Genetic evaluation with finite locus models

Liviu Radu Totir
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Genetics Commons

Recommended Citation
Totir, Liviu Radu, "Genetic evaluation with finite locus models " (2002). Retrospective Theses and Dissertations. 486.
https://lib.dr.iastate.edu/rtd/486
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®
Genetic evaluation with finite locus models

by

Liviu Radu Totir

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Animal Breeding and Genetics

Program of Study Committee:
Rohan Fernando, Major Professor
Alicia Carriquiry
Jack Dekkers
Albert Freeman
Hal Stern

Iowa State University
Ames, Iowa
2002

Copyright © Liviu Radu Totir, 2002. All rights reserved.
Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Liviu Radu Totir

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program
In memoria lui Haralambie Totir și a tuturor celor care au fost opriți să viseze.

In memory of Haralambie Totir and of all those who were not allowed to dream.
TABLE OF CONTENTS

LIST OF TABLES .................................................. vii
LIST OF FIGURES ................................................... xi
ACKNOWLEDGMENTS ................................................. xv
ABSTRACT ............................................................ xvi

CHAPTER 1. GENERAL INTRODUCTION .......................... 1
1.1 Objectives and Organization ................................ 7
Bibliography ....................................................... 8

CHAPTER 2. EFFECT OF USING APPROXIMATE GAMETIC VARIANCE COVARIANCE MATRICES ON MARKER ASSISTED SELECTION BY BLUP ................................................. 15
2.1 Introduction ................................................... 16
2.2 Methods ....................................................... 18
  2.2.1 Notation .................................................. 18
  2.2.2 IBD probabilities at the MQTL ......................... 19
  2.2.3 Approximate calculations of PDQ probabilities ....... 21
  2.2.4 Calculation of the inverse of the gametic variance covariance matrix .... 24
  2.2.5 Estimation of the exact genetic variance covariance matrix by MCMC ... 25
  2.2.6 Simulation study ........................................ 25
2.3 Results ....................................................... 27
  2.3.1 Validation of the use of ESIP to estimate the exact genetic variance covariance matrix ............................................. 27
2.3.2 Comparison of response to selection obtained with different MABLUP methods ........................................ 28

2.4 Discussion ........................................................................................................ 38

Bibliography ........................................................................................................... 40

CHAPTER 3. A COMPARISON OF ALTERNATIVE METHODS TO COMPUTE THE CONDITIONAL MEAN OF GENOTYPIC VALUES IN FINITE LOCUS MODELS ................................................. 44

3.1 Introduction ....................................................................................................... 45

3.2 Methods ........................................................................................................... 47

3.2.1 Elston-Stewart algorithm ........................................................................... 50

3.2.2 Iterative peeling ......................................................................................... 50

3.2.3 MCMC methods ......................................................................................... 50

3.2.4 Simulation study .......................................................................................... 55

3.3 Results ............................................................................................................ 58

3.3.1 Iterative peeling ......................................................................................... 58

3.3.2 Influence of the number of loci on computing efficiency ......................... 59

3.3.3 MCMC methods ......................................................................................... 60

3.3.4 Accuracy levels reached by the four methods under investigation .......... 68

3.3.5 Implementation of ESIP ............................................................................ 68

3.4 Discussion ....................................................................................................... 70

Bibliography ........................................................................................................... 72

CHAPTER 4. THE EFFECT OF THE NUMBER OF LOCI IN FINITE LOCUS MODELS FOR GENETIC EVALUATION ................................................................. 78

4.1 Introduction ....................................................................................................... 79

4.2 Methods ........................................................................................................... 80

4.3 Results ............................................................................................................ 89

4.3.1 Purebred case .............................................................................................. 90

4.3.2 Crossbred case ............................................................................................ 95
CHAPTER 5. A STRATEGY TO IMPROVE THE COMPUTATIONAL EFFICIENCY OF MARKER ASSISTED GENETIC EVALUATION UNDER FINITE LOCUS MODELS ........................................ 106

5.1 Introduction .................................................. 106
5.2 Materials and Methods ....................................... 108
5.3 Results and Discussion ...................................... 110
Bibliography .................................................. 111

CHAPTER 6. GENERAL DISCUSSION .............................. 113

APPENDIX. SUPPLEMENTARY RESULTS FOR CHAPTER 3 ........................................... 117
LIST OF TABLES

Table 2.1  Given the maternal marker haplotype inherited by \( i \), probability that MQTL allele \( Q_i^m \) descends from the parental allele \( Q_p^k \) (PDQ) 22

Table 2.2  Given the maternal marker allele inherited by \( i \), probability that MQTL allele \( Q_i^m \) descends from the parental allele \( Q_p^k \) (PDQ) 23

Table 2.3  Given the parental marker information, probability that marker allele \( M_i^1 \) descends from the parental allele \( M_p^k \) (PDM) 23

Table 2.4  Simulation details 26

Table 2.5  Maximum, mean and standard deviation of the elements of the matrices of absolute differences between the exact and the estimated genetic variance covariance matrices, for complete marker data with 4 alleles at the marker locus. 27

Table 2.6  Maximum, mean and standard deviation of the elements of the matrices of absolute differences between the exact and the estimated genetic variance covariance matrices, for incomplete marker data with 4 alleles at the marker locus. 28

Table 3.1  Situations simulated. 57

Table 3.2  Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using iterative peeling compared to estimates calculated using the Elston-Stewart algorithm. 59
Table 3.3 Computing time in seconds on a Dec Alphastation 500 for 1000 samples obtained with each of the three MCMC samplers for 1 2 and 3 locus models for situation 1.

Table 3.4 Number of samples obtained by blocking and scalar Gibbs in the same amount of time required by ESIP to generate 1 sample for the eight situations considered.

Table 3.5 Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using ESIP compared to estimates calculated using the Elston-Stewart algorithm.

Table 3.6 Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using blocking Gibbs compared to estimates calculated using the Elston-Stewart algorithm.

Table 3.7 Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using scalar Gibbs compared to estimates calculated using the Elston-Stewart algorithm.

Table 3.8 Differences between ESIP estimates, calculated based on 50000 samples, for different sizes of the maximum cutset size.

Table 4.1 Situations simulated for the purebred case for three different pedigrees.

Table 4.2 Parameters for the FLMS used to analyze purebred data.

Table 4.3 Situations simulated for the two-breed case for two pedigrees.

Table 4.4 Summary statistics of 200 values from 100 replicates of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(3) for purebred situation 7, calculated using the Elston-Stewart algorithm and ESIP.
Table 4.5 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2) and FLM(3), for purebred situation 1. ........................................ 90
Table 4.6 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 2. ......................... 91
Table 4.7 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2) and FLM(3), for purebred situation 3. ........................................ 91
Table 4.8 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 4. ......................... 92
Table 4.9 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), for purebred situation 5. ........................................ 92
Table 4.10 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 6. ......................... 93
Table 4.11 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), for purebred situation 7. ........................................ 93
Table 4.12 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 8. ......................... 94
Table 4.13 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3) and FLM(4), for purebred situation 9. ........................................ 94
Table 4.14  The value of $N_1$ and the corresponding mean scaled absolute difference between BLP and BP evaluations under $FLM(N_1)$ for the nine situations considered in the purebred case. .......................... 95

Table 4.15  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under $FLM(2,1)$, for two-breed situation 1. ................................. 96

Table 4.16  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under $FLM(2,1)$, $FLM(3,1)$ and $FLM(4,1)$, for two-breed situation 2. ................................. 97

Table 4.17  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under $FLM(2,1)$ for two-breed situation 3. ................................. 97

Table 4.18  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under $FLM(2,1)$, $FLM(3,1)$ and $FLM(4,1)$, for two-breed situation 4. ................................. 98

Table 5.1  Performance of various ESIP(1:k) ............................... 110
LIST OF FIGURES

Figure 2.1 Pedigree used. ........................................ 26
Figure 2.2 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 2 alleles at the flanking markers for complete marker data. ........................................ 29
Figure 2.3 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 4 alleles at the flanking markers for complete marker data. ........................................ 30
Figure 2.4 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 12 alleles at the flanking markers for complete marker data. ........................................ 31
Figure 2.5 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 2 alleles at the flanking markers for incomplete marker data. ........................................ 32
Figure 2.6 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 4 alleles at the flanking markers for incomplete marker data. ........................................ 33
Figure 2.7 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 12 alleles at the flanking markers for incomplete marker data. ........................................ 34
Figure 2.8 Summary of percent superiority of MABLUP estimated from 5,000 replicates of the simulation, by methods A, B, and E, and by the number of alleles at a marker for complete marker data. ........................................ 36
Figure 2.9 Summary of percent superiority of MABLUP estimated from 5,000 replicates of the simulation, by methods A, B, and E, and by the number of alleles at a marker for incomplete marker data. ......... 37

Figure 3.1 Base Pedigree. ................................................. 56

Figure 3.2 Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5,000 samples for ESIP, 65,000 samples for blocking Gibbs, or 130,000 samples for scalar Gibbs, for situation 6. ................................................. 62

Figure 3.3 Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5,000 samples for ESIP, 35,000 samples for blocking Gibbs, or 70,000 samples for scalar Gibbs, for situation 4. ................................................. 63

Figure 3.4 Mean of the scaled absolute differences between estimates obtained using each of the four methods under investigation and estimates obtained using the Elston-Stewart algorithm for each of the eight situations. ... 69

Figure 4.1 Simple purebred pedigree. ................................................. 85

Figure 4.2 Simple crossbred pedigree. ................................................. 88

Figure 4.3 Box plots of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6) for situation 2. ................................................. 96

Figure 4.4 Extended Pedigree ................................................. 100

Figure 4.5 Pedigree with loops ................................................. 101

Figure 5.1 Behavior of the mean sum of squares of the difference for ESIP(1:1) ... 109
Figure 6.1  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5 000 samples for ESIP, 35 000 samples for blocking Gibbs, or 70 000 samples for scalar Gibbs, for situation 1. ......................................................... 117

Figure 6.2  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5 000 samples for ESIP, 35 000 samples for blocking Gibbs, or 70 000 samples for scalar Gibbs, for situation 2. ......................................................... 118

Figure 6.3  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5 000 samples for ESIP, 35 000 samples for blocking Gibbs, or 70 000 samples for scalar Gibbs, for situation 3. ......................................................... 119

Figure 6.4  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5 000 samples for ESIP, 35 000 samples for blocking Gibbs, or 70 000 samples for scalar Gibbs, for situation 4. ......................................................... 120

Figure 6.5  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5 000 samples for ESIP, 35 000 samples for blocking Gibbs, or 70 000 samples for scalar Gibbs, for situation 5. ......................................................... 121
Figure 6.6  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5,000 samples for ESIP, 65,000 samples for blocking Gibbs, or 130,000 samples for scalar Gibbs, for situation 6. 

Figure 6.7  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5,000 samples for ESIP, 35,000 samples for blocking Gibbs, or 70,000 samples for scalar Gibbs, for situation 7. 

Figure 6.8  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5,000 samples for ESIP, 35,000 samples for blocking Gibbs, or 70,000 samples for scalar Gibbs, for situation 8.
ACKNOWLEDGMENTS

I am very grateful to my major professor Dr. Rohan Fernando for his patience, friendship, and constant support throughout my years of study. Without his support and guidance I would have never been able to write this dissertation. No words can truly describe my consideration for him so I will just simply say thank you for everything. My sincere thanks go to Dr. Jack Dekkers as well. He has made no difference between his own students and myself, making our interactions very enjoyable and extremely useful. My thanks go also to the other members of my committee: Dr. Alicia Carriquiry, Dr. Albert Freeman, and Dr. Hal Stern.

Throughout my years of study, I have been fortunate enough to interact with many graduate students, post-doctoral students, faculty, and visiting scientists. I thank all of them for their kindness and support. However, special thanks are due to Bruce Southey, Nerilson Terra Santos, Soledad Fernandez, and Bernt Guldbrandtsen.

I would like to thank my family, especially my parents, for their constant support and unconditional love. They have always supported me in my decisions regardless of the pain that these decisions might have caused them. My love and prayers go to them.

Throughout my years of study, my wife Claudia has been by my side in good times and in difficult times. Her love and care have made my work and life much easier and enjoyable. Her courage, love, and dedication have always inspired me. To her I say: I love you.

Finally, I would like to acknowledge the financial support received from Iowa State University, the Monsanto Company, and USDA. Without their financial support, this work would not have been possible.
ABSTRACT

The availability of genotypic data in recent years has resulted in increased interest in the use of marker assisted genetic evaluation (MAGE) in livestock species. Under additive inheritance, Henderson's mixed model equations (HMME) provide an efficient approach to obtain genetic evaluations by marker assisted best linear unbiased prediction (MABLUP) given pedigree relationships, trait, and marker data. For large pedigrees with many missing markers, however, it is not feasible to calculate the exact gametic variance covariance matrix required to construct HMME, and thus, approximations are used. By computer simulation we observed that the use of exact matrices would increase response to selection by 2.2% up to 11.7%. Marker assisted selection (MAS) is efficient especially for traits that have low heritability and non-additive gene action. BLUP methodology under non-additive gene action is not feasible for large inbred or crossbred pedigrees. It is easy to incorporate non-additive gene action in a finite locus model. Under such a model, the unobservable genotypic values can be predicted using the conditional mean of the genotypic values given the available data, which is also known as the best predictor (BP). The potential of alternative methods to compute BP under finite locus models was studied, and it was shown that Markov chain Monte Carlo (MCMC) methods that sample blocks of genotypes jointly hold most promise for such computations. The efficiency of MCMC methods for genetic evaluation by BP under finite locus models, depends on the number of loci considered in the model. Thus, the effect of the number of loci used in the finite locus model used for genetic evaluation by BP was studied by computer simulation. In our study, models with two to six loci yielded accurate BP evaluations for traits determined by 100 loci. Finally, we proposed a strategy to improve the computational efficiency of MAGE under finite locus models.
CHAPTER 1. GENERAL INTRODUCTION

The identification of the “best” animals for breeding is an essential step in any breeding program. Ideally this identification would be based on the breeding values of the candidate animals. Genotypic values, however, cannot be observed, and consequently they need to be predicted from the available information. Three sources of information can be used to predict the genotypic value of an animal: phenotypic, genotypic, and pedigree information.

In a classical paper, Henderson [32] discussed the properties and limitations of different predictors of genotypic values. The conditional mean of the genotypic value given the available information is the best predictor (BP) in the sense that it minimizes the mean squared error of prediction. When the genotypic and phenotypic values of candidates are identically and independently distributed, truncation selection using the BP maximizes the mean merit of the selected candidates [4]. It has been subsequently shown that, regardless of the joint distribution of the genotypic and phenotypic values, choosing k of n candidates using the BP maximizes the mean merit of the selected candidates [2, 11]. In order to calculate the conditional mean, however, the joint distribution of the genotypic and phenotypic values must be specified completely. Further, the BP might be non-linear and difficult to compute.

A common approach to simplify prediction is to consider only predictors that are linear in the data. Among linear predictors, the predictor that minimizes the mean square error of prediction is known as the best linear predictor (BLP). It can be shown that when the joint distribution of the genotypic and phenotypic values is multivariate normal, BLP is BP. According to Henderson [32], J. L. Lush was one of the first geneticists to use BLP methods for genetic evaluation [40]. After the introduction of conventional selection index theory in plant [9], and animal [27] genetics, BLP became the method of choice for genetic evaluation.
To compute BLP, only the first and second moments of the joint distribution of the genotypic and phenotypic values must be known.

