the posterior mean of the weights may provide an indication of the quality of the data originated on each herd. Herds having a high frequency of low values for the posterior mean of weights may have serious issues with the quality of the data collection. Furthermore, scatter plots of the weight variables over time may be useful to determine the profile of the data collection process within each herd; for example, if there is currently a problem or there was a problem in the past. Such plots could also help to check effectiveness of adjustment measures applied when data collection problems are identified.

Finally, it is important to note that the robust models automatically weight each observation for inferences, such that outlying records provide smaller contributions to parameter estimates.

### 3.3. Assessment of heteroskedasticity sources

#### *3.3.1. Fixed effects*

Posterior inference on “fixed” scaling factors for residual heterogeneity obtained by the G-HE and T-HE models are presented in Table 4.2. None of the considered “fixed”

Table 4.2. Posterior mean (PMEAN), posterior standard deviation (PSD), posterior mode (PMODE), 95% posterior probability intervals (PPI), and effective samples size (ESS) for "fixed" effects scaling factors and for the environmental heterogeneity parameter ($\eta$) on post-weaning gain residual variance, obtained from the Gaussian and Student *t* heteroskedastic models

| Effect (parameter) | PMEAN | PSD | PMODE | PPI | ESS |
|---|---|---|---|---|---|
| *Gaussian heteroskedastic model* | | | | | |
| Gender ($\lambda_1$) | 1.14 | 0.09 | 1.13 | (0.98, 1.32) | 1,712 |
| Nelore proportion ($\lambda_2$) | 0.94 | 0.35 | 0.75 | (0.43, 1.73) | 146 |
| Heterozygosity ($\lambda_3$) | 0.77 | 0.15 | 0.71 | (0.52, 1.10) | 142 |
| Environmental ($\eta$) | 3.46 | 0.25 | 3.44 | (3.00, 3.96) | 1,190 |
| *Student t heteroskedastic model* | | | | | |
| Gender ($\lambda_1$) | 1.13 | 0.09 | 1.11 | (0.97, 1.31) | 1,218 |
| Nelore proportion ($\lambda_2$) | 1.15 | 0.45 | 0.80 | (0.48, 2.20) | 160 |
| Heterozygosity ($\lambda_3$) | 0.70 | 0.16 | 0.61 | (0.46, 1.06) | 111 |
| Environmental ($\eta$) | 3.96 | 0.32 | 3.88 | (3.36, 4.63) | 1,109 |

factors turned out statistically "significant" since all 95% posterior probability intervals included 1.00, for both models.

The distribution of the gender scaling factor ($\lambda_1$) was very similar under the G-HE and T-HE models and it is to some extent surprising that there was no effect of sex on the residual variability of PWG, because other authors found gender as a significant factor causing heteroskedasticity on growth of beef cattle (Garrick et al., 1989; Rodriguez-Almeida et al., 1995). Males showed, however, a tendency for larger variability (e.g. the posterior $\Pr(\lambda_1 > 1) = 0.9378$ for the T-HE model), with absence of statistical significance perhaps attributed, at least in part, to lack of power and to poor environmental conditions, which did not allow males to express in full their extra growth potential (the average daily gain ± standard deviation was 0.432 ± 0.170 kg for males and 0.338 ± 0.131 kg for females).

In spite of these results, there were remarkable effects of heteroskedasticity on variance components, heritabilities and genetic effects inference (Sections 3.4 and 3.5). These effects may be, at least in part, due to the extra uncertainty introduced by Nelore proportion ($\lambda_2$) and heterozygosity ($\lambda_3$) effects on heteroskedasticity. The posterior $\Pr(\lambda_3 > 1) = 0.0449$ for the T-HE model provides some indication that as heterozygosity increases (and consequently heterosis) the residual variance decreases. This is consistent with the idea that heterozygosity acts as buffer against environmental variation (Lynch and Walsh, 1998).

### *3.3.2. Random effects*

Environmental residual heteroskedasticity was assessed by allowing contemporary group (CG) specific residual variances. Contemporary groups were formed such that animals included in the same CG were born in the same herd, year and season and were kept under the same environmental, management and feeding conditions throughout their productive life. Contemporary groups scaling factors were assumed to be random realizations of an inverted gamma distribution depending on the heterogeneity parameter $\eta$, as described in Equation [8]. Posterior inference on $\eta$ is presented on Table 4.2. The small range of values observed for $\eta$ in both specifications of heteroskedastic models, Gaussian and Student *t*, indicates that there is large heteroskedasticity among CG's. The largest scaling factor under T-HE was 5.57 and the smallest was 0.28, leading to ratio of about 20 times between the estimated residual variances of these two CG's.

The posterior mean of $\eta$ under the G-HE model of 3.46 was slightly smaller than the one for T-HE of 3.96. This was expected since the T-HE model attenuates the effects of outliers, which could inflate the variance within CG's. Using a structural multiplicative implementation analogous to G-HE to estimate herd specific residual variances on birth weights and calving ease scores of Piemontese cattle, Kizilkaya (2002) also found very low posterior means (<5) for $\eta$, concluding that there was heteroskedasticity across herds for these traits.

Although region was not explicitly fit as a fixed effect on the residual variance, regional heteroskedasticity can be inferred from the random CG scaling factors, because CG's are nested within region. Box plots of the random scaling factors for each of three different regions where PWG data were collected are presented in Figure 4.3. The average ± standard deviation of these random scaling factors was 0.93 ± 0.39, 0.97 ± 0.46 and 1.02 ± 0.57, respectively, for Regions 1, 2 and 3. The region specific box-plots (Figure 4.3) widely overlap and there is no indication of significant regional differences on average residual variances. The dispersion of the scaling factor, however, tends to increase from Region 1 to 2 and from 2 to 3, and clearly the number of extreme observation is larger in Region 3. This may be due, at least partially, to larger number of CG's on this region (621 in Region 3 versus 198 in Region 1 and 121 in Region 2), which increases the chance of observing extreme values of the random scaling factors inverse gamma distribution, but may also be due to true regional heteroskedasticity on random scaling factors. This possibility could be statistically verified by allowing region specific heterogeneity parameters ($\eta$) and comparing their posterior distributions.

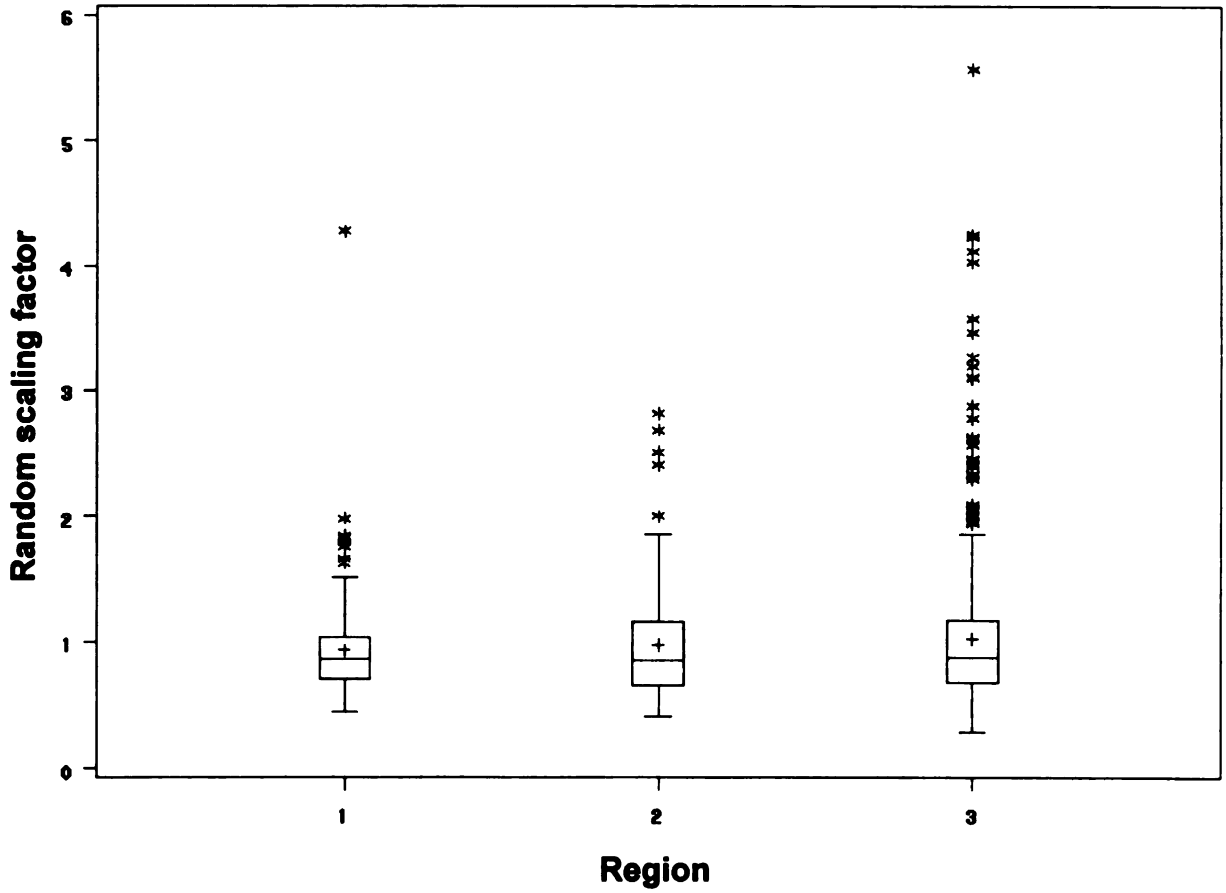

Figure 4.3. Box-plots of random contemporary group scaling factors posterior means according to the region of production: 1- located between 14°S and 16°S latitude; 2- located between 21°S and 23°S; and 3- located between 30°S and 32°S

Another possible cause of regional differences could be the level of production and environmental quality as, in general, growth traits variances tend to be proportional to means for beef cattle (Koots et al., 1994). Farms belonging to Region 1 are located in a poorer environment compared to Regions 2 and 3. This can be demonstrated by the average PWG ± standard deviation, which was 82.4 ± 21.2 kg, 105.5 ± 39.5 kg and 107.7 ± 39.3 kg, respectively for Regions 1, 2 and 3. Scaling factors tend to be larger as the mean of the respective contemporary groups increases with an estimated correlation coefficient between these two variables of 0.40. Furthermore, most of the contemporary groups with extreme residual variances $(\xi_l > 3)$ were composed exclusively by Hereford animals and located in Region 1.