The requirement to know the first moment of the joint distribution of the genotypic and phenotypic values was recognized by Henderson as a limitation to BLP. In order to remove this limitation, Henderson introduced best linear unbiased prediction (BLUP). Consider the following mixed linear model

$$y = X\beta + Za + e,$$

where $y$ is the vector of observations; $\beta$ is a vector of fixed effects; $X$ is an incidence matrix relating $\beta$ to $y$; $a \sim \mathcal{N}(0, G_a)$ is a vector of animal effects with variance-covariance matrix $G_a$; $Z$ is an incidence matrix relating $a$ to $y$; $e \sim \mathcal{N}(0, R)$ is the residual vector with variance-covariance matrix $R$. Henderson [29] maximized the joint density of $a$ and $y$ with respect to $\beta$ and $a$, and consequently obtained the following system of linear equations

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G_a^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix},$$

which are known as Henderson's mixed model equations (HMME). Subsequently it was shown that, regardless of the distribution of $a$ and $e$, the solution to HMME gives the best linear unbiased estimator (BLUE) of $\beta$ [30], and the best linear unbiased predictor (BLUP) of $a$ [31]. One of the requirements to obtain BLUP using HMME is to obtain the inverse of $G_a$.

When genetic evaluation is done under the assumption of additive inheritance and no genotypic information is included, the variance-covariance matrix of the vector of animal effects ($a$) can be modeled as

$$G_a = A\sigma_a^2,$$

where element $a_{ij}$ of the additive relationship matrix $A$ is the additive relationship coefficient between animals $i$ and $j$, and $\sigma_a^2$ is the additive variance. For $i$, not a descendant of $j$, we can
write

\[ a_{ij} = \frac{1}{2}(a_{isj} + a_{idj}) \quad (1.3) \]

\[ a_{ii} = 1 + \frac{a_{isj,iidj}}{2}, \]

where \( a_{isj} \) is the additive relationship coefficient between \( i \) and the sire of \( j \) (\( s_j \)) and \( a_{idj} \) is the additive relationship coefficient between \( i \) and the dam of \( j \) (\( d_j \)). Due to this simple recursive formula for \( a_{ij} \), the inverse of \( G_a \) can be computed very efficiently even with extremely large data sets [33, 47, 48]. Currently BLUP, obtained from HMME, is the most widely used method for genetic evaluation.

The increased availability of polymorphic genetic markers in the last ten years has allowed researchers to start mapping quantitative trait loci (QTL) [26]. Results from these mapping studies provide two sources of genotypic information: genotypes at QTL, and genotypes at marker loci linked to QTL. When genotypes at QTL are available, the corresponding effects can be incorporated as fixed effects in models for genetic evaluation and selection [16]. When genotypes at marker loci linked to QTL are available, these marker genotypes can be used to model the genotypic mean and the genotypic variance covariance matrix at the linked QTL, which are referred to as the marked QTL (MQTL). When the MQTL and the marker loci are in gametic phase (linkage) disequilibrium, marker information affects both the mean and the variance covariance matrix at the MQTL [60]. In contrast, when the MQTL and the marker loci are in gametic phase (linkage) equilibrium, marker information affects only the variance covariance matrix at the MQTL [12].

When marker information is available, it is convenient to write the vector \( \alpha \) of additive effects of \( n \) animals as

\[ \alpha = m + u, \quad (1.4) \]

where \( m \) is the vector of additive effects at the MQTL, and \( u \sim N(0, G_u) \) is the vector of additive effects of the remaining QTL (RQTL) of the \( n \) individuals. Assuming linkage equilibrium between the MQTL and the RQTL, the additive variance covariance matrix \( (G_{am}) \)
of the vector of animal effects \((a)\) can be written as

\[
G_{am} = G_m + G_u, \tag{1.5}
\]

where \(G_m\) is the variance-covariance matrix at the MQTL, and \(G_u\) is the variance-covariance matrix of the RQTL. It can be shown that, in general \(G_{am} \neq G_a\) [12]. BLUP of \(a\) can be obtained by solving the HMME given in (1.2) with \(G_a\) replaced with \(G_{am}\). Thus, the inverse of \(G_{am}\) must be computed. Chevalet et al. [3] have provided a general method to compute \(G_m\) given phenotypic data at a linked marker. \(G_u\) can be computed using the same recursive formulae given for \(G_a\). The inverse of \(G_{am}\), however, has a dense structure and consequently \(G_{am}\) cannot be inverted for large pedigrees. Thus, this strategy is not feasible for genetic evaluation by BLUP in large pedigrees with marker and trait data.

To accommodate large pedigrees, Fernando and Grossman [12] developed an alternative strategy to incorporate MQTL in BLUP methodology. They wrote the vector \(a\) of additive effects of \(n\) animals as

\[
a = K\nu + u, \tag{1.6}
\]

where \(\nu\) is a vector of \(2n\) gametic effects, each of the \(n\) individuals having a paternal MQTL effect \((\nu^p_i)\), and a maternal MQTL effect \((\nu^m_i)\); \(K_{n \times 2n}\) is an incidence matrix relating \(\nu\) to \(a\), the \(i\)th row of \(K\) having ones for the elements corresponding to the gametic effects at the MQTL of individual \(i\), and zeroes for the remaining elements. Thus, the mixed linear model becomes

\[
y = X\beta + ZK\nu + Zu + e, \tag{1.7}
\]

and, if we define \(W = ZK\), HMME can be written as

\[
\begin{bmatrix}
X'R^{-1}X & X'R^{-1}W & X'R^{-1}Z \\
W'R^{-1}X & W'R^{-1}W + G_u^{-1} & W'R^{-1}Z \\
Z'R^{-1}X & Z'R^{-1}W & Z'R^{-1}Z + G_u^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
\hat{\nu} \\
\hat{u}
\end{bmatrix}
= \begin{bmatrix}
X'R^{-1}y \\
W'R^{-1}y \\
Z'R^{-1}y
\end{bmatrix} \tag{1.8}
\]

Now, to obtain BLUP from the HMME given in (1.8), the inverses of \(G_\nu\) and \(G_u\) need to be
calculated. Both $G_v^{-1}$ and $G_u^{-1}$, however, have a sparse structure. Thus, $G_v$ and $G_u$ can be inverted efficiently for large pedigrees. Efficient algorithms to invert $G_v$ have been developed, both for the case when the origin of the marker genotypes can be inferred [12], and for the case when the origin of the marker genotypes is not known [35, 54, 57, 58]. $G_u$ can be inverted using the same algorithms developed for $G_u$ [33, 47, 48]. The Fernando and Grossman model was extended to multiple markers by Goddard [22].

The algorithms used to invert $G_v$ give exact results only when marker information is complete: when flanking markers are used, even the linkage phase between the markers is assumed known [22, 41]. In large pedigrees incomplete marker information is the rule rather than the exception. Theory developed by Wang et al. [58] explains how to compute the exact gametic variance covariance matrix for incomplete marker data. This theory, however, is computationally intensive, and as a result it is not feasible for large pedigrees. Wang et al. [58] suggest alternative methods that result in approximations of the exact variance-covariance matrix. The effect of these approximations on marker assisted BLUP, however, has not been investigated. Methods have also been developed to eliminate the MQTL equations for animals with missing marker genotypes [35, 42]. These methods also yield approximate results.

Marker assisted selection is efficient especially for traits that have low heritability. Several such traits, however, are also known to exhibit non-additive inheritance. Under non-additive inheritance, obtaining the inverse of the genetic variance-covariance matrix becomes computationally challenging [5]. Algorithms to invert the genetic variance-covariance matrix under dominance inheritance have been investigated [34, 49]. These algorithms, however, are not feasible for large inbred data sets [5]. Use of crossbred data further increases the complexity of the problem [17, 37, 38, 39]. Therefore, under non-additive inheritance BLUP methodology is difficult to implement [15].

It is easy to incorporate non-additive inheritance in finite locus models [15]. Furthermore, crossbred data do not add to the complexity of the model. In recent years, there has been an increased interest in the use of finite locus models for genetic evaluation [15, 23, 51] and for parameter estimation [6, 7, 45, 46]. Under a finite locus model, the conditional mean of the
genotypic values given the data can be easily calculated if the conditional probabilities of the possible genotypes can be calculated [15, 23, 51].

For simple pedigrees, these conditional probabilities can be calculated exactly [14], by using the Elston Stewart algorithm [8], which was developed in human genetics to compute the likelihood of alternative models for quantitative traits under monogenic and oligogenic inheritance.

For complex pedigrees, typically encountered in animal breeding, it has been proposed to use iterative peeling to approximate the conditional genotype probabilities [10, 53]. It has been shown that iterative peeling yields exact genotype probabilities for pedigrees without loops [13, 59]. For pedigrees with loops, iterative peeling has been shown to yield very accurate genotype probabilities for a biallelic disease locus [10].

Markov Chain Monte Carlo (MCMC) methods were proposed during the last decade to overcome computational problems encountered in analysis of complex pedigrees [24, 25, 36, 52]. These methods can be used to estimate conditional genotype probabilities to any degree of accuracy. The single site Gibbs sampler [18, 20] is one of the simplest MCMC methods for sampling random variables from a high dimensional space, and this sampler is still widely used in analysis of pedigree data [1, 7, 43, 44]. In this sampler, each variable is sampled from its conditional distribution given all the other variables. Although it is very easy to implement, the single site Gibbs sampler has been shown to have slow mixing in livestock pedigrees where large sibships are typical [36]. As a solution to this problem, it has been suggested to sample genotypes in blocks [36]. An alternative strategy is to sample genotypes jointly from the entire pedigree [10, 28]. For simple pedigrees, such samples can be obtained by using the Elston-Stewart algorithm and reverse peeling [10, 28]. For complex pedigrees, with many nested loops, this strategy is not feasible, and thus, approximate methods are used to generate candidate samples, which are then accepted or rejected by the Metropolis-Hastings algorithm [10, 28]. The Metropolis-Hastings algorithm provides an easy way to construct a Markov chain with a desired stationary distribution \( \pi \). The chain moves through the sample space based on the following predefined rules. To move from an arbitrary state \( X_t = X \) to the next state \( X_{t+1} \), a
proposal distribution \( q(Y \mid X) \) is used to generate a candidate point \( Y \). The proposed move is accepted with probability

\[
\alpha(X,Y) = \min \left( 1, \frac{\pi(Y)q(X \mid Y)}{\pi(X)q(Y \mid X)} \right).
\]  

(1.9)

If the move is accepted \( X_{t+1} \) becomes \( Y \). If the move is rejected \( X_{t+1} \) becomes \( X \). The Gibbs sampler, also known as the alternating conditional sampler \([19]\), is a special case of the Metropolis-Hastings algorithm. MCMC methods hold most promise for the efficient computation of the conditional probabilities needed for genetic evaluation by BP in finite locus models.

### 1.1 Objectives and Organization

As discussed above, BLUP can be used for marker assisted genetic evaluation. In large livestock pedigrees, however, when marker genotype information is not complete, approximations must be used. In chapter 2, the effect of using approximate gametic variance covariance matrices on response to selection by marker assisted BLUP is investigated. BLUP is not efficient under non-additive inheritance and thus, genetic evaluation by BP in finite locus models has been proposed as an alternative. In chapter 3, the potential of alternative methods to compute the conditional mean of the genotypic values in finite locus models is investigated. The efficiency of these methods for genetic evaluation by BP in finite locus models depends on the number of loci considered in the model. Thus, in chapter 4, the effect of the number of loci in finite locus models for genetic evaluation by BP is investigated. The application, in real livestock pedigrees, of marker assisted genetic evaluation by BP under finite locus models, depends on the computational efficiency of the method used to calculate the BP evaluations. In chapter 5, a strategy to improve the computational efficiency of marker assisted genetic evaluation under finite locus models is investigated. Each of the chapters 2 through 5 is structured as a journal article. Thus, each chapter contains its own list of literary citations. Chapter 6 summarizes the general conclusions of this dissertation.
Bibliography


CHAPTER 2. EFFECT OF USING APPROXIMATE GAMETIC VARIANCE COVARIANCE MATRICES ON MARKER ASSISTED SELECTION BY BLUP

Abstract

Under additive inheritance, Henderson's mixed model equations (HMME) provide an efficient way to obtain genetic evaluations by marker assisted best linear unbiased prediction (MABLUP) given pedigree relationships, trait, and marker data. For large pedigrees with many missing markers, however, it is not feasible to calculate the exact gametic variance covariance matrix required to construct HMME. In this paper we investigate two methods to approximate the gametic variance covariance matrix. The first method (Method A) completely discards the marker information when the linkage phase between two flanking markers is not known. The second method (Method B) makes use of the marker information at the most polymorphic marker locus when linkage phase between markers is not known. Data sets were simulated with complete or incomplete marker data for flanking markers with 2, 4, or 12 alleles. Response to selection by MABLUP using Method A or Method B were compared with that obtained by MABLUP using the exact genetic variance covariance matrix. The exact variance covariance matrix was estimated by using 15,000 independently distributed vectors of genotypic values obtained using a Markov Chain Monte Carlo sampler. For the simulated conditions, the superiority of MABLUP over BLUP based only on pedigree relationships and trait data varied between 0.5% and 10.6% for Method A, between 2.1% and 16.2% for Method B, and between 8% and 18.5% for the exact method.
2.1 Introduction

Over the last decade, as a result of extensive efforts to map quantitative trait loci (QTL) [12], genotypic data have become available for genetic evaluation. Two types of genotypic data can be used for genetic evaluation: genotypes at QTL, and genotypes at molecular markers linked to QTL. The effects of known QTL genotypes can be included as fixed effects in the mixed models used for genetic evaluation by best linear unbiased prediction (BLUP) [9]. Few causative QTL, however, have been identified so far [12]. In contrast, a large number of markers linked to QTL, which are referred to as the marked QTL (MQTL), have been identified. Genotypes at markers linked to MQTL can be used to model the genotypic mean and the genetic variance covariance matrix at the MQTL [8, 34]. Thus, the effects of the marker genotypes can be included as a fixed effects and the gametic effects of the MQTL as random effects in the mixed linear models used for genetic evaluation by BLUP [34]. Marker genotypes, however, affect the genotypic mean only if the markers and the MQTL are in gametic phase (linkage) disequilibrium [34].

For large pedigrees, Henderson's mixed model equations (HMME) [15] provide an efficient way to obtain BLUP. One of the requirements to obtain BLUP from HMME is to compute the inverses of the variance covariance matrices of the random effects in the model. When only pedigree and trait information are used for genetic evaluation, the inverse of the conditional variance covariance matrix of the vector of unobservable genotypic values given pedigree relationships needs to be computed. Under additive inheritance, efficient algorithms are available to invert this conditional variance covariance matrix [14, 25, 26]. When marker phenotypes are available, Chevalet et al. [3] provided a general method to compute the conditional variance covariance matrix of the vector of unobservable genotypic values given pedigree and marker information. The resulting genetic variance covariance matrix, however, has a dense structure, and thus, cannot be inverted efficiently for large pedigrees [33].

For large pedigrees, when marker genotype information is available, it is convenient to include in the mixed model, as a random effect, the gametic effect of the MQTL, in addition to the random additive genetic effect of the remaining QTL (RQTL) [8]. To construct the HMME
for such a mixed model, the inverse must be computed for the conditional variance covariance matrix of the vector of gametic effects at the MQTL given marker and pedigree information, which is referred to as the gametic variance covariance matrix at the MQTL. Furthermore, the inverse of the genetic variance covariance matrix at the RQTL must be computed. These inverse matrices are sparse. Efficient algorithms are available to invert the genetic variance covariance matrix at the RQTL [14, 25, 26]. Also, algorithms to invert the gametic variance covariance matrix at the MQTL have been developed for both the case when the marker origin is known [8] and the case when the marker origin is not known [16, 31, 32, 33].

The algorithms used to invert the gametic variance covariance matrix at the MQTL, yield exact results only if the marker genotypes and the linkage phase between markers are known, i.e. when the marker information is complete [17, 33]. In large pedigrees, however, incomplete marker information is the rule rather than the exception. Wang et al. [33] provided a formula to compute the exact gametic variance covariance matrix for incomplete marker data. Use of this formula, however, is computationally intensive and thus, not feasible for large pedigrees. For large pedigrees, when marker information is incomplete, approximations must be used.

When the linkage phase between two flanking markers is not known, a common practice is to completely ignore the marker information at these two markers [11, 21]. An alternative to this strategy is, when phase is not known, to use the information at one of the two flanking markers [23, 30]. Both of these strategies result in approximate gametic variance covariance matrices, but are computationally feasible for large pedigrees.

The objective of this study is to examine the consequences of using approximate gametic variance covariance matrices on response to selection by MABLUP. For small simulated data sets, response to selection by MABLUP computed using approximate gametic variance covariance matrices was compared to response to selection by MABLUP computed using exact genetic variance covariance matrices estimated by Markov chain Monte Carlo (MCMC).
2.2 Methods

2.2.1 Notation

Consider an MQTL (Q) closely linked to two polymorphic flanking markers (M and N). M and N are assumed to be in linkage equilibrium with Q and with each other. The following diagram shows the chromosomal segments containing Q, M, and N, for individual i with parents d and s, and for another individual j.

\[
\begin{array}{ccc}
M_d^m & Q_d^m & N_d^m \\
M_d^f & Q_d^f & N_d^f \\
M_s^m & Q_s^m & N_s^m \\
M_s^f & Q_s^f & N_s^f \\
M_i^m & Q_i^m & N_i^m \\
M_i^f & Q_i^f & N_i^f \\
M_j^m & Q_j^m & N_j^m \\
M_j^f & Q_j^f & N_j^f \\
\end{array}
\]

The paternal allele at a given locus is denoted by a superscript f, and the maternal allele by a superscript m. The genotypes at markers M and N may be observed, and thus, may be used for marker assisted genetic evaluation (MAGE). The genotypes at the MQTL (Q), however, cannot be observed. As discussed later, even if the marker genotypes are known, it is not always possible to infer the linkage phase between them.

The conditional covariance of the additive effects \( v_i^{k_i} \) and \( v_j^{k_j} \) of MQTL alleles \( Q_i^{k_i} \) and \( Q_j^{k_j} \) in individuals i and j, given the observable marker information \( (G_{obs}) \), is written as

\[
\text{Cov}(v_i^{k_i}, v_j^{k_j} \mid G_{obs}) = \Pr(Q_i^{k_i} = Q_j^{k_j} \mid G_{obs}) \sigma_v^2,
\]

where \( k_i \) and \( k_j \) are m or f if the MQTL allele origin is known [8], and 1 or 2 if the MQTL allele origin is not known [33]; \( \Pr(Q_i^{k_i} = Q_j^{k_j} \mid G_{obs}) \) is the conditional probability that \( Q_i^{k_i} \) is
identical by descent (IBD) to $Q_j^{kj}$ given $G_{obs}$; $\sigma_u^2$ is half of the variance of the additive effect of the MQTL.

### 2.2.2 IBD probabilities at the MQTL

Recursive formulae have been widely used to compute IBD probabilities [2, 4, 6, 10, 20, 24, 27, 28, 29]. These formulae are based on the principle that a priori the allele transmitted from a parent to an offspring is equally likely to be the parent’s maternal or paternal allele. Thus, the unconditional probability that $Q_i^m$, for example, is IBD to $Q_j^{kj}$ can be written as

$$\Pr(Q_i^m = Q_j^{kj}) = \frac{1}{2} \Pr(Q_d^m = Q_j^{kj}) + \frac{1}{2} \Pr(Q_d^f = Q_j^{kj}) \quad (2.2)$$

When genotype information is available at a single marker, but the origin of the marker genotypes is not known, following Wang et al. [33], the conditional probability that $Q_i^{ki}$ is IBD to $Q_j^{kj}$ given $G_{obs}$ for $i \neq j$, can be written as

$$\Pr(Q_i^{ki} = Q_j^{kj} | G_{obs}) = \Pr(Q_i^{ki} \leftarrow Q_{d}^{1}, Q_{d}^{1} = Q_j^{kj} | G_{obs}) + \Pr(Q_i^{ki} \leftarrow Q_{d}^{2}, Q_{d}^{2} = Q_j^{kj} | G_{obs}) + \Pr(Q_i^{ki} \leftarrow Q_{s}^{1}, Q_{s}^{1} = Q_j^{kj} | G_{obs}) + \Pr(Q_i^{ki} \leftarrow Q_{s}^{2}, Q_{s}^{2} = Q_j^{kj} | G_{obs}), \quad (2.3)$$

where for example, $\Pr(Q_i^{ki} \leftarrow Q_{d}^{1}, Q_{d}^{1} = Q_j^{kj})$ denotes the probability of the event that $Q_i^{ki}$ descended from $Q^1_d$ and $Q^1_d$ is IBD to $Q_j^{kj}$. Note that if the parental origin of marker genotypes is known, equation (2.3) becomes

$$\Pr(Q_i^{ki} = Q_j^{kj} | G_{obs}) = \Pr(Q_i^{ki} \leftarrow Q_{d}^{m}, Q_{d}^{m} = Q_j^{kj} | G_{obs}) + \Pr(Q_i^{ki} \leftarrow Q_{d}^{f}, Q_{d}^{f} = Q_j^{kj} | G_{obs}) + \Pr(Q_i^{ki} \leftarrow Q_{s}^{m}, Q_{s}^{m} = Q_j^{kj} | G_{obs}) + \Pr(Q_i^{ki} \leftarrow Q_{s}^{f}, Q_{s}^{f} = Q_j^{kj} | G_{obs}), \quad (2.4)$$

where if $k_i = f$ the first two terms of the equation (2.4) are 0, and if $k_i = m$ the last two terms of the equation (2.4) are 0.
If the marker genotypes of \(d\) and \(s\) are known and \(j\) is not a direct descendent of \(i\), the descent of allele \(Q^{k_i}_i\) from one of the alleles of \(d\) or \(s\), is independent of the event that alleles in \(j\) are identical by descent to alleles in \(d\) or \(s\). As a result, equation (2.3) becomes

\[
Pr(Q^{k_i}_i \equiv Q^{k_j}_j | G_{obs}) =
\]

\[
Pr(Q^{k_i}_i \leftarrow Q^1_d | G_{obs}) Pr(Q^1_j \equiv Q^{k_j}_j | G_{obs}) + Pr(Q^{k_i}_i \leftarrow Q^2_d | G_{obs}) Pr(Q^2_j \equiv Q^{k_j}_j | G_{obs}) +
\]

\[
Pr(Q^{k_i}_i \leftarrow Q^1_s | G_{obs}) Pr(Q^1_j \equiv Q^{k_j}_j | G_{obs}) + Pr(Q^{k_i}_i \leftarrow Q^2_s | G_{obs}) Pr(Q^2_j \equiv Q^{k_j}_j | G_{obs}),
\]

where for example, \(Pr(Q^{k_i}_i \leftarrow Q^1_d | G_{obs})\) denotes the probability of descent (PDQ) of \(Q^{k_i}_i\) from \(Q^1_d\). Note that if the parental origin at the marker genotypes is known, \(Q^1_d\), \(Q^2_d\), \(Q^1_s\), and \(Q^2_s\) in equation (2.5) become \(Q^m_d\), \(Q^l_d\), \(Q^m_s\) and \(Q^l_s\). When marker information for the parents is missing, the independence required to obtain equation (2.5) from equation (2.3) may not hold true. Thus, equation (2.5) may yield only approximate results when marker information is missing. When the parental origin at the marker genotype is not known, equation (2.5) cannot be used directly to compute IBD probabilities within an individual (\(i = j\)) [33]. For this situation, IBD probabilities can be computed using formula (11) in Wang et al [33], which makes use of equation (2.5).

When genotype information is available at markers flanking the MQTL, the conditional probability that \(Q^{k_i}_i\) is IBD to \(Q^{k_j}_j\) given \(G_{obs}\) for \(i \neq j\), can be obtained from (2.5), by computing PDQs conditional on this flanking marker information [11]. In this situation, however, when the linkage phase between the two flanking markers is not known, the independence required to obtain equation (2.5) from equation (2.3) may not hold true. Thus, even when marker genotypes are not missing, equation (2.5) may yield only approximate results when the linkage phase between flanking markers is not known.

For a single marker, Wang et al. [33] derived formulae for computing PDQs in terms of recombination rates and probabilities of descent for a marker allele (PDM), e.g. \(Pr(M^{k_i}_i \leftarrow M^1_d | G_{obs})\). When some marker genotypes are missing, however, computing the required PDMs may be computationally intensive. For example, when marker information is missing for the
parents \(d\) and \(s\) of \(i\), the PDM \(\Pr(M^1_i \leftarrow M^1_d \mid G_{\text{obs}})\) can be written as

\[
\Pr(M^1_i \leftarrow M^1_d \mid G_{\text{obs}}) = \sum_{G_d} \sum_{G_s} \sum_{G_i} \Pr(M^1_i \leftarrow M^1_d \mid G_d, G_s, G_i) \Pr(G_d, G_s, G_i \mid G_{\text{obs}}).
\] (2.6)

In equation (2.6), the calculation of \(\Pr(G_d, G_s, G_i \mid G_{\text{obs}})\) can be computationally demanding for a pedigree with a large number of missing marker genotypes. Thus, to make computations feasible for large pedigrees with many missing marker genotypes, \(\Pr(G_d, G_s, G_i \mid G_{\text{obs}})\) must be also approximated. Note that when flanking markers are used, PDMs are replaced by probabilities of descent of a haplotype [11]. Again, when the linkage phase between the flanking markers is not known, these probabilities must be approximated.

If the gametic variance covariance matrix is constructed using the recursive formula (2.5), then its inverse can also be obtained using a simple recursive formula [31, 33]. But, for large pedigrees with many missing markers, this requires the efficient computation of approximate PDQs. In the next section we discuss two strategies to compute approximate PDQs for large pedigrees given genotypes at two flanking markers.

### 2.2.3 Approximate calculations of PDQ probabilities

The genotype at a marker locus may be unobserved (missing) or observed. Based on the observable marker data for the entire pedigree, some of the unobserved marker genotypes can be inferred with certainty. In this paper, the genotype elimination algorithm by Lange and Goradia [19], was applied to the entire pedigree. This algorithm yields a list of possible genotypes for each of the unobserved genotypes. Whenever such a list contains only one possible genotype, the unobserved genotype is inferred with certainty and is treated as an observed genotype. An observed genotype is ordered if the parental origin of the alleles is known, or unordered if the parental origin is unknown.

One simple method to compute PDQs is to use marker information only when the genotypes are ordered at both flanking markers, i.e. when the linkage phase between the markers is known [11]. In this case, PDQs can be computed as described by Goddard [11]. For example, if we assume no double recombination between the flanking markers, the PDQ for MQTL allele \(Q^m_i\).
conditional on the maternal marker haplotype inherited by \( i \), can be calculated as shown in Table (2.1). The PDQ for MQTL allele \( Q^m_i \), conditional on the paternal marker haplotype inherited by \( i \), can be calculated in a similar manner. In this method, when the phase is not known marker information is ignored completely, and thus the PDQ for each of the parental alleles is equal to 0.5. This method will be referred to as Method A.

An alternative method that makes better use of the marker information is described below. This alternative method will be referred to as Method B. As in Method A, when the linkage phase between the markers is known, PDQs can be computed conditional on marker haplotypes [11]. When linkage phase between the markers is not known, genotype information at one of the two flanking markers can be used to compute PDQs [23, 30]. The genotype at the marker locus may be ordered or unordered, and these two cases are considered separately. When the marker genotype is ordered, PDQs can be computed as described by Fernando and Grossman [8]. For example, the PDQ for MQTL allele \( Q^m_i \), conditional on the maternal marker allele inherited by \( i \), can be calculated as shown in Table (2.2). The PDQ for MQTL allele \( Q^f_i \), conditional on the paternal marker allele inherited by \( i \) can be calculated in a similar manner. When marker genotypes of an offspring are unordered, marker information can be ignored [8, 23]. However, as discussed later, this results in loss of information. The genotype of an offspring at a marker locus may be unordered only if it is heterozygous at that locus. Given

<table>
<thead>
<tr>
<th>Haplotype inherited</th>
<th>( Q^m_d )</th>
<th>( Q^f_d )</th>
<th>( Q^m_s )</th>
<th>( Q^f_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M^m_d )</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>( M^m_d )</td>
<td>( \frac{r_1}{r_1 + r_2} )</td>
<td>( \frac{r_2}{r_1 + r_2} )</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>( M^f_d )</td>
<td>( \frac{r_1}{r_1 + r_2} )</td>
<td>( \frac{r_1}{r_1 + r_2} )</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>( M^f_d )</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>( M^f_d )</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Given the maternal marker allele inherited by $i$, probability that MQTL allele $Q_i^m$ descends from the parental allele $Q_p^k$ (PDQ), where $p$ is $d$ or $s$ and $k$ is $m$ or $f$. $M_d^*$ denotes unknown descent.

<table>
<thead>
<tr>
<th>Allele inherited</th>
<th>$Q_p^m$</th>
<th>$Q_d^m$</th>
<th>$Q_s^m$</th>
<th>$Q_f^m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_d^m$</td>
<td>$1 - r_1$</td>
<td>$r_1$</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$M_d^f$</td>
<td>$r_1$</td>
<td>$1 - r_1$</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$M_d^s$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

that the genotype of an individual is heterozygous, it will be unordered if both its parents are heterozygous for the same alleles, or one of the parents is heterozygous for the same alleles while the marker information at the other parent is missing, or if the marker information is missing in both parents. When the marker genotype is unordered, PDQs can be calculated as described by Wang et al. [33] by multiplying a $2 \times 4$ matrix of PDMs by a $4 \times 4$ matrix involving recombination rates. If the marker genotypes are observed for both parents, the PDMs are easily obtained from formula (A1) in Wang et al. [33]. For example, when both parents and the offspring have genotype $A_1A_2$, the PDMs for marker allele $M_1^d$ are given in row one of Table (2.3). When marker genotypes are missing in the parents, Wang et al.