### 3.4. Variance components and heritabilities

Despite the same variance-covariance structure for the random genetic effects across the six different models employed to analyze PWG, inference on genetic variance components changed considerably depending on the structural specification for residual variance. Posterior inferences based on the G-HO, T-HO, G-HE and T-HE models are presented in Tables 4.3 and 4.4 for variance components, and on Figure 4.4 for heritabilities of four breed composition groups: the purebreds Nelore and Hereford, the $F_1$ and the 3/8 Nelore 5/8 Hereford cross, which is the genotype of the Braford breed.

The 95% PPI of the marginal residual variance based on the four different models ($\sigma_E^2 = \bar{\sigma}_e^2$ for G-HO; $\sigma_E^2 = \bar{\sigma}_e^2 \lambda_1^{0.3197} \lambda_2^{0.2281} \lambda_3^{0.4347}$ for G-HE; $\sigma_E^2 = \frac{\nu}{\nu - 2} \bar{\sigma}_e^2$ for T-HO; and $\sigma_E^2 = \frac{\nu}{\nu - 2} \bar{\sigma}_e^2 \lambda_1^{0.3197} \lambda_2^{0.2281} \lambda_3^{0.4347}$ for T-HE, where 0.3197, 0.2281 and 0.4347 are, respectively, the average value of $p_{j1}$, $p_{j2}$ and $p_{j3}$) widely overlapped and point estimates (posterior means and models) were relatively constant across the different residual structure assumptions (Table 4.3). There was, nevertheless, a slight increase in the marginal residual variance using the G-HO compared to all structural models. This was not surprising since the structural models are more flexible to accommodate the extraneous variation and therefore a larger portion of this variability is expected to be included in residual variance than in other causal components of the model (Stranden and Gianola, 1999). A similar increase was observed on CG variances (Table 4.3).

Table 4.3. Posterior mean (PMEAN), posterior standard deviation (PSD), posterior mode (PMODE), 95% posterior probability intervals (PPI), and effective sample size (ESS) of additive genetic ($\sigma^2_{A_b}$ for breed $b$, $b$=1 for Nelores and $b$=2 for Herefords; $\sigma^2_{S_{12}}$ for between breed segregation), contemporary group ($\sigma^2_{cg}$) and marginal residual ($\sigma^2_E$) variance components (VC) estimated for post-weaning gain, obtained by different models.

| VC | PMEAN | PSD | PMODE | PPI | ESS |
|---|---|---|---|---|---|
| *Gaussian homoskedastic model* | | | | | |
| $\sigma^2_{A_1}$ | 23.80 | 6.86 | 22.44 | (13.02, 39.52) | 171 |
| $\sigma^2_{A_2}$ | 85.17 | 11.28 | 84.11 | (63.17, 108.45) | 223 |
| $\sigma^2_{S_{12}}$ | 8.42 | 5.84 | 4.71 | (2.32, 24.75) | 2,436 |
| $\sigma^2_{cg}$ | 897.98 | 43.31 | 888.59 | (816.69, 986.13) | 13,685 |
| $\sigma^2_E$ | 339.39 | 7.51 | 339.59 | (324.35, 353.98) | 259 |
| *Student t homoskedastic model* | | | | | |
| $\sigma^2_{A_1}$ | 46.24 | 10.90 | 47.45 | (26.77, 69.34) | 170 |
| $\sigma^2_{A_2}$ | 60.11 | 8.54 | 59.85 | (44.72, 77.46) | 191 |
| $\sigma^2_{S_{12}}$ | 7.48 | 4.76 | 4.41 | (2.27, 20.44) | 3,030 |
| $\sigma^2_{cg}$ | 946.15 | 45.43 | 938.06 | (861.00, 1038.84) | 20,324 |
| $\sigma^2_E$ | 352.08 | 8.32 | 352.62 | (335.67, 368.22) | 380 |
| *Gaussian heteroskedastic model* | | | | | |
| $\sigma^2_{A_1}$ | 119.74 | 23.52 | 123.46 | (76.97, 167.94) | 215 |
| $\sigma^2_{A_2}$ | 33.84 | 5.84 | 32.61 | (23.59, 46.69) | 213 |
| $\sigma^2_{S_{12}}$ | 9.04 | 6.65 | 4.73 | (2.35, 27.33) | 2,043 |
| $\sigma^2_{cg}$ | 927.13 | 44.79 | 919.76 | (843.80, 1018.97) | 26,191 |
| $\sigma^2_E$ | 354.82 | 12.66 | 354.18 | (330.81, 380.23) | 824 |
| *Student t heteroskedastic model* | | | | | |
| $\sigma^2_{A_1}$ | 124.87 | 21.75 | 125.55 | (82.35, 166.81) | 236 |
| $\sigma^2_{A_2}$ | 40.89 | 6.70 | 40.44 | (28.27, 54.97) | 194 |
| $\sigma^2_{S_{12}}$ | 8.03 | 5.53 | 4.53 | (2.31, 22.98) | 2,532 |
| $\sigma^2_{cg}$ | 949.30 | 46.07 | 943.69 | (863.35, 1042.94) | 25,394 |
| $\sigma^2_E$ | 346.19 | 12.18 | 345.83 | (323.05, 370.61) | 744 |

Table 4.4. Posterior mean (PMEAN), posterior standard deviation (PSD), posterior mode (PMODE), 95% posterior probability intervals (PPI), and effective sample size (ESS) of phenotypic variances ($\sigma^2_{P_g}$) estimated for post-weaning gain of different genotypes ($g$), obtained by four different models

| Parameter[a] | PMEAN | PSD | PMODE | PPI | ESS |
|---|---|---|---|---|---|
| *Gaussian homoskedastic model* | | | | | |
| $\sigma^2_{P_1}$ | 363.19 | 7.86 | 362.28 | (348.30, 379.23) | 351 |
| $\sigma^2_{P_2}$ | 424.56 | 6.44 | 424.55 | (412.15, 437.52) | 413 |
| $\sigma^2_{P_{F_1}}$ | 393.87 | 4.19 | 393.56 | (385.84, 402.29) | 1,025 |
| $\sigma^2_{P_{38N}}$ | 409.44 | 6.53 | 408.14 | (398.91, 425.27) | 1,802 |
| *Student t homoskedastic model* | | | | | |
| $\sigma^2_{P_1}$ | 398.31 | 10.63 | 398.61 | (378.35, 419.82) | 340 |
| $\sigma^2_{P_2}$ | 412.18 | 7.51 | 411.27 | (397.99, 427.39) | 989 |
| $\sigma^2_{P_{F_1}}$ | 405.25 | 7.24 | 403.71 | (391.67, 419.93) | 699 |
| $\sigma^2_{P_{38N}}$ | 413.99 | 7.98 | 413.77 | (399.61, 431.03) | 1,565 |
| *Gaussian heteroskedastic model* | | | | | |
| $\sigma^2_{P_1}$ | 507.06 | 144.60 | 431.47 | (297.88, 837.26) | 135 |
| $\sigma^2_{P_2}$ | 446.06 | 18.88 | 446.19 | (411.18, 485.25) | 2,364 |
| $\sigma^2_{P_{F_1}}$ | 370.21 | 13.51 | 368.19 | (345.21, 398.30) | 1,986 |
| $\sigma^2_{P_{38N}}$ | 420.03 | 19.56 | 420.24 | (383.34, 459.92) | 217 |
| *Student t heteroskedastic model* | | | | | |
| $\sigma^2_{P_1}$ | 588.31 | 182.89 | 443.20 | (316.66, 1008.29) | 157 |
| $\sigma^2_{P_2}$ | 444.61 | 18.41 | 442.08 | (410.21, 482.56) | 1,990 |
| $\sigma^2_{P_{F_1}}$ | 366.45 | 13.08 | 364.94 | (342.15, 393.66) | 1,102 |
| $\sigma^2_{P_{38N}}$ | 425.17 | 20.67 | 420.84 | (385.80, 466.29) | 185 |

[a] $g$ = Nelore (1), Hereford (2), $F_1$ and 3/8 Nelore (38N).