<table>
<thead>
<tr>
<th>Genotype of $M_p^k$</th>
<th>$M_d^1$</th>
<th>$M_d^2$</th>
<th>$M_s^1$</th>
<th>$M_s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1A_2$</td>
<td>$A_1A_2$</td>
<td>$A_1A_2$</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$A_1A_2$</td>
<td>-</td>
<td>$A_1A_2$</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>$A_1A_2$</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

[33] used equation (2.6) to compute the PDMs. But, this can be computationally demanding in large pedigrees with many missing genotypes. Thus, we compute the PDMs using only the marker genotypes that are observed in the parents. For example, if the marker genotype is missing in parent $s$ and is $A_1A_2$ for $d$ and $i$, the PDMs for marker allele $M_2^i$ ignoring all the
other marker information in the pedigree are given in row two in Table (2.3). Row three of Table (2.3) gives the PDMs for marker allele \( M_i \), ignoring all the other marker information in the pedigree, for the case when the marker genotype is missing for \( d \) and \( s \), and is \( A_1A_2 \) in \( i \). Thus, when marker genotypes of an offspring are unordered, PDMs of the type described above can be computed easily. Finally, when the genotypes at both markers are unobserved, the PDQ for each of the parental alleles is equal to 0.5.

It is important to note that if double recombination between markers is ignored some PDQs are equal to one (Table 2.1). When this occurs, the MQTL allele \( Q_i^m \), for example, is traced with certainty to MQTL allele \( Q_d^m \), and thus, \( \Pr(Q_i^m \equiv Q_d^m \mid G_{obs}) = 1 \). A similar situation will occur when, for example, \( \Pr(Q_i^m \equiv Q^e_d \mid G_{obs}) = 1 \). Recall that \( Q_i^m \) is either \( Q^m_d \) or \( Q_i^f \). Thus, regardless of the value of the PDQs, \( \Pr(Q_i^m \equiv Q_d^m \mid G_{obs}) = \Pr(Q_i^m \equiv Q^p_d \mid G_{obs}) = 1 \). When the IBD probability between any pair of MQTL alleles is one, the gametic variance covariance matrix will not be positive definite. To avoid this problem, if two alleles are IBD with probability one, only the effect of one of these two alleles is included in the mixed linear model. A side effect of this approach is the reduction in the number of equations in HMME and thus, an increase in the computational efficiency [11].

2.2.4 Calculation of the inverse of the gametic variance covariance matrix

The PDQs computed as described above can be used in formulae (18), (19), and (21) of Wang et al. [33] to obtain efficiently the inverse of the gametic variance covariance matrix. Formula (19) of Wang et al. [33] requires computing IBD probabilities between the MQTL alleles of parents. These were computed using the recursive formula (2.5), except for alleles within an individual with unordered markers. For individuals with unordered markers, IBD probabilities between their maternal and paternal alleles were computed using formula (11) in [33].

Recursive computation of the IBD probability between any pair of alleles may require IBD probabilities previously used in computing the IBD probability between other pairs of alleles. Thus, as in Abdel-Azim and Freeman [1], in order to avoid computing the same IBD
probability repeatedly, upon computation of an IBD probability it was stored for possible future use. While Abdel-Azim and Freeman [1] used linked lists to store the probabilities, we used a map container class of the C++ Standard Template Library. Each data item (an IBD probability in this case) stored in a map container class is indexed by a key. For element $i$ and $j$ of the IBD matrix, $i$ and $j$ were used as the key to store and retrieve this element.

2.2.5 Estimation of the exact genetic variance covariance matrix by MCMC

ESIP, an MCMC sampler that combines the Elston-Stewart algorithm with iterative peeling [7], was used to sample the genotypes for unobserved markers and all the MQTL genotypes jointly from the entire pedigree. Given genotypic effects and the sampled MQTL genotypes, a vector of genotypic values was obtained for the pedigree. The genetic variance covariance matrix was estimated from 15,000 independently distributed vectors of genotypic values. To validate this approach, the genetic variance covariance matrix estimated by ESIP was compared to the exact genetic variance covariance matrix calculated by using formula (27) of Wang et al. [33] for the case of a single marker linked to the MQTL.

2.2.6 Simulation study

Data were simulated for the hypothetical pedigree shown in Figure 2.1, to investigate the effect of approximating the gametic covariance matrix on response to selection by MABLUP. This pedigree spans four generations, has 96 individuals, four loops, and each of its nuclear families has 10 offspring.

In order to examine the effect of the number of alleles at the markers and the effect of missing marker data on the approximations, six experimental situations were considered. These consisted of simulations using 2, 4, or 12 alleles at the marker loci for complete and incomplete marker data. When incomplete marker data were simulated, individuals 1, 2, 3, 14, and 15 were assumed to have missing marker genotypes at both marker loci. For all six situations, phenotypic data were simulated only for individuals 1 to 46. Individuals 47 to 96, which represented the candidates for selection, did not have phenotypic data. The details of the simulation
are given in Table 2.4. For each of the six possible experimental situations, response to selection by MABLUP was calculated using approximate gametic variance covariance matrices obtained by the Method A and Method B, and by using the exact genetic variance covariance matrix estimated by MCMC.

<table>
<thead>
<tr>
<th>Table 2.4 Simulation details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of alleles per each marker</strong></td>
</tr>
<tr>
<td><strong>Marker allele frequencies, 2 alleles per marker</strong></td>
</tr>
<tr>
<td><strong>Marker allele frequencies, 4 alleles per marker</strong></td>
</tr>
<tr>
<td><strong>Marker allele frequencies, 12 alleles per marker</strong></td>
</tr>
<tr>
<td><strong>Recombination rate between the two markers and MQTL</strong></td>
</tr>
<tr>
<td><strong>Genetic variance explained by the MQTL</strong></td>
</tr>
<tr>
<td><strong>Phenotypic variance explained by the MQTL</strong></td>
</tr>
<tr>
<td><strong>Number of candidates for selection</strong></td>
</tr>
<tr>
<td><strong>Number of candidates selected</strong></td>
</tr>
</tbody>
</table>
2.3 Results

2.3.1 Validation of the use of ESIP to estimate the exact genetic variance covariance matrix

To validate ESIP as an accurate method to estimate the exact genetic variance covariance matrix, we used the simulated data for the situations where marker loci have 4 alleles for complete and incomplete marker data. However, because it is not feasible computationally to use formula (27) of Wang et al. [33] to calculate the exact genetic variance covariance matrix given information at flanking markers with missing genotypes, the marker data at one of the two marker loci was discarded. For each of the two situations considered, the exact genetic variance covariance matrix was calculated using formula (27) of Wang et al. [33], and it was estimated using 15,000 or 50,000 independently distributed vectors of genotypic values sampled by ESIP. The absolute difference between the exact genetic variance covariance matrix and the estimated variance covariance matrix was calculated for each of the two situations considered. For each matrix of absolute differences, the maximum, mean, and standard deviation of its elements were computed. These statistics were used to assess the accuracy of the variance covariance matrix estimated by ESIP. For both complete and incomplete marker data, the accuracy of the genetic variance covariance matrix estimated using 15,000 samples was considered sufficient (Tables 2.5 and 2.6).

Table 2.5 Maximum, mean and standard deviation of the elements of the matrices of absolute differences between the exact and the estimated genetic variance covariance matrices, for complete marker data with 4 alleles at the marker locus. The estimated matrices were calculated using 15,000 or 50,000 samples generated by ESIP.

<table>
<thead>
<tr>
<th>No. of samples</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,000</td>
<td>$7.0 \times 10^{-2}$</td>
<td>$1.3 \times 10^{-2}$</td>
<td>$1.0 \times 10^{-2}$</td>
</tr>
<tr>
<td>50,000</td>
<td>$3.0 \times 10^{-2}$</td>
<td>$6.7 \times 10^{-3}$</td>
<td>$5.0 \times 10^{-3}$</td>
</tr>
</tbody>
</table>
Table 2.6  Maximum, mean and standard deviation of the elements of the matrices of absolute differences between the exact and the estimated genetic variance covariance matrices, for incomplete marker data with 4 alleles at the marker locus. The estimated matrices were calculated using 15,000 or 50,000 samples generated by ESIP.

<table>
<thead>
<tr>
<th>No. of vectors used</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,000</td>
<td>$6.4 \times 10^{-2}$</td>
<td>$1.3 \times 10^{-2}$</td>
<td>$1.0 \times 10^{-2}$</td>
</tr>
<tr>
<td>50,000</td>
<td>$3.2 \times 10^{-2}$</td>
<td>$6.7 \times 10^{-3}$</td>
<td>$5.0 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

2.3.2 Comparison of response to selection obtained with different MABLUP methods

For each of the six experimental situations considered, the 5 highest ranking individuals were selected out of the 50 candidates for selection based on genetic evaluations obtained by: BLUP using only phenotypic data, MABLUP using the gametic variance covariance matrix calculated by Method A, MABLUP using the gametic variance covariance matrix calculated by Method B, MABLUP using the exact genetic variance covariance matrix estimated by ESIP (Method E). Response to selection obtained by BLUP using only phenotypic data was used as the reference value to compare methods A, B and E. For each of the three methods under investigation, the percent superiority in response by MABLUP over response by BLUP was calculated as

$$\frac{R_{MABLUP} - R_{BLUP}}{R_{BLUP}} \times 100,$$

where, for example, $R_{MABLUP}$ is the difference between the mean of the genotypic values of the candidates selected using MABLUP and the mean of the genotypic values of the candidates for selection. To visually determine the number of replicates of the simulation needed to compare methods A, B, and E, the running means of percent superiority of MABLUP by methods A, B, and E were plotted for each of the six experimental situations considered (Figures 2.2 to 2.7). For 12 alleles and incomplete marker data, we observed, based on 1650 replicates, that the
running mean was almost identical to the running mean of the complete marker data. Thus, for 12 alleles the same running mean was used for both complete and incomplete marker data. Based on these plots it was determined that 5,000 replicates of the simulation were sufficient to compare the methods.

![Graph](image)

Figure 2.2 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 2 alleles at the flanking markers for complete marker data.

2.3.2.1 Complete marker data

Figure 2.8 summarizes the percent superiority of MABLUP estimated from 5,000 replicates of the simulation, by methods A, B, and E, and by the number of alleles at a marker for complete marker data. Method A performed worst in all situations. Compared to Method
Figure 2.3 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 4 alleles at the flanking markers for complete marker data.
Figure 2.4  The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 12 alleles at the flanking markers for complete marker data.
Figure 2.5  The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 2 alleles at the flanking markers for incomplete marker data.
Figure 2.6  The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 4 alleles at the flanking markers for incomplete marker data.
Figure 2.7 The running means of percent superiority of MABLUP by methods A, B, and E for the situation with 12 alleles at the flanking markers for incomplete marker data.
Method B resulted in increased response to selection for all situations considered. For the situation with two alleles at the marker loci, the difference in superiority obtained with Method B versus superiority obtained with Method A was approximately 2.0%. For the situation with four alleles at the marker loci, the difference in superiority was approximately 3.7%, while for the situation with 12 alleles it was approximately 5.6%.

It can be seen from Figure 2.8 that as the number of alleles at the marker loci increases, the percent superiority of MABLUP by both method A and B increases. However, the increase for method A is more rapid, especially from 2 alleles to 4 alleles. Recall that in Method A marker information in a parent is used only if the markers are doubly heterozygous and their linkage phase is known. In contrast for Method B, even when only one of the two markers is heterozygous and even if the linkage phase is not known marker information in a parent may be used. In both methods the increase in percent superiority is due to the increase in the number of heterozygous genotypes at the two marker loci. A possible explanation for the more rapid change for Method A is that with 2 alleles most individuals were evaluated using only pedigree and trait information. When the number of alleles increased to 4, markers became informative for many of these individuals. In contrast for Method B, even with 2 alleles many individuals were evaluated using pedigree, trait and marker information at one locus. When the number of alleles increases to 4, many individuals that were originally evaluated using information at one marker are evaluated using information at both markers. The change in percent superiority of MABLUP is higher when going from using no marker information to using two markers, which is what happens with Method A, than when going from using one marker to using two markers, which is what happens with Method B.

Figure 2.8 also shows the effect of the increase in the number of alleles at the two marker loci on percent superiority of MABLUP by Method E. Again, the increase in the number of heterozygous genotypes at the marker loci resulted in an increase in superiority, for example, by ~ 6.8% when moving from 2 to 4 alleles. However, the rate of increase is higher for Method B than for Method E. A possible explanation for this result is that Method E always uses information at both marker loci, where phase information comes from the entire pedigree.
Thus, even with 2 alleles all informative markers are used. For Method E, the increase in percent response is due only to the increase in the number of informative markers.

Figure 2.8 Summary of percent superiority of MABLUP estimated from 5,000 replicates of the simulation, by methods A, B, and E, and by the number of alleles at a marker for complete marker data.

2.3.2.2 Incomplete marker data

When incomplete marker data were used (Figure 2.9), the percent superiority of MABLUP was smaller than when complete marker data were used for methods A and B. For Method E, a small reduction in response was observed for two and four alleles, but no reduction was observed for 12 alleles. For Method B, the change in percent superiority when going from two alleles to four alleles was lower than the corresponding change when complete marker data were
used. A possible explanation for this result is that, in calculating the PDQs for offspring

![Bar chart](image)

Figure 2.9 Summary of percent superiority of MABLUP estimated from 5,000 replicates of the simulation, by methods A, B, and E, and by the number of alleles at a marker for incomplete marker data.

of founders, the information at only one marker locus is used, and thus, when the marker genotypes of founders are missing, the effect of an increased number of heterozygous founders individuals is lost.

### 2.4 Discussion

This paper investigates the effect of using approximate gametic variance covariance matrices on response to selection by MABLUP. Two approximate methods to compute the gametic
variance covariance matrix conditional on marker information at two flanking markers were considered. For Method A, when the linkage phase between markers was known, marker information was used as described Goddard [11], and when linkage phase was not known, marker information was ignored completely. For Method B, when the linkage phase between markers was known, marker information was also used as described by Goddard [11], but when linkage phase was not known the methodology described by Fernando and Grossman [8], was combined with the methodology described by Wang et al. [33] to make use of the marker information at one of the two flanking markers.

Pong-Wong et al. [23] used a method that is similar to our Method B to compute the approximate gametic variance covariance matrix. Method B, however, differs from the method used by Pong-Wong et al. [23] as described below. First, when marker genotypes are missing, Pong-Wong et al. [23] used only the marker genotypes of parents and offspring to determine the missing marker genotypes. In contrast, in Method B genotype elimination [19] was used to determine the missing marker genotypes given the observable marker data for the entire pedigree. Second, Pong-Wong et al. [23] use a deterministic approach developed by Knott and Haley [18] to calculate IBD probabilities between offspring of founders conditional on marker information of both flanking loci. When this approach is used, however, the inverse of the gametic variance covariance matrix cannot be obtained efficiently. In Method B, to compute the PDQs for offspring from founders, we used only information from one marker locus, and all IBD probabilities were computed using the recursive formula (2.5). Thus, the gametic variance covariance obtained using Method B could be inverted efficiently. Finally, when the linkage phase was not known in non-founders both the method used by Pong-Wong et al. [23] and Method B used only one of the two flanking markers to obtain IBD probabilities. However, if the genotype at the marker locus used is unordered, Pong-Wong et al. [23] ignored the marker information at this locus. In contrast, for this situation, we used the approach of Wang et al. [33] to calculate IBD probabilities. The benefit of using marker information in this situation is described below. Consider, for example, the covariance between the MQTL effects of half-sibs that receive different marker alleles from their sire. If the marker genotype
in the sire is unordered. Pong-Wong et al. used only pedigree information to compute this covariance between these half-sibs, which is equal to $\frac{1}{2}\sigma^2_e$. However, if marker information is used as described by Wang et al. [33] the covariance between half-sibs will be $2(1 - r)r\sigma^2_e$.

Methods A and B yield approximate gametic variance covariance matrices due to the following reasons. The gametic variance covariance matrix is constructed in both methods using equation (2.5). However, when the marker genotypes for parents are missing, or even when marker data are complete but the linkage phase between flanking markers is unknown, this recursive equation yields approximate IBD probabilities. Furthermore, the PDQs required to calculate IBD probabilities using equation (2.5) are approximated to be able to accommodate large pedigrees with many missing markers. For all situations considered, Method B yielded a higher response to selection and thus indicates better use of the available marker information. However, the results obtained with Method E indicate that even Method B does not adequately utilize the observable marker information. At present, in Method B PDQs are computed based only on the marker information of the individual and its parents. Method B could be improved by computing the PDQs conditional on observable marker data from all "closely" related individuals [33]. This can be done by exact methods such as the Elston-Stewart algorithm [5], or by MCMC. When PDQs are estimated conditional on all the observable marker data, they are exact. However, computing exact PDQs by the Elston-Stewart algorithm would be feasible only for small pedigrees. In contrast, MCMC can be used to estimate exact PDQs for large pedigrees.

From the comparisons in this paper it is not possible to determine how much of the loss in response is due to the violation of the independence condition that is required to obtain equation (2.5) as opposed to the use of approximate PDQs. Use of exact PDQs in Method B in comparison to Method E would allow us to determine the loss in response caused by the use of equation (2.5).
Bibliography


CHAPTER 3. A COMPARISON OF ALTERNATIVE METHODS TO COMPUTE THE CONDITIONAL MEAN OF GENOTYPIC VALUES IN FINITE LOCUS MODELS

A paper submitted for publication to Genetic Selection Evolution

Liviu R. Totir, Rohan L. Fernando, Jack C.M. Dekkers, Soledad A. Fernández and Bernt Guldbrandtsen

Abstract

Increased availability of genotypes at marker loci has prompted the development of models that include the effect of individual genes. Selection based on these models is known as marker-assisted selection (MAS). MAS is known to be efficient especially for traits that have low heritability and non-additive gene action. BLUP methodology under non-additive gene action is not feasible for large inbred or crossbred pedigrees. It is easy to incorporate non-additive gene action in a finite locus model. Under such a model, the unobservable genotypic values can be predicted using the conditional mean. Simulated data were used to assess the performance of iterative peeling and of three Markov chain Monte Carlo (MCMC) methods (scalar Gibbs, blocking Gibbs, and ESIP) to calculate the conditional mean of the genotypic value under finite locus models. For pedigrees with loops, the accuracy of estimates obtained by iterative peeling decreases as the number of loci in the model increases. Also, computing time is exponentially related with the number of loci in the model. For MCMC methods, a linear relationship can be maintained by sampling genotypes one locus at a time. Out of the three MCMC methods considered, ESIP performed best while scalar Gibbs performed worst.
3.1 Introduction

Best linear unbiased prediction (BLUP) is currently the most widely used method for genetic evaluation. In large populations, BLUP can be obtained efficiently by solving Henderson's mixed model equations (HMME) [30]. To construct HMME, however, the inverse of the covariance matrix is needed for each random effect in the model. Under additive inheritance, efficient algorithms are available to compute the inverse of the covariance matrix for the genotypic effects in both purebred [28, 29, 39, 40] and crossbred populations [10, 34].

BLUP theory can also be used for marker-assisted genetic evaluation (MAGE), where trait phenotypes as well as marker genotypes are used to predict the unobservable genotypic values. Under additive inheritance, efficient algorithms are available to invert the matrices required in HMME for MAGE in purebred [16, 23, 31, 51] and in crossbred populations [53].

MAGE is most useful for traits with low heritability [36, 41], but such traits are known to have non-additive gene action. Under non-additive inheritance, however, BLUP is difficult to implement, especially when inbreeding is present [7]. This difficulty arises from the need to obtain the inverse of the genetic variance covariance matrix under non-additive inheritance. Although algorithms to invert the genetic variance covariance matrix under non-additive inheritance have been investigated [32, 42, 48], such algorithms are not feasible for large inbred or crossbred populations [14, 35].

To overcome the computing problems associated with BLUP under non-additive gene action, it has been proposed to predict the unobservable genotypic values using the conditional mean calculated under the assumption of a finite locus model [17, 24, 44]. Furthermore, crossbred data do not increase the complexity of this type of prediction. The conditional mean of the genotypic values is also known as the best predictor (BP) because it minimizes the mean square error of prediction, and it has been shown to be optimal as a selection criterion [4, 15]. Under a finite locus model, the conditional mean of the genotypic values can be calculated exactly by the Elston-Stewart algorithm [9], approximated by iterative peeling [13, 47], or estimated by Markov Chain Monte Carlo (MCMC) methods [17, 24, 44].

The Elston-Stewart algorithm was introduced in human genetics to compute the likelihood
of alternative models for quantitative traits under monogenic and oligogenic inheritance [9]. However, this algorithm is computationally feasible only for simple pedigrees [18]. Further, computing the conditional mean using the Elston-Stewart algorithm is not feasible for models with more than about three loci.

Van Arendonk et al. [47] introduced iterative peeling to approximate genotype probabilities of quantitative traits under monogenic inheritance in the complex pedigrees that are commonly encountered in livestock populations. Iterative peeling yields exact probabilities for pedigrees without loops [18, 52], and for pedigrees with loops, very accurate genotype probabilities have been obtained in the analysis of phenotypes of a biallelic disease locus [13]. The performance of iterative peeling for computing the conditional mean under finite locus models with more than one locus has not been studied.

During the last decade, the potential of MCMC methods has been recognized to overcome computational problems in analyses of genetic data [25, 26, 43, 49, 50]. This is especially true for analyses that use mixtures of distributions such as segregation analyses [9, 33] and combined segregation and linkage analyses [45]. One of the simplest and most widely used MCMC methods is the scalar Gibbs sampler [19, 20]. Guo and Thompson [25, 26] have discussed the use of the scalar Gibbs sampler in human genetics. Janss et al. [33] studied the potential of using the Gibbs sampler to analyze quantitative traits in animal genetics. They found that the scalar Gibbs sampler has mixing problems in pedigrees that contain large sibships. This is due to the dependence between the genotypes of parents and offspring [33]. Scalar Gibbs is, however, still one of the most widely used MCMC methods for genetic analyses [1, 8, 37, 38]. Blocking Gibbs was recommended as an alternative to scalar Gibbs in order to overcome the dependence problem [33]. The blocking scheme suggested by Janss et al. [33], samples the genotype of a sire jointly with the genotypes of its terminal offspring. A more extreme alternative is to use peeling and reverse peeling to sample jointly the genotypes of all animals in a pedigree [13, 27]. This strategy, however, is not feasible when the pedigree contains many nested loops. For such pedigrees, it has been proposed to use an approximate method to obtain candidate samples and accept or reject these by the Metropolis-Hastings algorithm [13, 27].
MCMC sampler, known as ESIP, combines the Elston-Stewart algorithm with iterative peeling to obtain candidate samples from the entire pedigree, which are then accepted or rejected using a Metropolis-Hastings algorithm [13].

The objective of this paper is to study the performance of iterative peeling, scalar Gibbs, blocking Gibbs, and ESIP when used to calculate the conditional mean of the genotypic value for a quantitative trait in finite locus models. Simulated data are used to assess the performance of the methods.

3.2 Methods

Consider a trait determined by $N$ segregating quantitative trait loci (QTL) with two alleles at each locus. For a population of $n$ individuals, a given genotypic configuration of this trait can be written as a matrix $G$ of dimension $n \times N$

$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1N} \\ g_{21} & g_{22} & \cdots & g_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nN} \end{bmatrix}, \quad (3.1)$$

where $g_{ij}$ denotes the genotype of individual $i$ at locus $j$. $G$ can also be written as

$$G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_i \\ \vdots \\ g_n \end{bmatrix}, \quad (3.2)$$

where $g_i$ is the $1 \times N$ vector of genotypes of individual $i$, or as

$$G = \begin{bmatrix} c_1 & c_2 & \cdots & c_j & \cdots & c_N \end{bmatrix}, \quad (3.3)$$
where \( e_j \) is the \( n \times 1 \) column vector of genotypes at locus \( j \). When only additive and dominance gene actions are present, the vector \( v \) of genotypic values of \( n \) individuals can be modeled as

\[
v = \mathbf{1} \eta + \sum_{j=1}^{N} v_j = \mathbf{1} \eta + \sum_{j=1}^{N} Q_j \delta_j, \tag{3.4}
\]

where \( \mathbf{1} \) is a \( n \times 1 \) vector of ones; \( \eta \) is the trait mean \([11]\); \( v_j \) is the \( n \times 1 \) vector of genotypic values at locus \( j \) deviated from the trait mean; \( Q_j \) is an \( n \times 3 \) incidence matrix relating the genotypic deviations at locus \( j \) to the corresponding individuals, with each row of \( Q_j \) being one of the vectors \([1 \, 0 \, 0]^T\), \([0 \, 1 \, 0]^T\), or \([0 \, 0 \, 1]^T\); and \( \delta_j \) is a \( 3 \times 1 \) vector that contains the genotypic effects at locus \( j \): \([a_j \, d_j \, -a_j]^T\) \([11]\). The vector \( y \) of phenotypic values of \( n \) individuals under a finite locus model can be written as

\[
y = X \beta + Z (1 \eta + Q \delta) + e, \tag{3.5}
\]

where \( X \) is the incidence matrix relating the vector \( \beta \) of fixed effects to \( y \); \( Z \) is the incidence matrix relating \( v \) to \( y \); \( Q = [Q_1 \, Q_2 \, \ldots \, Q_N] \); \( \delta = [\delta_1 \, \delta_2 \, \ldots \, \delta_N]^T \); \( e \) is the vector of residuals. The parameters of this model are: \( \beta, \eta \), the genotypic effects \( a_j \) and \( d_j \), and gene frequency \( p_j \) for locus \( j = 1, \ldots, N \), and the residual variance \( \sigma^2 \). In this paper, we assume all parameters are known. The only unknowns are the genotypes at the \( N \) loci.

The conditional mean of the vector of genotypic values given phenotypic values, which is also the best predictor \((BP)\), can be written as

\[
E(v \mid y) = \mathbf{1} \eta + \sum_G v_G \Pr(G \mid y), \tag{3.6}
\]

where \( v_G \) is the vector of genotypic deviations that corresponds to the genotypic configuration
Under a finite locus model, the phenotypic values are assumed to be independent given the genotypes. As a result the conditional probability of the phenotypic values given $G$ can be written as

$$
\Pr(y \mid G) = \prod_{i=1}^{n} \Pr(y_i \mid g_i),
$$

(3.8)

where $\Pr(y_i \mid g_i)$ is the penetrance function for individual $i$. If individuals are numbered such that ancestors precede descendants, and if the founder genotypes are assumed to be independent, the probability of a given genotypic configuration can be written as

$$
\Pr(G) = \prod_{i \in F} \Pr(g_i) \prod_{i \in C} \Pr(g_i \mid g_{mi} \cdot g_{fi}),
$$

(3.9)

where $F$ is the set of founder individuals and $C$ is the set of nonfounders. For $i \in F$, the probability of the vector $g_i$ of genotypes for individual $i$ can be written as

$$
\Pr(g_i) = \prod_{j=1}^{N} \Pr(g_{ij}),
$$

(3.10)

where $\Pr(g_{ij})$ is equal to the population frequency of $g_{ij}$. Assuming the QTL are unlinked, for $i \in C$ the conditional probability that offspring $i$ will have the genotype vector $g_i$ given the parents of $i$ have the genotype vectors $g_{mi}$ and $g_{fi}$ can be written as

$$
\Pr(g_i \mid g_{mi}, g_{fi}) = \prod_{j=1}^{N} \Pr(g_{ij} \mid g_{mij}, g_{fij}),
$$

(3.11)

where $\Pr(g_{ij} \mid g_{mij}, g_{fij})$ is the conditional probability that offspring $i$ will have genotype $g_{ij}$ at locus $j$ given the parents of $i$ have genotypes $g_{mij}$ and $g_{fij}$ at locus $j$ [2, 9].

The key problem in any implementation of genetic evaluation using a finite locus model is...
the correct and efficient calculation of the sum over all possible genotypic configurations \( G \) in equation (5.1). The following methods are used here: the Elston-Stewart algorithm, iterative peeling, and three different MCMC methods (scalar Gibbs, blocking Gibbs, and ESIP).

### 3.2.1 Elston-Stewart algorithm

For simple pedigrees and models with up to three loci, the Elston-Stewart algorithm [9] can be used to compute efficiently the sum over all genotypic configurations and obtain exact genetic evaluations. These exact genetic evaluations were used here as reference values to assess the performance of the four methods under investigation.

### 3.2.2 Iterative peeling

Iterative peeling applied to pedigrees has been discussed by several authors [18, 47, 52]. When pedigrees have loops, iterative peeling results in an extended pedigree [52]. Fernandez et al. [13] describe iterative peeling using directed graphs to represent pedigrees. They provide general expressions that allow the use of iterative peeling in arbitrary directed graphs. Fernandez et al. [13] implemented iterative peeling for the analysis of phenotypic data of a biallelic disease locus. For this type of inheritance, the genotype completely determines the phenotype, and thus, the penetrance function is a simple indicator function. For the purpose of this paper, we use the approach of Fernandez et al. [13], but for models with different numbers of independent loci. For these models, the calculation of transition probabilities is done as shown in equation (3.11). The penetrance function is also modified to accommodate these types of models.

### 3.2.3 MCMC methods

#### 3.2.3.1 General considerations

Monte Carlo integration can be used to estimate expectations of random variables [22]. The BP can be estimated by simple Monte Carlo integration if we can draw independent samples from \( \Pr(\mathbf{G} \mid \mathbf{y}) \). In most cases, however, it is not feasible to draw independent samples from
this distribution. Often it is feasible to generate samples from a Markov chain with \( \Pr(G \mid y) \) as its stationary distribution. Monte Carlo integration using samples from a Markov chain is called MCMC. All three MCMC methods under investigation (scalar Gibbs, blocking Gibbs, and ESIP) give accurate results if the Markov chains are sufficiently long. The efficiency of these methods is characterized by the computing time needed to obtain accurate results. Various convergence diagnostics are used to determine the length required for accurate results [3, 22]. However, none of the available convergence diagnostics is foolproof [3, 22]. For all situations considered in this paper, exact evaluations of BP can be calculated by the Elston-Stewart algorithm. Thus, we do not need to rely on convergence diagnostics to determine the length of the chain required to obtain accurate results.