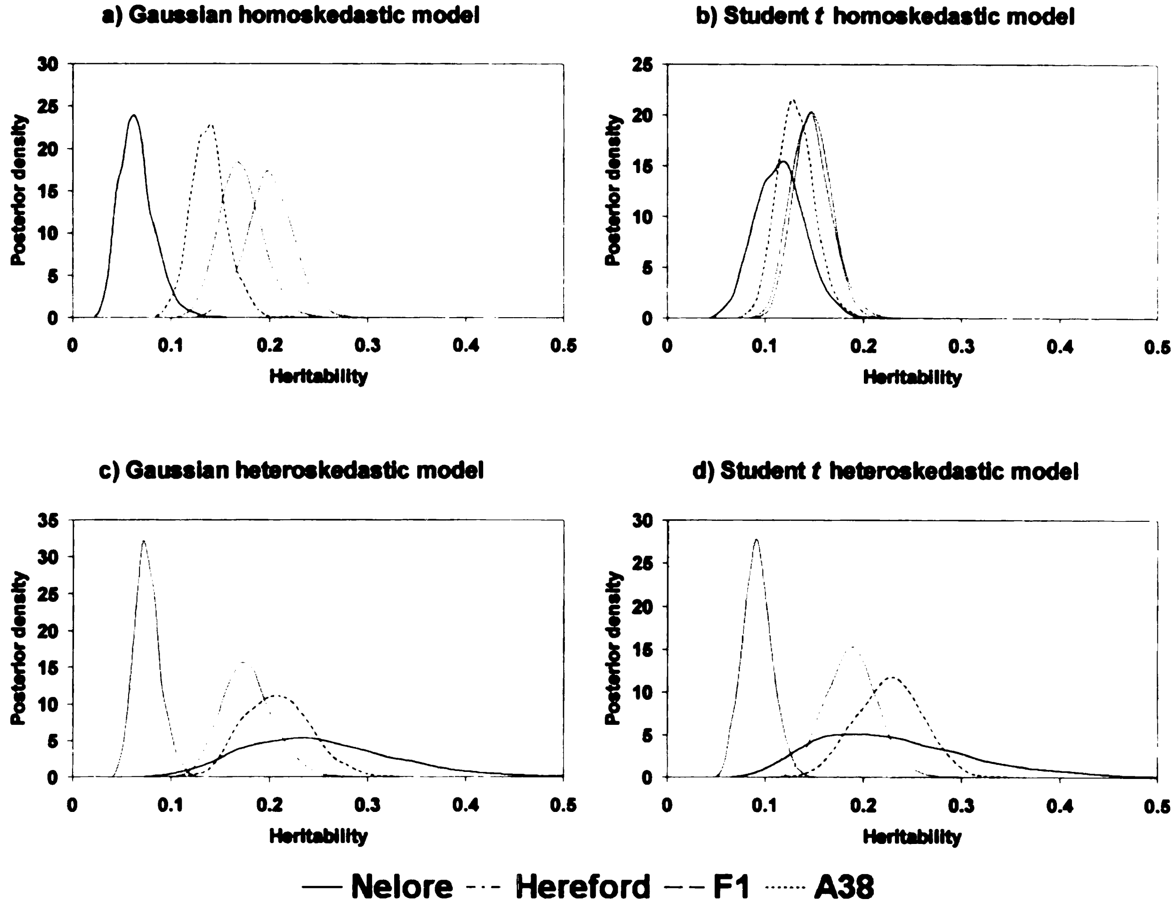

Figure 4.4. Posterior density of additive heritabilities of post-weaning gain for different breed composition groups, Nelore, Hereford, $F_1$ and Advance 3/8 Nelore (A38), obtained by a) Gaussian homoskedastic, b) Student *t* homoskedastic, c) Gaussian heteroskedastic and d) Student *t* heteroskedastic models

Among the genetic variance components, the segregation variance $\sigma^2_{S_{12}}$ was the least affected by the different model specifications and its posterior mean and mode were quite similar across the four models (Table 4.3). The genetic variances for the Nelore and Hereford breeds were widely affected by the different structural specification. Allowing for heteroskedasticity had a larger impact on inferences than allowing for robustness. The Herefords had a larger genetic variance compared to the Nelores in the G-HO and T-HO models, whereas the opposite situation was observed for the G-HE and T-HE models, i.e. the Nelores were more variable than the Hereford under these models (Table 4.3).

Despite a non significant Nelore proportion scaling factor ($\lambda_1$) under both G-HE and T-HE models (Table 4.2), the impact of heteroskedasticity is appreciable. The wide 95% PPI for the phenotypic variance of the Nelore breed obtained by the heteroskedastic models indicated poor precision for inferring upon this parameter (Table 4.4). This situation could be anticipated, since purebred Nelores are only represented by parents without records and all the information to estimate their genetic and residual variances derives solely from crossbred progeny. Due to this data structure limitation, greater uncertainty on phenotypic variances using the G-HE and T-HE models appears to be more realistic than the relatively sharp 95% PPI obtained from the G-HO and T-HO models (Table 4.4). For breed composition groups with data, the phenotypic variance is expected to be somewhat constant across models, because it has to reflect the variation noticed on the records (Falconer and Mackay, 1996). This was, to some extent, observed on Herefords, $F_1$'s and 3/8 Nelores in Table 4.4; the largest difference between the minimum and maximum posterior mean of this component was 9.6% observed for $F_1$'s between the T-HO and T-HE models. This difference is most likely due to heterozygosity effect on decreasing the residual variance of $F_1$'s.

The changes in heritability inference between models reflect the differences on genetic and phenotypic variances. The relationship between genetic variance of Hereford and Nelore breeds is further shown in the posterior density of heritabilities for these two breeds under the four different models considered (Figure 4.4). Heritability for the Nelore breed was larger than for the Hereford breed under the heteroskedastic models (G-HE and T-HE) and vice-versa for the homoskedastic models (G-HO and T-HO). The F₁'s tend to have larger heritability under the heteroskedastic models (G-HE and T-HE) compared to homoskedastic models (G-HO and T-HO), as a consequence of larger genetic variance ($\sigma^2_{A_{F_1}} = 0.5\sigma^2_{A_1} + 0.5\sigma^2_{A_2}$) and smaller $F_1$ phenotypic variance (Table 4.4) under the heteroskedastic models. The heritability of the 3/8 Nelore group was similar across the four different models. The extra uncertainty in the Nelore variance components introduced by the G-HE and T-HE models is also demonstrated for heritability inference, which presented flatter posterior distributions (Figure 4.4).

One possible reason for the dramatic change in variance components and heritabilities of Nelores and Herefords between the homoskedastic and heteroskedastic models is that most of the contemporary groups with extreme residual variances $(\xi_l > 3)$ were composed exclusively by Hereford animals. It is reasonable to assume that most of this extra variation would be captured by the Hereford genetic variance when the residual variance is assumed homoskedastic across breed groups.

### 3.5. Random additive genetic effects

Inferences on random additive genetic effects have two major purposes in animal breeding programs; they serve to rank animals for selection of parents of future generations and to predict the expected progeny difference (EPD) – a differential observed on the progeny of a particular animal relative to the population mean when this animal is mated at random to other individuals in the population. The Spearman rank correlation between additive genetic effects for the most relevant combinations of the G-HO, G-HE, T-HO and T-HE models are presented in Table 4.5, overall and for the most frequent breed compositions in the dataset. The rank correlation among the Gaussian and Student *t* counterpart models (G-HO vs. T-HO and G-HE vs. T-HE) were considerably high overall and within genotype, being always greater than 0.95 for homoskedastic models and greater than 0.98 for the heteroskedastic models. However, when we only consider animals ranked in the top 10% using the G-HO model, the rank correlation among the genetic values of these top animals by G-HO and T-HO models decreased considerably (Table 4.5). This is an expected consequence of accounting for outliers; as outlying observations often lead to extreme genetic value prediction on the animal corresponding to the record and close relatives, particularly when the outlying record is the main source of information for these individuals. This is further evident in Figure 4.5 (top graph), where posterior means additive genetic effects for Nelores, Herefords and $F_1$'s obtained by the G-HO model are plotted against the corresponding predictions using the T-HO model. In this plot we observed that several animals with extreme genetic effects under the G-HO model are shifted towards the center of their distribution under the T-HO model; e.g. the $F_1$ animal associated with observation 3 (extreme outlier described in section 3.2) has a posterior mean of -13 kg under the G-HO model but of only of -4 kg under the T-HO model (Figure 4.5 – top graph). A lower rank correlation was also observed between the G-HE and T-HE models when considering only the top 10% animals to that for all animals (Table 4.5); however, the magnitude of the decrease in correlation and the degree of change in genetic effects predictions (Figure 4.6 – top graph) were not as significant as they were between the G-HO and T-HO models. Rank correlations between the G-HO and G-HE models tend to be even smaller than between the G-HO and T-HO models (Table 4.5). One possible reason for the relatively low

Table 4.5. Spearman rank correlation between posterior mean of additive genetic effects on post-weaning gain for different combinations of the Gaussian homoskedastic (G-HO), Student *t* homoskedastic (T-HO), Gaussian heteroskedastic (G-HE) and Student *t* heteroskedastic (T-HE) models for all animals and for animals ranked in the top ten percentile for G-HO within the most frequent genotypes.

| Breed composition | N | G-HO vs. T-HO | G-HO vs. G-HE | G-HE vs. T-HE | T-HO vs. T-HE |
|---|---|---|---|---|---|
| *Including all animals within genotype* | | | | | |
| Overall | 40,082 | 0.95 | 0.83 | 0.99 | 0.92 |
| Nelore | 7445 | 0.99 | 0.97 | 0.99 | 0.98 |
| Hereford | 19,976 | 0.96 | 0.91 | 0.98 | 0.96 |
| $F_1$ | 8,718 | 0.96 | 0.86 | 0.99 | 0.93 |
| 3/8 Nelore | 1,452 | 0.96 | 0.95 | 0.99 | 0.98 |
| *Considering only animals ranked in the top 10% by G-HO within genotype* | | | | | |
| Overall | 4,008 | 0.58 | 0.39 | 0.95 | 0.77 |
| Nelore | 745 | 0.36 | 0.56 | 0.85 | 0.53 |
| Hereford | 1,998 | 0.47 | 0.44 | 0.92 | 0.85 |
| $F_1$ | 872 | 0.70 | 0.60 | 0.94 | 0.82 |
| 3/8 Nelore | 145 | 0.48 | 0.41 | 0.77 | 0.73 |

correlation between genetic predictions obtained by the G-HO and G-HE models on the top 10% animals is that, in accounting for heteroskedasticity, the G-HE model allow for more balanced selection of animals across environments, whereas a larger proportion of animals from the most variable environments tend be ranked near the top when heterogeneous variances are ignored (Gianola, 1986; Gianola et al., 1992). Moreover, the remarkable difference in genetic variances among G-HO and G-HE (Table 4.3) affects the dispersion of genetic prediction within different breed composition (Figure 4.5 – bottom graph) and consequently the manner in which these predictions overlap in the whole population, thereby decreasing the overall rank correlation. Several Hereford