For each of the three MCMC methods under investigation, an initial sample from \( \Pr(G \mid y) \) is needed. To obtain this, the genotypes of ancestors are sampled before those of descendants. For founders, genotypes are sampled using the cumulative distribution function (cdf) of \( (g_i \mid y_i) \). For nonfounders, genotypes are sampled using the cdf of \( (g_i \mid g_m, g_f, y_i) \). Once an initial sample has been obtained, new genotype samples are generated one locus at a time conditional on the genotypes at all the other loci. Before moving to the next locus, genotypes are sampled within the current locus for all individuals. The three MCMC methods differ in the way the genotypes are sampled within a locus.

### 3.2.3.2 Scalar Gibbs

For scalar Gibbs, each \( g_{ij} \) is sampled conditional on \( y \) and all the other genotypes \( (G_{ij-}) \). Due to the Markovian nature of genetic data, however, the genotype of an individual is completely determined by the genotypes of the individuals that form its neighborhood: parents, spouses, and descendants. As a result, the genotype \( g_{ij}^t \) of nonfounder \( i \) at locus \( j \) in step \( t \) is sampled from

\[
\Pr(g_{ij} \mid y, G_{ij-}) = \frac{\Pr(g_{ij} \mid g_{mij}, g_{fij}) \Pr(y_i \mid g_{ij}^t) \prod_{k \in O_i} \Pr(g_{kj}^t \mid g_{ij}^t, g_{oj}^t)}{\sum_{g_{ij}} \text{numerator}}, \quad (3.12)
\]
where \( g_{mij} \) and \( g_{fij} \) represent the current genotypes of the parents of \( i \):

\[
g_i^t = [g_{i1}^t \ g_{i2}^t \ \ldots \ g_{ij-1}^t \ g_{ij} \ g_{ij+1} \ \ldots \ g_{iN}^{t-1}];
\]

(3.13)

\( O_i \) is the set of offspring of \( i \); \( g_{kj}^t \) is the current genotype of offspring \( k \) at locus \( j \); \( g_{oij}^t \) is the current genotype of the other parent of \( k \) at locus \( j \). For founders the same formula is used except that \( \Pr(g_{kj}^t | g_{mij}^t, g_{fij}^t) \) is replaced with \( \Pr(g_{kj}^t) \). This sampling process is repeated for all individuals within locus \( j \). Once all individuals have been sampled within locus \( j \), the same process is repeated for locus \( j + 1 \).

3.2.3.3 Blocking Gibbs

For blocking Gibbs, genotypes at locus \( j \) are sampled using the blocking scheme suggested by Janss et al. [33], where the genotypes of sires and their terminal offspring are sampled jointly. For sire \( i \) with a set \( T_i \) of terminal offspring, \( g_{ij} \) is sampled conditional on \( y \) and all other genotypes except the genotypes at locus \( j \) for the terminal offspring \( (G_{ij,T_i,j-}) \). Thus, the genotype \( g_{ij}^t \) of a nonfounder sire \( i \) at locus \( j \) in step \( t \) is sampled from

\[
\Pr(g_{ij} | y, G_{ij,T_i,j-}) =
\]

\[
\frac{\Pr(g_{ij}^t | g_{mij}^t, g_{fij}^t) \Pr(y | G^t) \prod_{k \in N_i} \Pr(g_{kj}^t | g_{oij}^t) \prod_{l \in T_i} \sum_{g_{lji}^t} \Pr(g_{li}^t | g_{oij}^t, y_l) \Pr(y_l | G^t)}{\sum_{g_{lji}^t} \text{numerator}},
\]

(3.14)

where \( N_i \) is the set of non terminal offspring of \( i \); \( g_{oij}^t \) is the current genotype of the other parent of \( k \) at locus \( j \); \( g_{oij}^t \) is the current genotype of the other parent of \( l \) at locus \( j \);

\[
g_i^t = [g_{i1}^t \ g_{i2}^t \ \ldots \ g_{ij-1}^t \ g_{ij} \ g_{ij+1} \ \ldots \ g_{iN}^{t-1}].
\]

(3.15)

For founder sires the same formula is used except that \( \Pr(g_{ij} | g_{mij}^t, g_{fij}^t) \) is replaced with \( \Pr(g_{ij}) \). For terminal offspring \( l \) of sire \( i \), \( g_{ij}^t \) is sampled from the cdf of \( (g_{ij} | g_{ij}^t, g_{oij}^t, y_l) \). For other individuals, \( g_{ij}^t \) is sampled according to (3.12). Once all individuals have been sampled
within locus \( j \), the same process is repeated for locus \( j + 1 \).

### 3.2.3.4 ESIP

For ESIP, genotypes at locus \( j \) are sampled as described by Fernandez et al. [13], where joint genotype samples from the entire pedigree are obtained by reverse peeling [13, 27]. For example, a sample in step \( t \) is obtained by sampling sequentially

\[
\begin{align*}
  g_{1j}^t & \text{ from } \Pr(g_{1j}^t \mid y, G_{j-1}^t), \\
  g_{2j}^t & \text{ from } \Pr(g_{2j}^t \mid y, G_{j-1}^t, g_{1j}^t), \\
  g_{3j}^t & \text{ from } \Pr(g_{3j}^t \mid y, G_{j-1}^t, g_{1j}^t, g_{2j}^t), \\
  & \vdots \\
  g_{nj}^t & \text{ from } \Pr(g_{nj}^t \mid y, G_{j-1}^t, g_{1j}^t, g_{2j}^t, g_{3j}^t, \ldots, g_{n-1j}^t). \\
\end{align*}
\]

(3.16)

where \( G_{j-1}^t = \begin{bmatrix} c_1^t \ldots c_{j-1}^t \ldots c_{j}^{t-1} \ldots c_{N}^{t-1} \end{bmatrix} \) is the current genotype configuration at all the other loci except locus \( j \) at step \( t \). Note that the resulting sample comes from

\[
\Pr(g_{1j}, g_{2j}, g_{3j}, \ldots, g_{nj} \mid y, G_{j-1}^t) = \Pr(c_j \mid y, G_{j-1}^t).
\]

(3.17)

where \( c_j \) is the genotype configuration at locus \( j \). The Elston-Stewart algorithm can be used to calculate the probabilities needed in the sampling process [5, 9]. In the Elston-Stewart algorithm, intermediate results must be stored in multidimensional tables called cutsets [13]. For pedigrees without loops, only two-dimensional tables are generated. For pedigrees with many nested loops, the dimension of the cutsets may increase to the point that the Elston-Stewart algorithm may not be feasible anymore. As a result, the Elston-Stewart algorithm cannot be used for this type of pedigrees. Fernandez et al. [13] have combined the Elston-Stewart algorithm with iterative peeling to make the joint sampling of genotypes feasible for arbitrary pedigrees. In this combined approach, the Elston-Stewart algorithm is used while the cutset size is small enough, and iterative peeling is used for the remainder of the pedigree. It can be shown that results from iterative peeling are equivalent to those obtained by the Elston-
Stewart algorithm for a modified pedigree [52]. By using the combined approach candidate samples from a modified pedigree are generated. These candidate samples are then accepted or rejected through a Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm used corresponds to the special case of independence sampling [13]. For this case, the acceptance probability of a move from the genotype configuration \( c_j^{t-1} \) to genotype configuration \( c_j^t \) is given by

\[
\alpha(c_j^{t-1}, c_j^t | G_{j-}) = \min \left( 1, \frac{\pi(c_j^t | G_{j-}) \times q(c_j^{t-1} | G_{j-})}{\pi(c_j^{t-1} | G_{j-}) \times q(c_j^t | G_{j-})} \right).
\]  

(3.18)

where

\[
\pi(c_j^t | G_{j-}) = \Pr(c_j^t | y, G_{j-})
\]

(3.19)

is the target probability of the genotype configuration \( c_j^t \);

\[
\pi(c_j^{t-1} | G_{j-}) = \Pr(c_j^{t-1} | y, G_{j-})
\]

(3.20)

is the target probability of the genotype configuration \( c_j^{t-1} \);

\[
q(c_j^t | G_{j-}) = \Pr_M(c_j^t | y, G_{j-})
\]

(3.21)

is the probability of the candidate sample, where the subscript \( M \) is used to denote that, if iterative peeling is used, this sample is drawn from a modified pedigree. Finally,

\[
q(c_j^{t-1} | G_{j-}) = \Pr_M(c_j^{t-1} | y, G_{j-})
\]

(3.22)

is the probability of \( c_j^{t-1} \), if \( c_j^{t-1} \) would be sampled from the same distribution as \( c_j^t \). The target probability of genotype configuration \( c_j^t \), for example, is calculated as follows

\[
\pi(c_j^t | G_{j-}) \propto \prod_{i \in F} \Pr(g_{ij}^t) \Pr(y_i | g_i^t) \prod_{i \in C} \Pr(g_{ij}^t | g_{i,mij}^t, g_{fij}^t) \Pr(y_i | g_i^t).
\]

(3.23)
Next consider the calculation of \( q(c_j^t \mid G_{j-}^t) \). This can be done as follows

\[
q(c_j^t \mid G_{j-}^t) = \Pr_M(g_{ij}^t \mid y, G_{j-}^t) \times \Pr_M(g_{ij}^t \mid y, G_{j-}^t, g_{ij}^t) \\
\times \Pr_M(g_{ij}^t \mid y, G_{j-}^t, g_{ij}^t, g_{ij}^t) \times \ldots \\
\times \Pr_M(g_{ij}^t \mid y, G_{j-}^t, g_{ij}^t, g_{ij}^t, \ldots, g_{ij}^t).
\]

(3.24)

where \( g_{ij}^t \) denotes the genotype sampled for animal \( i \) at locus \( j \) in step \( t \). Note that all probabilities that form the product in equation (3.24) were already calculated in the reverse peeling process used to sample \( c_j^t \). Now consider the calculation of \( q(c_{j-}^{t-1} \mid G_{j-}^t) \). This is not as straightforward because \( c_{j-}^{t-1} \) was sampled from \( \Pr_M(c_j \mid y, G_{j-}^{t-1}) \), while what we need to calculate is \( q(c_{j-}^{t-1} \mid G_{j-}^t) \). This probability can be calculated as follows

\[
q(c_{j-}^{t-1} \mid G_{j-}^t) = \Pr_M(g_{ij}^{t-1} \mid y, G_{j-}^t) \times \Pr_M(g_{ij}^{t-1} \mid y, G_{j-}^t, g_{ij}^{t-1}) \\
\times \Pr_M(g_{ij}^{t-1} \mid y, G_{j-}^t, g_{ij}^{t-1}, g_{ij}^{t-1}) \times \ldots \\
\times \Pr_M(g_{ij}^{t-1} \mid y, G_{j-}^t, g_{ij}^{t-1}, g_{ij}^{t-1}, \ldots, g_{ij}^{t-1}).
\]

(3.25)

where \( g_{ij}^{t-1} \) denotes the genotype sampled for animal \( i \) at locus \( j \) in step \( t - 1 \). The probabilities that form the left-hand side product in equation (3.25) are calculated using the same intermediate results from the Elston-Stewart algorithm that were used to calculate the probabilities that form the left-hand side product of equation (3.24).

Finally, note that if only the Elston-Stewart algorithm is used to calculate the probabilities needed in the sampling process, \( q \) is the same as \( \pi \), and as a result all samples are accepted.

### 3.2.4 Simulation study

Three hypothetical pedigrees were used to assess the performance of the four methods under investigation. The first hypothetical pedigree is shown in Figure 3.1. This pedigree has 96 individuals, four loops, and each of its nuclear families has 10 offspring. This pedigree will be referred to as the base pedigree. The second pedigree is an extension of the base pedigree. The extension was done by duplicating the structure of the base pedigree in the
next three generations. In order to obtain the extension from the base pedigree, we assigned to individuals 66, 67, 87, 77, 56 the same parental role as that of individuals 1, 2, 3, 14, 15. As a result, the second pedigree has seven generations and 187 individuals and will be referred to as the extended pedigree. Finally, a third pedigree with a family structure typical for a poultry population was considered. This pedigree consists of one male mated to eight females each mating producing 15 offspring. It has 129 individuals and no loops and will be referred to as the poultry pedigree.

In order to examine the effect of pedigree structure, missing data, number of loci in the model, and genetic parameters on the accuracy of genetic evaluations, eight situations were simulated (Tab. 3.1). For each situation, the simulation model and the analysis models were identical. The simulation study was designed so that the Elston-Stewart algorithm can be used to obtain exact genetic evaluations for each situation considered. All loci of a given finite locus model had the same parameters. Thus, all loci had equal gene frequencies and additive and dominance effects. Situation 3 was used as the reference situation in the design of the simulation study. The genetic parameters for this situation are similar to estimates reported.
Table 3.1 Situations simulated. No. missing denotes the number of parents with missing phenotypic information. \( h_n^2 \) denotes the narrow sense heritability and \( h_b^2 \) denotes the broad sense heritability.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Pedigree</th>
<th>No. loci</th>
<th>No. missing</th>
<th>( h_n^2 )</th>
<th>( h_b^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>base</td>
<td>1</td>
<td>15</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>base</td>
<td>1</td>
<td>15</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>base</td>
<td>2</td>
<td>15</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>base</td>
<td>2</td>
<td>15</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>base</td>
<td>2</td>
<td>0</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>6</td>
<td>extended</td>
<td>2</td>
<td>15</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>7</td>
<td>poultry</td>
<td>2</td>
<td>9</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>8</td>
<td>poultry</td>
<td>3</td>
<td>9</td>
<td>0.04</td>
<td>0.08</td>
</tr>
</tbody>
</table>

in the animal science literature for low heritable traits that exhibit non-additive gene action [6]. For this situation, all parents in the base pedigree (15 individuals) are assumed to have missing phenotype information.

The first four situations of Table 3.1 were designed to consider all possible combinations of two heritabilities (0.04 and 0.4) and two values for the number of loci in the model (one and two). This design allowed us to examine the main effects of heritability and number of loci in the model, as well as the effect of their interaction, for the base pedigree. Situation 5, which differs from situation 3 only in the number of missing phenotypes, was considered to examine the effect of missing data. Situations 6 and 7, which differ from situation 3 only in the pedigree structure, were considered to examine the effect of the pedigree. Situation 8, which differs from situation 7 only in the number of loci, was considered to examine the effect of the number of loci in the poultry pedigree. For the base and extended pedigree, only models with one or two loci were considered due to the computational limitations of the Elston-Stewart algorithm.

Equation (5.1) was used to obtain estimates of genotypic values. In (5.1), the sum over the possible genotypic configurations was calculated exactly when the Elston-Stewart algorithm was used. When iterative peeling was used, the sum was calculated exactly for pedigrees without loops and approximated for pedigrees with loops. Finally, when the MCMC methods
were used, the sum was estimated by sampling.

For each situation, the scaled absolute difference between estimates of genotypic values obtained with each of the four methods under investigation (iterative peeling, scalar Gibbs, blocking Gibbs, and ESIP) and estimates obtained with the Elston-Stewart algorithm was calculated for each individual. For a given situation, the maximum and mean of the scaled absolute differences, as well as the scaled square root of the mean square error, were used to investigate the performance of the four methods under investigation. The scaling factor used was the genetic standard deviation for each situation considered.