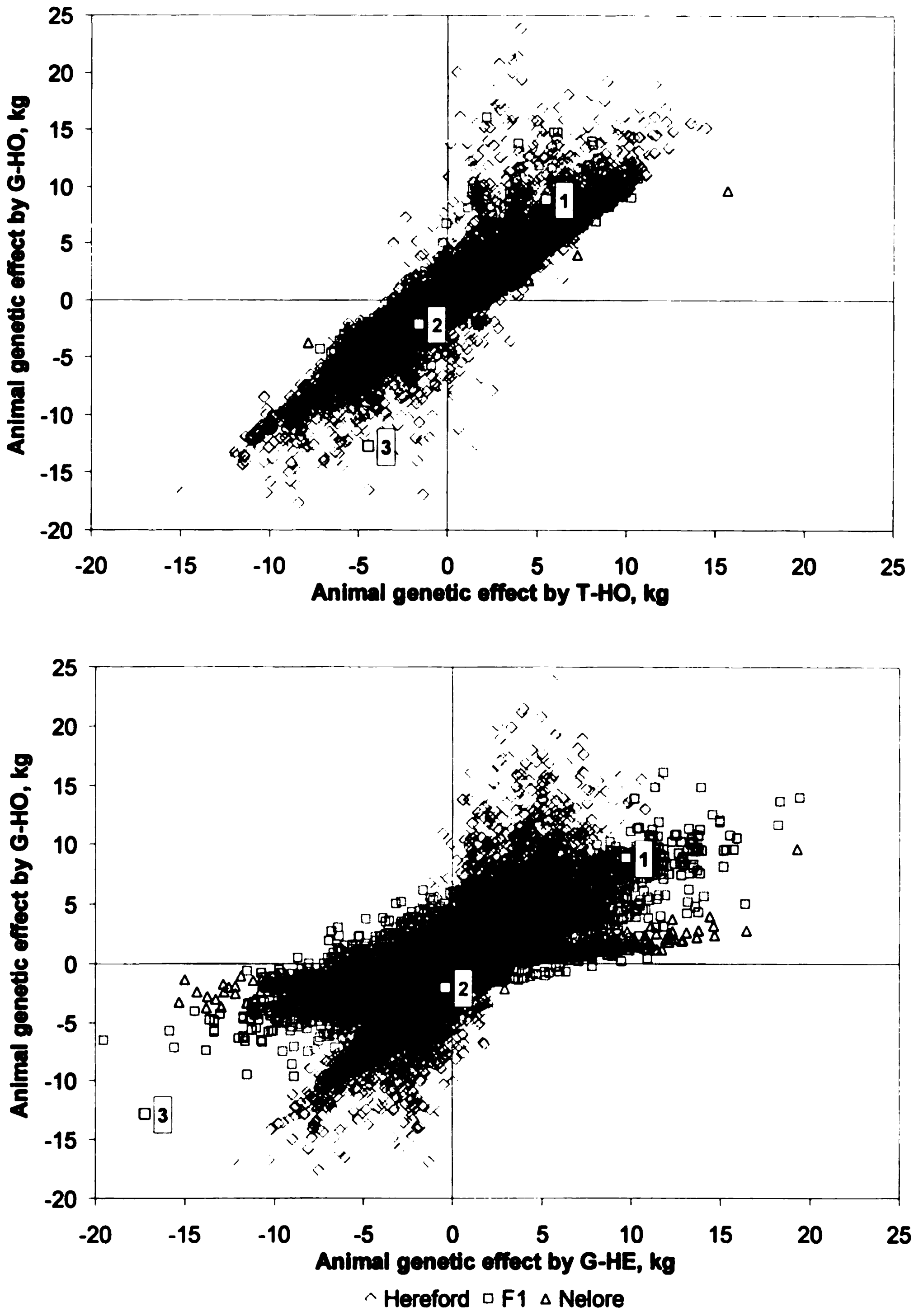

Figure 4.5. Scatter plot of posterior means of additive genetic effects for post-weaning gain obtained by the Gaussian homoskedastic (G-HO) and Student *t* homoskedastic (T-HO) models (top) and, by the G-HO and Gaussian heteroskedastic (G-HE) models (bottom), for the Nelore, Hereford and $F_1$ breed composition groups

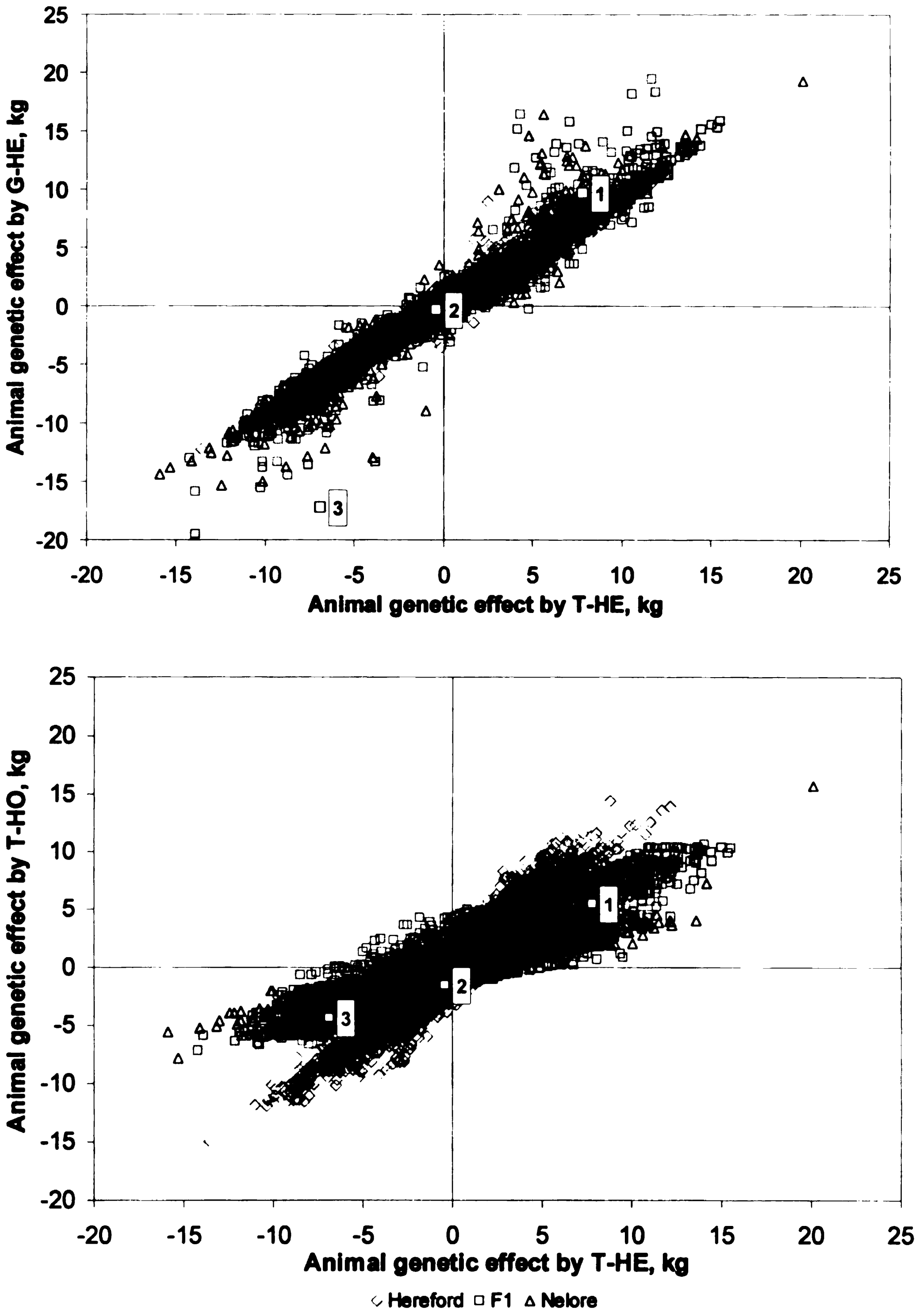

Figure 4.6. Scatter plot of posterior means of additive genetic effects for post-weaning gain obtained by the Gaussian heteroskedastic (G-HE) and Student *t* heteroskedastic (T-HE) models (top) and, by the Student *t* homoskedastic (T-HO) and T-HE models (bottom), for the Nelore, Hereford and $F_1$ breed composition groups

animals had their genetic effects predictions shifted towards the center of the distribution when comparing the G-HE and G-HO models. However, this was not the case of the genetic effect of the animal associated with observation 3, which was extreme under both Gaussian models (G-HE and G-HO).

Furthermore, the results from Table 4.5 provide additional evidence that there is overlap in accounting for heteroskedasticity and robustness, as rank correlations were higher between the Gaussian and Student *t* models under the heteroskedastic models (G-HE and T-HE) compared to their homoskedastic counterparts (G-HO and T-HO), and between homoskedastic and heteroskedastic models under robustness (T-HO and T-HE) than under Gaussian specification (G-HO and G-HE). Similar scenarios are observed in Figures 4.5 and 4.6, where the correspondence between the Gaussian and Student *t* model under the heteroskedasticity (G-HE and T-HE) was larger than under homoskedasticity (G-HO and T-HO) and was larger between homoskedastic and heteroskedastic models under robustness (T-HO and T-HE) than under Gaussian specification (G-HO and G-HE).

The relationship (slopes) among the genetic effects under different models within genotype (Figure 4.6) reflects change in genetic variability between models. As genetic variability increases the range and the spread of genetic effects also increases, e.g. the Nelore genetic variance is much larger under the G-HE model than under the G-HO model and so is the dispersion of posterior mean genetic effects of Nelores. The standard deviations (SD) of genetic effects follow similar relationships with the genetic variance as that of the genetic effects; therefore, comparisons between models in terms of these

SD follow according to their differences in genetic variances and, for brevity, results are not shown here.

### 4. Final remarks

The heteroskedastic robust models presented here combine features of the structural models presented by Foulley et al. (1992) and Kizilkaya and Tempelman (2002) with those of some heavy-tailed distributions of the Normal/independent family (Lange and Sinsheimer, 1993; Rosa et al., 2003) in a general framework. In addition to the application shown for an animal breeding problem, these models have potentially important uses in other research areas, for example gene expression with microarray data (Ibrahim et al., 2002).

We concentrated our attention on heteroskedasticity of residual variances; nonetheless, there is no conceptual difficulty to extend the structural specification to other random components of the model, such as genetic variances (as in SanCristobal et al., 1993) and contemporary groups, which have shown to be the largest source of variation in the PWG data (Table 4.3).

The use of robust models will increase the stability of the model based predictions (e.g. genetic evaluations), providing a much more appropriate treatment of outliers than simply deleting extreme records. In animal breeding, the edits used in determining which records are outliers are somewhat ad-hoc in nature and often hard to justify. For example, the ratio record/mean of its class approach advocated by Bertrand and Wiggans (1998) if applied to our PWG data would have resulted on the deletion of 1,517 records (6.7% of the data), which fall outside the range of 60% to 140%. However, only a fraction of these records should corresponds to true outliers and the approach seems too strict for this particular data set.

Finally, results from Section 3.4 and 3.5 indicate the importance of properly accounting for sources of heteroskedasticity and outliers to reliably infer upon genetic merit of crossbred animals. In our Nelore-Hereford population, inference based on the typical assumption of Gaussian homoskedastic errors (G-HO model) led to remarkably different ranking of top animals for selection compared to most appropriate Student *t* heteroskedastic errors specification (T-HE model) – the rank correlation between genetic effects obtained by these two models for animals ranked in the top 10% by the G-HO model was 0.36 – thereby hindering genetic progress.

### Appendix

*Fully conditional densities (FCD)*

In what follows the FCD are presented using the notation "*ELSE*" to denote the data vector **y** and all other parameters treated as known in the FCD in question.

Let $\boldsymbol{\theta}' = [\boldsymbol{\beta}' \quad \boldsymbol{\gamma}' \quad \mathbf{a}']$; $\mathbf{X}_1 = \{\mathbf{x}'_{1j}\}$, $\mathbf{X}_2 = \{\mathbf{x}'_{2j}\}$, and $\mathbf{Z} = \{\mathbf{z}'_j\}$, $j \in S$; Moreover, let

$$\mathbf{R}^{-1} = diag\left\{ w_j \left( \bar{\sigma}_e^2 \times \left[ \prod_{k=1}^{K} \lambda_k^{p_{jk}} \right] \times \left[ \prod_{l=1}^{L} \xi_l^{q_{jl}} \right] \right)^{-1} \right\}; \qquad \mathbf{r} = \begin{bmatrix} \mathbf{X}'_1\mathbf{R}^{-1}\mathbf{y} + \mathbf{V}_{\beta}^{-1}\boldsymbol{\beta}_o \\ \mathbf{X}'_2\mathbf{R}^{-1}\mathbf{y} + \mathbf{V}_{\gamma}^{-1}\boldsymbol{\gamma}_o \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad \text{and}$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'_1\mathbf{R}^{-1}\mathbf{X}_1 + \mathbf{V}_{\beta}^{-1} & \mathbf{X}'_1\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{X}'_1\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{X}'_2\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{X}'_2\mathbf{R}^{-1}\mathbf{X}_2 + \mathbf{V}_{\gamma}^{-1} & \mathbf{X}'_2\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + (\mathbf{G}(\boldsymbol{\varphi}))^{-1} \end{bmatrix}^{-1}.$$

Following Wang et al. (1994b), it can be shown the location parameters have the following multivariate Normal distribution,

$$\boldsymbol{\theta} \mid ELSE \sim N\left(\hat{\boldsymbol{\theta}}, \mathbf{C}\right), \qquad [A1]$$

where $\hat{\boldsymbol{\theta}} = \mathbf{Cr}$.

The FCD for the "mean" residual variance can be shown to have the following inverted gamma density:

$$p\left(\bar{\sigma}_e^2 \mid ELSE\right) \propto \left(\bar{\sigma}_e^2\right)^{-\left(\frac{n}{2}+\bar{\alpha}_e+1\right)} \exp\left(-\frac{1}{\bar{\sigma}_e^2}\left(\sum_{j\in S}\frac{w_j\left(y_j - \mathbf{x}'_{1j}\boldsymbol{\beta} - \mathbf{x}'_{2j}\mathbf{g} - \mathbf{z}'_j\mathbf{a}\right)^2}{2\prod_{k=1}^{K}\lambda_k^{p_{jk}}\prod_{l=1}^{L}\xi_l^{q_{jl}}} + \bar{\beta}_e\right)\right) \qquad [A2]$$

Moreover, the "fixed" dispersion parameters have the following FCD:

$$p(\lambda_k \mid ELSE) \propto (\lambda_k)^{-\left(\frac{\sum_{j\in S} p_{jk}}{2}+\alpha_{(\lambda)k}+1\right)} \exp\left(-\sum_{\substack{j\in S\\(p_{jk}>0)}}\frac{1}{\lambda_k^{p_{jk}}}\left(\frac{w_j\left(y_j - \mathbf{x}'_{1j}\boldsymbol{\beta} - \mathbf{x}'_{2j}\mathbf{g} - \mathbf{z}'_j\mathbf{a}\right)^2}{2\bar{\sigma}_e^2\prod_{\substack{k'=1\\k'\neq k}}^{K}\lambda_{k'}^{p_{jk'}}\prod_{l=1}^{L}\xi_l^{q_{jl}}} + \beta_{(\lambda)k}\right)\right)$$

$k$=1,..., $K$; [A3]

which is of inverted gamma form if $p_{jk}$ is an indicator variable or, equivalently, $k$ corresponds to a classification effect (e.g. gender); however [A3] is not of recognizable form if $p_{jk}$ corresponds to a continuous variable (as e.g. breed proportion or heterozygosity). In the latter case, the sampling process requires a Metropolis-Hastings (MH) step. This step was based on random walk algorithm (Chib and Greenberg, 1995) using an inverted Gamma proposal density distribution with scale parameter equal to the value of the parameter in the previous cycle times its shape parameter. The shape parameter was tuned during the burning period such that the acceptance of proposal values was intermediate for optimal MCMC mixing (Chib and Greenberg, 1995).

The FCD of random multiplicative effects on residual variance are of inverted gamma form (Kizilkaya and Tempelman, 2002), as follows:

$$p(\xi_l \mid ELSE) \propto (\xi_l)^{-\left(\frac{\sum_{j\in S} q_{jl}}{2}+\eta+1\right)} \exp\left(-\frac{1}{\xi_l}\left(\sum_{\substack{j\in S\\ (q_{jl}=1)}} \frac{w_j\left(y_j-\mathbf{x}'_{1j}\boldsymbol{\beta}-\mathbf{x}'_{2j}\mathbf{g}-\mathbf{z}'_j\mathbf{a}\right)^2}{2\bar{\sigma}_e^2\prod_{k'=1}^{K}\lambda_k^{p_{jk}}}+\eta-1\right)\right)$$

$l$=1,..., $L$; [A4]

The FCD for the weights $w_j$'s depend on the choice of $p\left(w_j \mid \nu\right)$. Adopting [9], i.e. the Student $t$ specification, these FCD correspond to series of Gamma densities, given by:

$$p\left(w_j \mid ELSE\right) \propto w_j^{\left(\frac{\nu+1}{2}-1\right)} \exp\left(-w_j\frac{1}{2}\left(\nu+\frac{\left(y_j-\mathbf{x}'_{1j}\boldsymbol{\beta}-\mathbf{x}'_{2j}\mathbf{g}-\mathbf{z}'_j\mathbf{a}\right)^2}{\bar{\sigma}_e^2\prod_{k=1}^{K}\lambda_k^{p_{jk}}\prod_{l=1}^{L}\xi_l^{q_{jl}}}\right)\right),$$

$j \in S, w_j$ >0. [A5]

On the other hand, adopting [10] as the $p\left(w_j \mid \nu\right)$, we have truncated Gamma densities, given by:

$$p\left(w_j \mid ELSE\right) \propto w_j^{\left(\frac{\nu+1}{2}-1\right)} \exp\left(-w_j\frac{\left(y_j-\mathbf{x}'_{1j}\boldsymbol{\beta}-\mathbf{x}'_{2j}\mathbf{g}-\mathbf{z}'_j\mathbf{a}\right)^2}{2\bar{\sigma}_e^2\prod_{k=1}^{K}\lambda_k^{p_{jk}}\prod_{l=1}^{L}\xi_l^{q_{jl}}}\right),$$

$j \in S, 0 < w_j \leq 1.$ [A6]

One suitably and relatively "noninformative" prior on $\eta$ is $\eta \sim Gamma\left(\alpha_{(\eta)}, \beta_{(\eta)}\right)$, $\eta > 1$, with small values of $\alpha_{(\eta)}$ and $\beta_{(\eta)}$ and $\beta_{(\eta)}$ << $\alpha_{(\eta)}$ (Liu, 1996). Using this specification, the FCD of $\eta$ is given by:

$$p(\eta \mid ELSE) \propto \frac{(\eta-1)^{\eta L}\eta^{\alpha_{(\eta)}-1}}{\Gamma(\eta)^{L}}\prod_{l=1}^{L}(\xi_l)^{-(\eta+1)}\exp\left(-(\eta-1)\sum_{l=1}^{L}\xi_l^{-1}-\beta_{(\eta)}\eta\right), \qquad [A7]$$

which does not have recognizable and also requires MH sampling. In this case we used a transformation strategy, and sampled $\psi = \log(\eta)$ using a random walk sampler with a Gaussian proposal density centered at the value of $\psi$ in the previous cycle and with variance tuned during MCMC burn-in for optimal MCMC mixing (Chib and Greenberg, 1995).

The FCD for the robustness parameter $\nu$ depends on the choices of $p(w_j \mid \nu)$ and $p(\nu)$. Similarly to $\eta$, we chose the prior on $\nu$ to be $\nu \sim Gamma\left(\alpha_{(\nu)}, \beta_{(\nu)}\right)$, $\nu > 0$. Under the Student *t* specification (i.e. $p(w_j \mid \nu)$ as in [9]), the FCD of $\nu$ is given by:

$$p(\nu \mid ELSE) \propto \frac{\nu^{n\nu/2+\alpha_{(\nu)}-1}}{\Gamma(\nu/2)^{n}\, 2^{n\nu/2}}\prod_{j\in S} w_j^{\nu/2}\exp\left(-\nu/2\left(\sum_{j\in S} w_j+\beta_{(\nu)}\right)\right), \qquad [A8]$$

which does not have a recognizable form, but can be sampled using a MH algorithm similar to the one used on [A7].

Alternatively, adopting the Slash specification (i.e. $p(w_j \mid \nu)$ as in [10]), the FCD of $\nu$ is given by :

$$p(\nu \mid ELSE) \propto \nu^{n+\alpha_{(\nu)}-1}\exp\left(-\nu\left(\beta_{(\nu)}-\sum_{j\in S}\log(w_j)\right)\right), \qquad [A9]$$

which is a gamma distribution .