Estimates obtained using MCMC methods depend on the number of samples used to calculate them. To have a fair comparison between the three MCMC methods, equal computing time was allocated to each method. The mean sum of squares of the unscaled absolute differences was used as the criterion to decide the amount of the computing time. The mean sum of squares was calculated at different stages of the chains so that the mixing properties of the three MCMC methods could be assessed as well.

3.3 Results

3.3.1 Iterative peeling

Table 3.2 summarizes the scaled absolute difference between estimates of genetic values obtained using iterative peeling and estimates obtained using the Elston-Stewart algorithm. For iterative peeling, the estimates were obtained using five iterations. The effect of a larger number of iterations on the accuracy of genetic evaluations was negligible. Fernandez et al. [13] showed that iterative peeling yields very good approximations for conditional genotype probabilities in the case of a recessive disease trait. The results obtained for the one-locus models (situations 1 and 2) indicate that a similar conclusion holds true for quantitative traits with dominance action, with low or high heritability. However, the results for low heritability (situations 1 and 3) showed that the approximation was not as good as with high heritability (situations 2 and 4). The approximations obtained for two locus models (situations 3 and 4) were also inferior to those obtained for one-locus models (situations 1 and 2). The most notable
Table 3.2 Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using iterative peeling compared to estimates calculated using the Elston-Stewart algorithm. In situations 1 and 2, the genetic standard deviation was 2.83 (see Tab. 3.1). In situations 3, 4, 5 and 6, the genetic standard deviation was 4.0 (see Tab. 3.1).

<table>
<thead>
<tr>
<th>Situation</th>
<th>Maximum</th>
<th>Mean</th>
<th>√MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.5 × 10^{-2}</td>
<td>2.4 × 10^{-2}</td>
<td>3.2 × 10^{-2}</td>
</tr>
<tr>
<td>2</td>
<td>3.6 × 10^{-3}</td>
<td>1.4 × 10^{-4}</td>
<td>5.1 × 10^{-4}</td>
</tr>
<tr>
<td>3</td>
<td>1.3 × 10^{-1}</td>
<td>4.2 × 10^{-2}</td>
<td>5.2 × 10^{-2}</td>
</tr>
<tr>
<td>4</td>
<td>5.4 × 10^{-2}</td>
<td>2.4 × 10^{-3}</td>
<td>7.7 × 10^{-3}</td>
</tr>
<tr>
<td>5</td>
<td>1.3 × 10^{-1}</td>
<td>3.6 × 10^{-2}</td>
<td>4.7 × 10^{-2}</td>
</tr>
<tr>
<td>6</td>
<td>1.8 × 10^{-1}</td>
<td>6.0 × 10^{-2}</td>
<td>7.3 × 10^{-2}</td>
</tr>
</tbody>
</table>

result for situation 3, however, was the magnitude of the maximum of the scaled differences: 0.13 genetic standard deviations. This value corresponded to individual 13, which is one of the parents.

For the base pedigree, missing phenotypic records had limited impact, as seen by comparing results of situation 3 with situation 5. Iterative peeling performed worst for the extended pedigree of situation 6, which has a larger number of loops. Note that, for this situation, the maximum of the scaled differences was 0.18 genetic standard deviations and corresponded to a parent. Iterative peeling yielded exact results for situations 7 and 8 because the poultry pedigree has no loops.

### 3.3.2 Influence of the number of loci on computing efficiency

As described below, the exponential relationship between computing efficiency and the number of loci in the model restricts the practical use of iterative peeling to models with about three loci. With iterative peeling, genotype probabilities must be calculated for every multilocus genotype. Given two alleles at each locus, the number of possible genotypes is $3^N$. Iterative peeling involves working with a three-dimensional table of conditional probabilities for the genotype of an offspring given the genotypes of its parents. Thus the number of
computations required is proportional to

\[(3^N)^3 \times n \times i,\]  

(3.26)

where \(i\) is the number of iterations. In contrast, when MCMC samplers are used, a linear relationship between computing efficiency and the number of loci in the model can be maintained by sampling genotypes one locus at a time. Table 3.3 reflects this linear relationship for each of the three MCMC samplers under investigation. Next we assess the performance of the

<table>
<thead>
<tr>
<th>Sampler</th>
<th>No. of loci</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Esip</td>
<td>83</td>
</tr>
<tr>
<td>Blocking Gibbs</td>
<td>12</td>
</tr>
<tr>
<td>Scalar Gibbs</td>
<td>6</td>
</tr>
</tbody>
</table>

three MCMC samplers for each of the eight situations considered.

### 3.3.3 MCMC methods

#### 3.3.3.1 Conversion of computing time to chain length

ESIP was used as the reference sampler to determine the number of samples to be generated with each MCMC method. First, ESIP was run until the mean sum of squares of the difference between estimates obtained by ESIP and estimates obtained by the Elston-Stewart algorithm was below 0.0001. The other two samplers were then used for the same amount of computing time by converting computing time to the number of samples that can be generated. For the eight situations considered, Table 3.4 shows the number of samples that were generated by blocking and scalar Gibbs, in the same amount of time required by ESIP to generate one sample. Note that the sampling speed was affected by the pedigree structure. The sampling
Table 3.4  Number of samples obtained by blocking and scalar Gibbs in the same amount of time required by ESIP to generate 1 sample for the eight situations considered.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Blocking Gibbs</th>
<th>Scalar Gibbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 4</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>7,8</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

speed of blocking and scalar Gibbs increased with the number of loops in the model. The presence or absence of missing phenotypes had an effect on sampling speed as well.

In order to investigate the mixing behavior of the three MCMC samplers, the mean sum of squares of the difference between estimates obtained by MCMC and estimates obtained by the Elston-Stewart algorithm was calculated at different stages of each MCMC run. For seven of the eight situations considered, the mean sum of squares of the differences was the lowest for ESIP at all stages of the chain. Figure 3.2 shows this type of behavior for situation 6. For situation 4, which represents the exception, all three samplers reached a high level of accuracy in a short period of time and the mean sum of squares had a similar behavior for the first two-thirds of the allocated computing time for all samplers (Fig. 3.3). In the last third of the run, however, ESIP had the lowest mean sum of squares of the differences and reached the convergence criterion of 0.0001. The performance of the three MCMC samplers, based on the estimates of genotypic values obtained at the end of the allocated computing time, was assessed using the same statistics as used for iterative peeling.

3.3.3.2 ESIP

Table 3.5 summarizes the scaled absolute difference between estimates of genetic values obtained by ESIP at the end of the allocated computing time and estimates obtained by the Elston-Stewart algorithm. Due to the fact that ESIP was used as the reference sampler, the accuracy of ESIP estimates were similar for all situations. It is of interest, however, to examine the difference in the number of samples needed to reach the desired level of accuracy.
Figure 3.2  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 65000 samples for blocking Gibbs, or 130000 samples for scalar Gibbs, for situation 6. □ ESIP; + Blocking Gibbs;△ Scalar Gibbs.
Figure 3.3  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 35000 samples for blocking Gibbs, or 70000 samples for scalar Gibbs, for situation 4. □ ESIP; + Blocking Gibbs; △ Scalar Gibbs.
Table 3.5  Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using ESIP compared to estimates calculated using the Elston-Stewart algorithm. In situation 1 and 2, the genetic standard deviation was 2.8284 (see Tab. 3.1). In situations 3, 4, 5, 6 and 7, the genetic standard deviation was 4.0 (see Tab. 3.1). In situation 8 the genetic standard deviation is equal to 4.9 (see Table 3.1).

<table>
<thead>
<tr>
<th>Situation</th>
<th>No. samples included</th>
<th>Maximum</th>
<th>Mean</th>
<th>(\sqrt{\text{MSE}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>75 000</td>
<td>(1.1 \times 10^{-2})</td>
<td>(2.8 \times 10^{-3})</td>
<td>(3.5 \times 10^{-3})</td>
</tr>
<tr>
<td>2</td>
<td>3 500</td>
<td>(1.8 \times 10^{-2})</td>
<td>(1.5 \times 10^{-3})</td>
<td>(3.2 \times 10^{-3})</td>
</tr>
<tr>
<td>3</td>
<td>195 000</td>
<td>(8.0 \times 10^{-3})</td>
<td>(1.9 \times 10^{-3})</td>
<td>(2.5 \times 10^{-3})</td>
</tr>
<tr>
<td>4</td>
<td>250 000</td>
<td>(1.5 \times 10^{-2})</td>
<td>(8.6 \times 10^{-4})</td>
<td>(2.5 \times 10^{-3})</td>
</tr>
<tr>
<td>5</td>
<td>180 000</td>
<td>(8.0 \times 10^{-3})</td>
<td>(1.8 \times 10^{-3})</td>
<td>(2.5 \times 10^{-3})</td>
</tr>
<tr>
<td>6</td>
<td>175 000</td>
<td>(6.2 \times 10^{-3})</td>
<td>(2.0 \times 10^{-3})</td>
<td>(2.4 \times 10^{-3})</td>
</tr>
<tr>
<td>7</td>
<td>175 000</td>
<td>(7.1 \times 10^{-3})</td>
<td>(2.0 \times 10^{-3})</td>
<td>(2.5 \times 10^{-3})</td>
</tr>
<tr>
<td>8</td>
<td>230 000</td>
<td>(5.6 \times 10^{-3})</td>
<td>(1.6 \times 10^{-3})</td>
<td>(2.0 \times 10^{-3})</td>
</tr>
</tbody>
</table>

for the eight situations considered. In general, all things being equal, as the amount of genetic information increased, the number of samples needed decreased. For example, situations 1 and 2 differ only in the heritability of the traits modeled. Situation 2, which corresponds to a highly heritable trait, needed a smaller number of samples compared with situation 1, which corresponds to a lowly heritable trait. For a highly heritable trait, the distribution of the genotypic values given the phenotypes is narrow. As a result a small number of samples was needed to obtain accurate estimates for the conditional mean of the genotypic values given the phenotypes. To reach the same level of accuracy for a lowly heritable trait, however, a larger number of samples was needed, because now the distribution of the genotypic values given the phenotypes is more dispersed. Situations 3 and 4, however, contradicted this pattern. Situation 4, which corresponds to a highly heritable trait, needed a larger number of samples compared with situation 3, which corresponds to a lowly heritable trait. For these two situations, however, a two-locus model was used. The high number of samples needed in situation 4 indicated the presence of a mixing problem. This type of behavior has been reported when sampling tightly linked loci, and has been referred to as horizontal dependence [46]. Although in this paper the
trait loci were unlinked, horizontal dependence was generated through the penetrance function when sampling one locus at a time and heritability was high. Consider, for example, the genotypes $[0 \ 1]$ and $[1 \ 0]$. If the two loci that form each genotype vector have equal gene frequencies and genotypic effects, the two genotypes will have equal genotypic values. As a result, these two genotypes should be sampled in equal proportions given the data. When sampling genotypes one locus at a time, however, it is not possible to move from $g^t_i = [0 \ 1]$ to $g^{t+k}_{i} = [1 \ 0]$ in one step (i.e., $k=1$). An intermediate step through either genotype $g^{t+k'}_{i} = [0 \ 0]$ or genotype $g^{t+k'}_{i} = [1 \ 1]$, where $k' < k$, needs to occur first. The genotypic values of $[0 \ 0]$ and $[1 \ 1]$ are different from the genotypic value of $[0 \ 1]$ and $[1 \ 0]$. For a trait with low heritability the penetrance function is dispersed, that generates overlaps for different genotypic values. This makes the required intermediate move from $[0 \ 1]$ to $[0 \ 0]$ or $[1 \ 1]$ more likely.

The difference in the number of samples needed in situation 1 versus situation 3, or 7 versus 8 (Tab. 3.5), emphasizes a second effect caused by the increase in the number of loci in the model. As the number of loci increased, the number of samples needed to reach the same level of accuracy increased as well because of the larger number of genotype probabilities that needed to be estimated. For practical purposes, however, the loss in accuracy due to horizontal dependence and the number of genotype probabilities to be estimated was negligible, because ESIP reached a high level of accuracy very fast.

### 3.3.3.3 Blocking Gibbs

Table 3.6 summarizes the scaled absolute difference between estimates of genetic values obtained using blocking Gibbs at the end of the allocated computing time and estimates obtained using the Elston-Stewart algorithm. For the eight situations considered, blocking Gibbs (Tab. 3.6) yielded estimates that were between two and ten times less accurate than estimates obtained by ESIP (Tab. 3.5). For the level of accuracy reached by blocking Gibbs, the estimates for some of the parental individuals in situations 1 to 5 were about 0.1 genetic
Table 3.6 Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using blocking Gibbs compared to estimates calculated using the Elston-Stewart algorithm. For all situations, 5000 initial samples were discarded as burnin. In situations 1 and 2, the genetic standard deviation was 2.8284 (see Tab. 3.1). In situations 3, 4, 5, 6 and 7, the genetic standard deviation was 4.0 (see Tab. 3.1). In situation 8 the genetic standard deviation was 4.9 (see Tab. 3.1).

<table>
<thead>
<tr>
<th>Situation</th>
<th>No. samples included</th>
<th>Maximum</th>
<th>Mean</th>
<th>√MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>520 000</td>
<td>9.7 x 10^{-2}</td>
<td>1.2 x 10^{-2}</td>
<td>1.9 x 10^{-2}</td>
</tr>
<tr>
<td>2</td>
<td>20 000</td>
<td>1.1 x 10^{-1}</td>
<td>3.8 x 10^{-3}</td>
<td>1.4 x 10^{-2}</td>
</tr>
<tr>
<td>3</td>
<td>1 360 000</td>
<td>1.0 x 10^{-1}</td>
<td>1.5 x 10^{-2}</td>
<td>2.2 x 10^{-2}</td>
</tr>
<tr>
<td>4</td>
<td>1 745 000</td>
<td>7.6 x 10^{-2}</td>
<td>2.3 x 10^{-3}</td>
<td>8.8 x 10^{-3}</td>
</tr>
<tr>
<td>5</td>
<td>1 075 000</td>
<td>7.5 x 10^{-2}</td>
<td>8.8 x 10^{-3}</td>
<td>1.3 x 10^{-2}</td>
</tr>
<tr>
<td>6</td>
<td>2 270 000</td>
<td>4.4 x 10^{-2}</td>
<td>9.0 x 10^{-3}</td>
<td>1.3 x 10^{-2}</td>
</tr>
<tr>
<td>7</td>
<td>525 000</td>
<td>3.8 x 10^{-2}</td>
<td>9.7 x 10^{-3}</td>
<td>1.3 x 10^{-2}</td>
</tr>
<tr>
<td>8</td>
<td>685 000</td>
<td>2.1 x 10^{-2}</td>
<td>5.9 x 10^{-3}</td>
<td>7.4 x 10^{-3}</td>
</tr>
</tbody>
</table>

standard deviations away from the estimates obtained by the Elston-Stewart algorithm.