The reader interested in the FCD of genetic variances is referred to Chapter 3, which presented details on these FCD and a MH scheme. More details on the MH implementations and a FORTRAN 90 code can be obtained from the author upon request.

# CONCLUSIONS

## 1. This study in the context of beef cattle breeding

Genetic evaluation programs for multiple-breed beef cattle populations selected for various economically important traits are required to improve efficiency and competitiveness in the modern beef industry. Beef cattle performance programs are usually carried out on diverse production systems and environments, with measurements and data quality often compromised by the occurrence of recording error, preferential treatment, the effect of injury or disease, and pedigree errors. Hierarchical models present a general framework to address problems arising from the nature of field data structure; a variety of multistage propositions have been advocated to handle issues such as uncertain paternity (Foulley et al., 1987; Henderson, 1988), heterogeneity of variance (Foulley et al., 1992; Foulley and Quaas, 1995; Gianola et al., 1992; SanCristobal et al., 1993), and outlying observations (Rosa, 1999; Stranden and Gianola, 1998, 1999), for instance. In Chapters 1 and 2, methodology for genetic evaluation using a fully Bayesian approach was proposed and applied to the prediction of genetic merit of animals having uncertain paternity. Similar to the empirical Bayes sire model method of Foulley et al. (1987), the procedure combines data and prior information to determine posterior probabilities of sire assignments, whereas Henderson's method (Henderson, 1988) is solely based on prior information. Nevertheless, our method represents an important extension since it uses more recently developed MCMC tools to provide small sample (i.e. non-asymptotic) inference based on the animal model, the most common model for current genetic evaluations. In Chapter 4, we present hierarchical Bayes models combining in a general framework features of the structural models presented by Foulley et al. (1992) and

Kizilkaya and Tempelman (2002) with heavy-tailed distributions of the Normal/independent family (Lange and Sinsheimer, 1993) to allow heteroskedastic and robust inference on genetic merit. Heteroskedasticity and robustness have been addressed individually in the past.

Crossbreeding is a key tool available to increase the efficiency of production through heterosis and complementarity between breeds (Gregory et al., 1999). A hierarchical Bayes model to predict performance on crossbred beef populations, based on additive and non-additive genotypic effects and additive genetic individual deviations was proposed in Chapter 3. The ability to combine data and literature information and the implementation of a more realistic modeling of the additive genetic variability and correlation between relatives on crossbred populations (Lo et al., 1993) are distinctive features of this model that will help to improve accuracy of genetic predictions and, consequently, selection response (Falconer and Mackay, 1996).

## 2. Objectives revisited

1) To develop and apply a hierarchical Bayes model for genetic evaluation of animals originated from multiple-sire mating systems.

This objective is particularly relevant for extensive beef cattle production systems, which often rely upon multiple-sire mating to increase the probability of pregnancy, as the size of breeding groups is often too large to be sired by single bulls. Statistical modeling developments and validation referring to this objective were accomplished by

Chapter 1, and an example of application to beef cattle genetic evaluation was presented in Chapter 2.

2) To develop and apply a hierarchical Bayes model for genetic evaluation of animals in multiple-breed populations.

This objective was fully addressed by Chapter 3. In this chapter, a multistage hierarchical Bayes construction of the multiple-breed animal model from Lo et al. (1993) was proposed to estimate genotypic effects and individual additive deviations when breed and segregation variance components are unknown. This model was validated using simulated data and applied to a dataset of post-weaning gains on purebred and crossbred animals derived from Hereford and Nelore cattle and raised in diverse environments. One limitation was the computational time required to obtain the necessary number of samples to do MCMC based inference. However, an empirical Bayes approach could be adopted for genetic evaluation of large beef populations. In this case, variance components could be estimated from a data subset using MCMC, followed by the use of Henderson's mixed model equations (Equation [A1] in the appendix to Chapter 3) to provide empirical best linear unbiased estimates (BLUE) for genotypic effects and prediction (BLUP) of individual additive deviations.

3) To extend the genetic evaluation models in 1) and 2) to account for heterogeneity of residual variances across environments and provide greater robustness to outliers.

A structural framework extending hierarchical models to account for residual heteroskedastic and to provide robustness to outliers was proposed in Chapter 4. Despite our developments being concentrated on the hierarchical multiple-breed animal model (Objective 2), these are readily extendable to the case of uncertain paternity, merging the developments in Chapters 1 and 4.

### 3. Implications for genetic improvement of beef cattle

Beef cattle herds raised on pastoral conditions are often subjected to multiple-sire mating. Currently, about 25-30% of the calves evaluated by the beef cattle improvement programs in Brazil derive from multiple-sire matings. For these herds, the uncertain paternity hierarchical model (HIER) proposed in this study represents an important alternative for genetic prediction. In addition to the incorporation of prior probabilities on sire assignments, the HIER model allows for the integration of the uncertainty about these prior probabilities in the prediction of genetic merit. Genetic markers, for example, represent an important objective source of prior information. Results from the simulation study (Chapter 1) and from the analysis of growth records on Brazilian Herefords (Chapter 2) indicated that the HIER model provided a better fit to data characterized by the presence of uncertain paternity compared to a model based on the average numerator relationship matrix (ANRM) (Henderson, 1988). For genetic evaluations, there may be pragmatically little difference between models for rankings of predicted genetic values. The main advantage of the HIER model is in terms of properly accounting for reduced precision on genetic merit inference due to uncertainty on sire assignments, thereby providing a better risk assessment for the decision process in terms of selection and the number of mates assigned to each selected animal. The estimated reliabilities associated with genetic values of animals with uncertain paternity would tend to be appreciably lower using the HIER model compared to the ANRM model, since the latter assumes that the true probabilities of paternity are known.

Most of the beef produced in the US, Brazil and other countries is derived form crossbred animals. Genetic evaluation of multiple-breed populations is, however, complicated by the different genetic backgrounds and degrees of crossing present in these populations. Confoundedness and multicollinearity between the coefficients for genotypic effects makes it difficult to precisely estimate such effects solely from data on multiple-breed cattle. Moreover, in order to predict animal additive genetic effects, it is crucial to properly model genetic covariances between crossbred relatives as specified by Lo et al. (1993). The hierarchical multiple-breed animal model presented in Chapter 3 effectively combines data and prior information to predict genetic merit and provides a useful framework for inference on multiple-breed genetic variances. This model specifies the additive genetic variance of each breed composition group as a function of breed-specific and segregation variances, thereby sufficiently characterizing the genetic heteroskedasticity of these groups in crossbred populations. In contrast, the conventional animal model assumes constant genetic variances across groups and no segregation variance. Accordingly, the proposed hierarchical model enhances flexibility for modeling the dispersion of genetic merit within breed groups, thereby having important implications for improved precision on prediction of genetic merit. Furthermore, prior information on genotypic effects, as it is available from the literature, might be useful for analyses of poorly structured datasets as is common for crossbred beef cattle (Quaas and Pollak, 1999) and might also further mitigate the effects of multicollinearity amongst genetic effects coefficients.

The appropriate treatment of sources of heteroskedasticity and to outliers provided by the hierarchical Bayes model with structural residual variances (Chapter 4) would increase the stability of genetic evaluations. This is particularly relevant for beef cattle populations, since the diversity of environmental, management and feeding conditions to which the animals are subjected during their productive life creates several possible sources of heteroskedasticity and other data perturbations. Results from Chapter 4 have provided strong evidence to the importance of properly accounting for sources of heteroskedasticity and outliers to accurately infer upon genetic merit of crossbred animals. In the studied Nelore-Hereford population, inference based on the typical assumption of Gaussian homoskedastic errors led to remarkable rerankings of animals for selection compared to most appropriate Student *t* heteroskedastic errors specification. These results have potentially important implications for genetic improvement programs based on conventional genetic evaluation models.

### 4. Opportunities for further studies

The hierarchical Bayes model presented in Chapter 1 provides a general framework to account for uncertain paternity. One potentially relevant extension of this model is to directly integrate genetic marker information in the model as was proposed through prior distribution specifications. This would allow, for example, to assess the quality of the marker information contributing to determine sire assignment (Rosa et al., 2002).

For large scale genetic evaluations on a national or breed association level, it appears that a computationally tractable empirical Bayes or "plug-in" strategy may be advisable and would likely lead to potentially very little or no difference in estimated breeding values and standard errors of prediction relative to MCMC based inference. This implementation deserves further investigation with the marginalization over prior probabilities of sire assignments employed by Foulley et al. (1987) as a suitable strategy to accomplish this task.

The multiple-breed animal model of Chapter 3 is presented in a single trait context; however, generalization for the case of multiple-traits or additive-maternal genetic effects could be attained by using multiple-breed variance-covariance genetic matrices as proposed by Cantet and Fernando (1995) following Lo et al. (1993) and a Wishart proposal density in the Metropolis-Hastings algorithm.

Given that uncertain paternity is also seen in multiple-breed populations, another promising development is to combine the features of the uncertain paternity model of Chapter 1 with those of the multiple-breed animal model of Chapter 3. The hierarchical structure of these models facilitates this task. Under multiple-sire mating, uncertainty is introduced on various elements of $\mathbf{G}(\boldsymbol{\varphi})$, thru $\mathbf{P}$ and $\boldsymbol{\Omega}(\boldsymbol{\varphi})$, and possibly on elements of $\mathbf{X}_2$, if not all possible sires have the same breed composition. In this situation, the sampling density (first stage) would be conditioned on the sire assignment, i.e. $s_j^* = s_j^{(k)}$, $1 \leq k \leq v_j$, for animal $j$. For example, Equation [3] of Chapter 3 would be replaced by:

$$y_j \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{a}, s_j^* = s_j^{(k)}, \sigma_e^2 \sim N\left(\mathbf{x}'_{1j}\boldsymbol{\beta} + \mathbf{x}'^{(k)}_{2j}\boldsymbol{\gamma} + \mathbf{z}'_j\mathbf{a}, \sigma_e^2\right), \quad j \in S,$$

where $\mathbf{x}_{2j}^{\prime(k)}$ is $\mathbf{x}_{2j}^{\prime}\big|_{s_j^*=s_j^{(k)}}$, i.e. with coefficients of genetic "fixed" effects based on the sire assignment $s_j^* = s_j^{(k)}$. The prior distribution of $\mathbf{a}$ (in the second stage) would also change from $\mathbf{a}\,|\,\boldsymbol{\varphi} \sim N\left(\mathbf{0},\mathbf{G}(\boldsymbol{\varphi})\right)$ to $\mathbf{a}\,|\,\boldsymbol{\varphi},\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})} \sim N\left(\mathbf{0},\mathbf{G}^{(\mathbf{k})}(\boldsymbol{\varphi})\right)$. Here, the notation is that of Chapter 1, and the extra developments required as a consequence of the conditioning of $\mathbf{G}(\boldsymbol{\varphi})$ on sire assignments for all animals with uncertain paternity $\mathbf{s}^* = \mathbf{s}^{(\mathbf{k})}$, given that $\mathbf{G}^{(\mathbf{k})}(\boldsymbol{\varphi}) = \mathbf{G}(\boldsymbol{\varphi})\big|_{\mathbf{s}^*=\mathbf{s}^{(\mathbf{k})}}$, would follow directly from the decomposition of $\mathbf{G}(\boldsymbol{\varphi})$. Moreover, the additional stages required to sample sire assignments would be analogous to those presented in Chapter 1.

Finally, despite the attention in Chapter 4 being centered on heteroskedasticity of residual variances, there is no conceptual difficulty to extend the structural specification to other random components of the model, such as genetic variances (as in SanCristobal et al., 1993) and the contemporary group variance, which have shown to be the largest source of variation in the Nelore-Hereford post-weaning data.

# BIBLIOGRAPHY

Arnold, J. W., Bertrand, J. K., and Benyshek, L. L. 1992. Animal-model for genetic evaluation of multibreed data. Journal of Animal Science, 70(11): 3322-3332.

Arthur, P. F., Hearnshaw, H., and Stephenson, P. D. 1999. Direct and maternal additive and heterosis effects from crossing *Bos indicus* and *Bos taurus* cattle: cow and calf performance in two environments. Livestock Production Science, 57(3): 231-241.

Bertrand, J. K. and Wiggans, G. R. 1998. Validation of data and review of results from genetic evaluation systems for US beef and dairy cattle. Paper presented at the Proceedings of 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia.

Birchmeier, A. N., Cantet, R. J. C., Fernando, R. L., Morris, C. A., Holgado, F., Jara, A., and Cristal, M. S. 2002. Estimation of segregation variance for birth weight in beef cattle. Livestock Production Science, 76(1-2): 27-35.

Blasco, A. 2001. The Bayesian controversy in animal breeding. Journal of Animal Science, 79(8): 2023-2046.

Cantet, R. J. C., Gianola, D., Misztal, I., and Fernando, R. L. 1993. Estimates of dispersion parameters and of genetic and environmental trends for weaning weight in Angus cattle using a maternal animal-model with genetic grouping. Livestock Production Science, 34(3-4): 203-212.

Cantet, R. J. C. and Fernando, R. L. 1995. Prediction of breeding values with additive animal-models for crosses from 2 Populations. Genetics Selection Evolution, 27(4): 323-334.

Cardoso, F. F., Cardellino, R. A., and Campos, L. T. 2001. (Co)Variance components and genetic parameters for weaning production traits of Angus calves raised in the state of Rio Grande do Sul. Brazilian Journal of Animal Science, 30(1): 41-48.

Cardoso, F. F. and Tempelman, R. J. 2001. Bayesian inference on uncertain paternity for prediction of genetic merit. Journal of Animal Science, 79 Suppl. 1: 111.

Chib, S. and Greenberg, E. 1995. Understanding the Metropolis-Hastings algorithm. American Statistician, 49(4): 327-335.

Cunningham, E. P. 1987. Crossbreeding - the Greek temple model. Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie, 104(1-2): 2-11.

de Mattos, D., Misztal, I., and Bertrand, J. K. 2000. Variance and covariance components for weaning weight for Herefords in three countries. Journal of Animal Science, 78(1): 33-37.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via EM algorithm. Journal of the Royal Statistical Society Series B-Methodological, 39(1): 1-38.

DeNise, S. 1999. Using parentage analysis in commercial beef operations. Paper presented at the Beef Improvement Federation 31st Annual Research Symposium & Annual Meeting, Roanoke, Virginia.

Dickerson, G. E. 1969. Experimental approaches in utilising breed resources. Animal Breeding Abstracts, 37: 191-202.

Dickerson, G. E. 1973. Inbreeding and heterosis in animals. Paper presented at the Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. J. L. Lush, Champaign, IL.

Eler, J. P., Vanvleck, L. D., Ferraz, J. B. S., and Lobo, R. B. 1995. Estimation of variances due to direct and maternal effects for growth traits of Nelore cattle. Journal of Animal Science, 73(11): 3253-3258.

Elzo, M. A. 1994. Restricted maximum-likelihood procedures for the estimation of additive and nonadditive genetic variances and covariances in multibreed populations. Journal of Animal Science, 72(12): 3055-3065.

Elzo, M. A., Manrique, C., Ossa, G., and Acosta, O. 1998. Additive and nonadditive genetic variability for growth traits in the Turipana Romosinuano-Zebu multibreed herd. Journal of Animal Science, 76(6): 1539-1549.

Elzo, M. A. and Wakeman, D. L. 1998. Covariance components and prediction for additive and nonadditive preweaning growth genetic effects in an Angus- Brahman multibreed herd. Journal of Animal Science, 76(5): 1290-1302.

Falconer, D. S. and Mackay, T. F. C. 1996. Introduction to quantitative genetics (4 ed.). Harlow: Longman Group Ltd.

Famula, T. R. 1992. Simple and rapid inversion of additive relationship matrices incorporating parental uncertainty. Journal of Animal Science, 70(4): 1045-1048.

Famula, T. R. 1993. The contribution of progeny of uncertain paternity to the accuracy of sire evaluation. Journal of Animal Science, 71(5): 1136-1141.

Fernandez, C. and Steel, M. F. J. 1998. On Bayesian modeling of fat tails and skewness. Journal of the American Statistical Association, 93(441): 359-371.

Fernando, R. L. 1999. Theory for analysis of multi-breed data. Paper presented at the 7th Genetic Prediction Workshop, Kansas City, MO.

Foulley, J. L., Gianola, D., and Planchenault, D. 1987. Sire evaluation with uncertain paternity. Genetics Selection Evolution, 19(1): 83-102.

Foulley, J. L., Thompson, R., and Gianola, D. 1990. On sire evaluation with uncertain paternity. Genetics Selection Evolution, 22(3): 373-376.

Foulley, J. L., Cristobal, M. S., Gianola, D., and Im, S. 1992. Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. Computational Statistics & Data Analysis, 13(3): 291-305.

Foulley, J. L. and Quaas, R. L. 1995. Heterogeneous variances in Gaussian linear mixed models. Genetics Selection Evolution, 27(3): 211-228.

Garrick, D. J., Pollak, E. J., Quaas, R. L., and Vanvleck, L. D. 1989. Variance heterogeneity in direct and maternal weight traits by sex and percent purebred for Simmental-sired calves. Journal of Animal Science, 67(10): 2515-2528.

Gelfand, A. E. and Smith, A. F. M. 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85(410): 398-409.

Gelfand, A. E. 1996. Model determination using sampling-based methods. In W. R. Gilks and S. Richardson and D. J. Spiegelhalter (Eds.), Markov Chain Monte Carlo in practice, 1st ed.: 145-161. London: Champman & Hall.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(6): 721-741.

Geyer, C. J. 1992. Practical Markov Chain Monte Carlo. Statistical Science, 7(4): 473-511.

Gianola, D. 1986. On selection criteria and estimation of parameters when the variance is heterogeneous. Theoretical and Applied Genetics, 72(5): 671-677.

Gianola, D. and Fernando, R. L. 1986. Bayesian methods in animal breeding theory. Journal of Animal Science, 63(1): 217-244.

Gianola, D., Foulley, J. L., Fernando, R. L., Henderson, C. R., and Weigel, K. A. 1992. Estimation of heterogeneous variances using empirical Bayes methods - Theoretical considerations. Journal of Dairy Science, 75(10): 2805-2823.

Gilks, W. R., S. Richard, and D. J. Spiegelhalter. 1996. Markov Chain Monte Carlo in practice. New York: Chapman and Hall.

Gregory, K. E., Cundiff, L. V., and Koch, R. M. 1999. Composite breeds to use heterosis and breed differences to improve efficiency of beef production: 75. Clay Center, NE: MARC-USDA-ARS.

Hasting, W. K. 1970. Monte Carlo sampling methods using Markov Chains and their applications. Biometrika, 57: 97-109.

Henderson, C. R. 1973. Sire evaluation and genetic trends. Paper presented at the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush, Champaign, IL.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 31: 423-447.

Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics, 32: 69-83.

Henderson, C. R. 1988. Use of an average numerator relationship matrix for multiple-sire joining. Journal of Animal Science, 66(7): 1614-1621.

Hill, W. G. 1982. Dominance and epistasis as components of heterosis. Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics, 99(3): 161-168.

Hobert, J. P. 2000. Hierarchical models: A current computational perspective. Journal of the American Statistical Association, 95(452): 1312-1316.

Ibrahim, J. G., Chen, M. H., and Gray, R. J. 2002. Bayesian models for gene expression with DNA microarray data. Journal of the American Statistical Association, 97(457): 88-99.

Im, S. 1992. Mixed linear-model with uncertain paternity. Applied Statistics-Journal of the Royal Statistical Society Series C, 41(1): 109-116.

Jensen, J., Wang, C. S., Sorensen, D. A., and Gianola, D. 1994. Bayesian-inference on variance and covariance components for traits influenced by maternal and direct genetic-effects, using the Gibbs sampler. Acta Agriculturae Scandinavica Section a-Animal Science, 44(4): 193-201.

Kerr, R. J., Graser, H. U., Kinghorn, B. P., and Johnston, D. J. 1994a. Implications of using an average relationship matrix in genetic evaluation for a population using multiple-sire matings. Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie, 111(3): 199-208.

Kerr, R. J., Hammond, K., and Kinghorn, B. P. 1994b. Effects of multiple-sire matings on genetic evaluations, selection response and rates of inbreeding. Livestock Production Science, 38(3): 161-168.

Kinghorn, B. 1980. The expression of recombination loss in quantitative traits. Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie-Journal of Animal Breeding and Genetics, 97(2): 138-143.

Kinghorn, B. P. 1987. The nature of 2-locus epistatic interactions in animals - Evidence from Sewall Wright Guinea-Pig data. Theoretical and Applied Genetics, 73(4): 595-604.

Kizilkaya, K. 2002. Hierarchical Bayesian threshold models applied to the quantitative genetic analysis of calving ease scores of Italian Piemontese cattle. Michigan State University, East Lansing.

Kizilkaya, K., Banks, B. D., Carnier, P., Albera, A., Bittante, G., and Tempelman, R. J. 2002. Bayesian inference strategies for the prediction of genetic merit using threshold models with an application to calving ease scores in Italian Piemontese cattle. Journal of Animal Breeding and Genetics, 119(4): 209-220.

Kizilkaya, K. and Tempelman, R. J. 2002. Bayesian heteroskedastic generalized linear models for animal breeding applications. Paper presented at the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France.

Kizilkaya, K. and Tempelman, R. J. 2003. An assessment of heteroskedastic t-error linear mixed models for the analysis of field data collected from diverse environments. Paper presented at the International Biometric Society - ENAR Spring Meeting, Tampa, FL.

Klei, L., Quaas, R. L., Pollak, E. J., and Cunningham, B. E. 1996. Multiple-breed evaluation. Paper presented at the Beef Improvement Federation 28th Annual Research Symposium & Annual Meeting, Birmingham, AL.

Koch, R. M., Dickerson, G. E., Cundiff, L. V., and Gregory, K. E. 1985. Heterosis retained in advanced generations of crosses among Angus and Hereford cattle. Journal of Animal Science, 60(5): 1117-1132.

Koots, K. R., Gibson, J. P., Smith, C., and Wilton, J. W. 1994. Analyses of published genetic parameter estimates for beef cattle production traits. 1. Heritability. Animal Breeding Abstracts, 62(5): 309-338.

Lange, K. and Sinsheimer, J. S. 1993. Normal/independent distributions and their applications in robust regression. Journal of the American Statistical Association, 2(2): 175-198.

Liu, C. H. 1996. Bayesian robust multivariate linear regression with incomplete data. Journal of the American Statistical Association, 91(435): 1219-1227.

Lo, L. L., Fernando, R. L., and Grossman, M. 1993. Covariance between relatives in multibreed populations - Additive-model. Theoretical and Applied Genetics, 87(4): 423-430.

Lo, L. L., Fernando, R. L., Cantet, R. J. C., and Grossman, M. 1995. Theory for modeling means and covariances in a 2-breed population with dominance inheritance. Theoretical and Applied Genetics, 90(1): 49-62.

Lo, L. L., Fernando, R. L., and Grossman, M. 1997. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. Journal of Animal Science, 75(11): 2877-2884.

Lutaaya, E., Misztal, I., Mabry, J. W., Short, T., Timm, H. H., and Holzbauer, R. 2001. Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. Journal of Animal Science, 79(12): 3002-3007.

Lynch, M. and Walsh, B. 1998. Genetics and analysis of quantitative traits (1st ed.). Sunderland, MA: Sinauer Associates, Inc.

Metropolis, N., A. W. Rosenbulth, A. H. Teller, E. Teller. 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics, 21: 1087.

Meuwissen, T. H. E. and Luo, Z. 1992. Computing inbreeding coefficients in large populations. Genetics Selection Evolution, 24(4): 305-313.

Meyer, K. 1992. Variance-components due to direct and maternal effects for growth traits of Australian beef-cattle. Livestock Production Science, 31(3-4): 179-204.

Meyer, K. 1993. Covariance matrices for growth traits of Australian Polled Hereford cattle. Animal Production, 57: 37-45.

Miller, S. P. and Wilton, J. W. 1999. Genetic relationships among direct and maternal components of milk yield and maternal weaning gain in a multibreed beef herd. Journal of Animal Science, 77(5): 1155-1161.

Natarajan, R. and Kass, R. E. 2000. Reference Bayesian methods for generalized linear mixed models. Journal of the American Statistical Association, 95(449): 227-237.

Nunez-Dominguez, R., Vanvleck, L. D., and Cundiff, L. V. 1995. Prediction of genetic values of sires for growth traits of crossbred cattle using a multivariate animal-model with heterogeneous variances. Journal of Animal Science, 73(10): 2940-2950.

Perez-Enciso, M. and Fernando, R. L. 1992. Genetic evaluation with uncertain parentage - A comparison of methods. Theoretical and Applied Genetics, 84(1-2): 173-179.

Quaas, R. L. and Pollak, E. J. 1980. Mixed model methodology for farm and ranch beef cattle testing programs. Journal of Animal Science, 51(6): 1277-1287.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. Journal of Dairy Science, 71(5): 1338-1345.

Quaas, R. L. and Pollak, E. J. 1999. Application of a multi-breed genetic evaluation. Paper presented at the 7th Genetic Prediction Workshop, Kansas City, MO.

Reverter, A., Tier, B., Johnston, D. J., and Graser, H. U. 1997. Assessing the efficiency of multiplicative mixed model equations to account for heterogeneous variance across herds in carcass scan traits from beef cattle. Journal of Animal Science, 75(6): 1477-1485.

Rodriguez-Almeida, F. A., Vanvleck, L. D., Cundiff, L. V., and Kachman, S. D. 1995. Heterogeneity of variance by sire breed, sex, and dam breed in 200-day and 365-day weights of beef-cattle from a top cross experiment. Journal of Animal Science, 73(9): 2579-2588.

Rogers, W. H. and Tukey, J. W. 1972. Understanding some long-tailed distributions. Statistica Neerlandia, 26: 211-226.

Rosa, G. J. M. 1999. Robust mixed linear models in quantitative genetics: Bayesian analysis via Gibbs sampling. Paper presented at the International symposium on animal breeding and genetics, Vicosa, MG, Brazil.

Rosa, G. J. M., Yandell, B. S., and Gianola, D. 2002. A Bayesian approach for constructing genetic maps when markers are miscoded. Genetics Selection Evolution, 34(3): 353-369.

Rosa, G. J. M., Padovani, C. R., and Gianola, D. 2003. Robust linear mixed models with Normal/Independent distributions and Bayesian MCMC implementation. Biometrical Journal, 45(4): 573-590.

Roso, V. M. and Fries, L. A. 1998. Maternal and individual heterozygosities and heterosis on preweaning gain of Angus x Nelore calves. Paper presented at the World Congress On Genetics Applied To Livestock Production, Armidale.

SanCristobal, M., Foulley, J. L., and Manfredi, E. 1993. Inference about multiplicative heteroskedastic components of variance in a mixed linear Gaussian model with an application to beef-cattle breeding. Genetics Selection Evolution, 25(1): 3-30.

Searle, S. R. 1971. Linear models (1st ed.). New York: John Wiley & Sons, Inc.

Sorensen, D. A., Wang, C. S., Jensen, J., and Gianola, D. 1994. Bayesian-analysis of genetic change due to selection using Gibbs sampling. Genetics Selection Evolution, 26(4): 333-360.

Sorensen, D. A., Andersen, S., Gianola, D., and Korsgaard, I. 1995. Bayesian-inference in threshold models using Gibbs sampling. Genetics Selection Evolution, 27(3): 229-249.

Sorensen, D. A. and Gianola, D. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics (1 ed.). New York: Springer-Verlag New York, Inc.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van derLinde, A. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology, 64: 583-616.

Stranden, I. and Gianola, D. 1998. Attenuating effects of preferential treatment with Student-t mixed linear models: a simulation study. Genetics Selection Evolution, 30(6): 565-583.

Stranden, I. and Gianola, D. 1999. Mixed effects linear models with t-distributions for quantitative genetic analysis: a Bayesian approach. Genetics Selection Evolution, 31(1): 25-42.

Sullivan, P. G. 1995. Alternatives for genetic evaluation with uncertain parentage. Canadian Journal of Animal Science, 75(1): 31-36.

Sullivan, P. G., Wilton, J. W., Miller, S. P., and Banks, L. R. 1999. Genetic trends and breed overlap derived from evaluations of beef cattle for multiple-breed genetic growth traits. Journal of Animal Science, 77(8): 2019-2027.

Van-Arendonk, J. A. M., Spelman, R. S., Vander Waaij, E. H., Bijma, P., and Bovenhuis, H. 1998. Livestock breeding schemes: Challenges and opportunities. Paper presented at the 6th World Congress on Genetics Applied to Livestock Production, Armidale, Australia.

Wang, C. S., Gianola, D., Sorensen, D. A., Jensen, J., Christensen, A., and Rutledge, J. J. 1994a. Response to selection for litter size in Danish landrace pigs - A Bayesian-analysis. Theoretical and Applied Genetics, 88(2): 220-230.

Wang, C. S., Rutledge, J. J., and Gianola, D. 1994b. Bayesian-analysis of mixed linear-models via Gibbs sampling with an application to litter size in Iberian pigs. Genetics Selection Evolution, 26(2): 91-115.

Wang, C. S., Quaas, R. L., and Pollak, E. J. 1997. Bayesian analysis of calving ease scores and birth weights. Genetics Selection Evolution, 29(2): 117-143.

Westell, R. A., Quaas, R. L., and Vanvleck, L. D. 1988. Genetic groups in an animal-model. Journal of Dairy Science, 71(5): 1310-1318.

Willham, R. L. 1972. The role of maternal effects in animal breeding: III. Biometrical aspects of maternal effects in animals. Journal of Animal Science, 35(6): 1288-1293.

Winkelman, A. and Schaeffer, L. R. 1988. Effect of heterogeneity of variance on dairy sire evaluation. Journal of Dairy Science, 71(11): 3033-3039.

Wolf, J., Distl, O., Hyanek, J., Grosshans, T., and Seeland, G. 1995. Crossbreeding in farm-animals .5. Analysis of crossbreeding plans with secondary crossbred generations. Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie, 112(2): 81-94.