3.3.3.4 Scalar Gibbs

Table 3.7 summarizes the scaled absolute difference between estimates of the genetic values obtained using scalar Gibbs at the end of the allocated computing time and estimates obtained using the Elston-Stewart algorithm. For situation 1, scalar Gibbs had almost the same accuracy as blocking Gibbs (Tab. 3.6) and was approximately ten times less accurate than ESIP (Tab. 3.5). For situation 2, scalar Gibbs exhibited poor mixing, some of the estimates being 2.63 genetic standard deviations away from the estimates obtained by the Elston-Stewart algorithm. Note that the only difference between situations 1 and 2 was the heritability of the trait. The low heritability in situation 1 helped overcome the mixing problem due to the vertical dependence between parents and offspring. The results for situations 3 and 4 were very similar to those obtained with blocking Gibbs (Tab. 3.6). The mixing problem observed in situation 2 disappeared in situation 4, where a two-locus model was used. In this case,
Table 3.7  Maximum and mean of the absolute errors and square root of the mean square error (all in genetic standard deviation units) of estimates of genotypic values calculated using blocking Gibbs compared to estimates calculated using the Elston-Stewart algorithm. For all situations, 5000 initial samples where discarded as burnin. In situations 1 and 2, the genetic standard deviation was 2.8284 (see Tab. 3.1). In situations 3, 4, 5, 6 and 7, the genetic standard deviation was 4.0 (see Tab. 3.1). In situation 8 the genetic standard deviation was 4.9 (see Tab. 3.1).

<table>
<thead>
<tr>
<th>Situation</th>
<th>No. samples included</th>
<th>Maximum</th>
<th>Mean</th>
<th>√MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 045 000</td>
<td>9.8 x 10^{-2}</td>
<td>1.8 x 10^{-2}</td>
<td>2.6 x 10^{-2}</td>
</tr>
<tr>
<td>2</td>
<td>44 000</td>
<td>2.63</td>
<td>3.4 x 10^{-1}</td>
<td>9.3 x 10^{-1}</td>
</tr>
<tr>
<td>3</td>
<td>2 725 000</td>
<td>4.7 x 10^{-2}</td>
<td>1.1 x 10^{-2}</td>
<td>1.4 x 10^{-2}</td>
</tr>
<tr>
<td>4</td>
<td>3 495 000</td>
<td>5.1 x 10^{-2}</td>
<td>2.4 x 10^{-3}</td>
<td>7.0 x 10^{-3}</td>
</tr>
<tr>
<td>5</td>
<td>2 155 000</td>
<td>7.6 x 10^{-2}</td>
<td>1.7 x 10^{-2}</td>
<td>2.3 x 10^{-2}</td>
</tr>
<tr>
<td>6</td>
<td>4 545 000</td>
<td>1.1 x 10^{-1}</td>
<td>2.4 x 10^{-2}</td>
<td>3.1 x 10^{-2}</td>
</tr>
<tr>
<td>7</td>
<td>1 045 000</td>
<td>2.23</td>
<td>7.4 x 10^{-1}</td>
<td>8.5 x 10^{-1}</td>
</tr>
<tr>
<td>8</td>
<td>1 605 000</td>
<td>2.02</td>
<td>8.5 x 10^{-1}</td>
<td>9.6 x 10^{-1}</td>
</tr>
</tbody>
</table>

the benefit of breaking the vertical dependence by increasing the number of loci outweighed the loss in accuracy caused by the introduction of horizontal dependence. For situation 5, the results were again very similar to those obtained with blocking Gibbs. The extension of the base pedigree in situation 6 increased the vertical dependence between parents and offspring. For this situation, a loss in accuracy was observed when compared with the level of accuracy reached for situation 3.

It has been suggested that for samplers, such as scalar Gibbs, that update one variable at a time, the mixing time tends to be exponential in the number of variables [21]. To investigate the relationship between computing time and pedigree size, the time needed by scalar Gibbs to reach a value of the mean sum of squares of the differences below 0.09 was recorded for situations 3 and 5. For the base pedigree, which contained 96 animals, 1 hour and 20 minutes were needed to reach the desired accuracy. This is 50 seconds per animal. For the extended pedigree, which contained 187 animals, 23 hours and 48 minutes were needed to reach the desired accuracy, or 450 seconds per animal. So, while the increase in the number of animals
in the pedigree was twofold, the increase in computing time per animal was ninefold.

Slow mixing was also evident for situations 7 and 8, situations in which the strong vertical dependence between parents and offspring is obvious. Note that for the poultry pedigree, neither low heritability nor an increase in the number of loci (two and three, respectively) could alleviate the mixing problem generated by the vertical dependence between parents and offspring (Tab. 3.7).

3.3.4 Accuracy levels reached by the four methods under investigation

Figure 3.4 summarizes the level of accuracy reached by each of the four methods under investigation at the end of the allocated computing time, in terms of the mean of the scaled absolute differences. For situations 2, 7, and 8, the mean obtained using scalar Gibbs was not represented in Figure 3.4; because of slow mixing, values were outside the range of the Y axis (0.34, 0.74, and 0.85 genetic standard deviations).

3.3.5 Implementation of ESIP

The results presented so far for ESIP were obtained by using only the Elston-Stewart algorithm. Thus, all proposed samples were accepted. The Elston-Stewart algorithm can be used as long as the cutset size is not too large for efficient computations. Once the cutset size becomes too large, iterative peeling is used and the proposed samples come from a modified pedigree. As a result, some of the proposed samples will be rejected. For situation 3, Table 3.8 contrasts the behavior of ESIP in the case when only the Elston-Stewart algorithm was used versus the case when both the Elston-Stewart algorithm and iterative peeling were used. For each case, 50,000 samples were averaged to calculate the statistics used in the comparison. For the pedigree used in situation 3, the maximum size of a cutset generated by the Elston-Stewart algorithm was five. As a result, the Elston-Stewart algorithm could be used for all individuals in the pedigree. When the maximum cutset size was set to four, iterative peeling was used for six of the 96 individuals. Finally, when the maximum cutset size was set to three, iterative peeling was used for nine of the 96 individuals.
Figure 3.4  Mean of the scaled absolute differences between estimates obtained using each of the four methods under investigation and estimates obtained using the Elston-Stewart algorithm for each of the eight situations. × Iterative peeling; □ ESIP; + Blocking Gibbs; ∆ Scalar Gibbs. For scalar Gibbs, the mean was outside the range of the Y axis for situations 2, 7 and 8.

Table 3.8 Differences between ESIP estimates, calculated based on 50000 samples, for different sizes of the maximum cutset size. The genetic standard deviation was 4.0 (see Tab. 3.1). The computer used was a Pentium Pro-333.

<table>
<thead>
<tr>
<th>Cutset size</th>
<th>Computing time (sec)</th>
<th>Rejection Rate</th>
<th>Maximum</th>
<th>Mean</th>
<th>(\sqrt{\text{MSE}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3596</td>
<td>0.00</td>
<td>(1.1 \times 10^{-2})</td>
<td>(3.2 \times 10^{-3})</td>
<td>(4.0 \times 10^{-3})</td>
</tr>
<tr>
<td>4</td>
<td>6168</td>
<td>0.05</td>
<td>(1.0 \times 10^{-2})</td>
<td>(3.6 \times 10^{-3})</td>
<td>(4.4 \times 10^{-3})</td>
</tr>
<tr>
<td>3</td>
<td>8944</td>
<td>0.12</td>
<td>(1.6 \times 10^{-2})</td>
<td>(4.3 \times 10^{-3})</td>
<td>(5.4 \times 10^{-3})</td>
</tr>
</tbody>
</table>
The computing time needed to generate 50,000 samples increased as the number of individuals to be iteratively peeled increased (Tab. 3.8). As explained by Fernandez et al. [13], however, this problem can be eliminated by explicitly cutting loops. Once the loops are cut, samples can be obtained by applying the Elston-Stewart algorithm [12]. By using this approach, sampling time should not be affected by the number of individuals that are iteratively peeled.

From Table 3.8 we can see that the accuracy of the estimates decreased with increasing number of individuals that must be iteratively peeled. Even by using only 50,000 samples and iteratively peeling some members of the pedigree, however, ESIP yielded more accurate estimates than scalar Gibbs did based on 2,750,000 samples (Tab. 3.7) or blocking Gibbs did based on 1,360,000 samples (Tab. 3.6).

3.4 Discussion

This paper investigates the performance of four methods used to compute conditional means of genotypic values for a quantitative trait based on phenotypic data using finite locus models. Simulated data were used to assess the performance of iterative peeling, scalar Gibbs, blocking Gibbs, and ESIP.

Iterative peeling yielded exact results for pedigrees without loops regardless of the number of loci considered. For a pedigree with loops, iterative peeling yielded good approximations of the conditional means for one-locus and two-locus models. As the number of loops in the pedigree increased, the accuracy of the estimates obtained using iterative peeling decreased. When loops were present, missing data and low heritability further reduced the level of accuracy. Iterative peeling has a serious limitation due to the exponential relationship between computing time and the number of loci in the model. On the other hand, a linear relationship between computing efficiency and the number of loci can be maintained for MCMC methods by sampling one locus at a time.

Out of the three MCMC methods considered, ESIP performed best. With the exception of one situation, the behavior of ESIP was similar for all situations considered. During the course
of a run, ESIP reached a high level of accuracy rapidly, and the subsequent improvement in accuracy was almost negligible. This was, however, not true for a highly heritable trait in a two-locus model. In this situation, ESIP was consistently more accurate than the other two samplers only on the last third of the run. For this situation, however, all three MCMC methods reached a high level of accuracy in a short period of time.

Iterative peeling was the only method to yield more accurate estimates than ESIP for three of the eight situations considered. These correspond to the situations where iterative peeling gives exact estimates due to absence of loops in the poultry pedigree and for a highly heritable trait with a one-locus model for the base pedigree.

Neither blocking nor scalar Gibbs reached the desired level of accuracy in the allocated computing time in any of the eight situations considered. Out of the three MCMC methods considered, scalar Gibbs had the poorest performance overall. For the base pedigree (situation 2), and for the poultry pedigree (situations 7 and 8), this method showed poor mixing due to vertical dependence between parents and offspring. For the base pedigree, however, when either heritability was low or the number of loci was greater than one, the mixing problem was alleviated. For the poultry pedigree, neither low heritability nor a larger number of loci in the model improved mixing.

The results presented in this paper suggest that genetic evaluation, based on conditional means of genotypic values given the phenotypes, is feasible under a finite locus model. ESIP proved to be an efficient as well as a reliable method for computing conditional means under such a model. This approach could provide a solution to the problem of genetic evaluation of traits that show non-additive inheritance, even in crossbred populations.

Acknowledgements

This work was partially funded by award no. 2002-35205-1156 of the National Research Initiative Competitive Grants Program of the USDA.
Bibliography


CHAPTER 4. THE EFFECT OF THE NUMBER OF LOCI IN FINITE LOCUS MODELS FOR GENETIC EVALUATION

Abstract

For a finite locus model, Markov Chain Monte Carlo (MCMC) methods can be used to estimate the conditional mean of genotypic values given phenotypes, which is also known as the best predictor (BP). Successful application of MCMC methods for genetic evaluation using finite locus models depends, among other factors, on the number of loci assumed in the model. The effect of the assumed number of loci on accuracy of genetic evaluations was investigated using data simulated with 100 loci. For several small pedigrees, genetic evaluations obtained by best linear prediction (BLP) were compared to genetic evaluations obtained by BP. For BLP evaluation, only the first and second moments of the joint distribution of the genotypic and phenotypic values must be known. These moments were calculated from the gene frequencies and genotypic effects used in the simulation model. For BP evaluation the complete distribution must be known. The gene frequencies and genotypic effects of each model used for BP evaluation were derived such that the genotypic mean, the additive variance, and the dominance variance were the same as in the simulation model. These gene frequencies and genotypic effects completely specify the distribution of the genotypic values needed for the BP evaluation. For traits with low heritability, models with up to three loci yielded very accurate genetic evaluations by BP for both purebred and crossbred data. For traits with high heritability, models with up to six loci yielded accurate genetic evaluations by BP for both purebred and crossbred data.
4.1 Introduction

One of the most important objectives of animal breeding is the identification of the “best” animals for breeding. To be able to select the parents of the next generation, animal breeders need to assess the genotypic value of candidate animals. However, genotypic values cannot be observed and consequently they need to be predicted from the available information.

Best linear unbiased prediction (BLUP), which can be obtained efficiently by solving Henderson’s mixed model equations (HMME) [20], is currently the most widely used method for genetic evaluation. When only phenotypic and pedigree information is available, one of the requirements for building HMME is to calculate the inverse of the variance covariance matrix of the genotypic values. Under additive inheritance, efficient algorithms to calculate the required inverse have been developed for both purebred [18, 19, 26, 25] and crossbred [8, 22] populations. Under non-additive inheritance, algorithms to calculate the required inverse have been investigated as well [21, 27, 31], but these algorithms are not feasible for large inbred populations [5]. This is especially true for crossbred populations [23]. However some traits of interest, for example reproductive or disease resistance traits, are known to have low heritability. Lowly heritable traits suggest non-additive gene action. Also, the breeding strategies used in several livestock species result in crossbred pedigrees. Thus, efficient methods for genetic evaluation under non-additive inheritance for purebred and especially for crossbred populations must be developed.

Finite locus models can easily accommodate non-additive inheritance as well as crossbred data. The use of the conditional mean of genotypic values given phenotypes, calculated under the assumption of a finite locus model, has been suggested as an alternative to BLUP [12, 15, 28]. The conditional mean is also known as the best predictor (BP) because it minimizes the mean square error of prediction and maximizes the mean of the selected candidates [2, 11]. Given a finite locus model, the BP can be calculated exactly using Elston Stewart type algorithms [7], approximated using iterative peeling [30] or it can be estimated using Markov Chain Monte Carlo (MCMC) methods [12, 15, 28]. The computational efficiency of these methods is directly related to the number of loci considered in the finite locus model. For
peeling type algorithms, this relationship is exponential whereas for MCMC methods a linear relationship can be maintained by sampling genotypes one locus at a time.

The exact number of quantitative trait loci (QTL) responsible for the genetic variation of a quantitative trait is not known. However, after performing a meta-analysis on published results from various QTL mapping experiments, Hayes and Goddard estimate that between 50 and 100 loci are segregating in dairy cattle and swine populations [17]. For the large pedigrees encountered in real livestock populations, genetic evaluation by BP using a finite locus model with 50 to 100 loci is computationally unfeasible. Therefore, in this paper, the feasibility of genetic evaluation using finite locus models with a small number of loci (FLMS) is investigated by computer simulation.

FLMS with two through six loci were considered for genetic evaluation. To study the accuracy of evaluations using these FLMS, data sets were generated using finite locus models with a large number of loci (FLML). Data were simulated using FLML with about 100 loci. For such data, the distribution of the genotypic values is well approximated by a multivariate normal, and thus, the BP becomes the BLP [20]. The accuracy of BP evaluations under the assumption of these FLMS is assessed by comparison to BLP evaluations.

4.2 Methods

Consider a trait determined by \( N \) segregating quantitative trait loci (QTL) with two alleles at each locus. When only additive and dominance gene action is present, the vector \( \mathbf{u} \) of genotypic values of \( n \) individuals can be modeled as

\[
\mathbf{u} = \mathbf{1} \eta + \sum_{i=1}^{N} \mathbf{u}_i
\]

\[
= \mathbf{1} \eta + \sum_{i=1}^{N} \mathbf{Q}_i \delta_i, \tag{4.1}
\]

where \( \mathbf{1} \) is an \( n \times 1 \) vector of ones; \( \eta \) is the trait mean in a reference breed; \( \mathbf{u}_i \) is the \( n \times 1 \) vector of genotypic values at locus \( i \) in the reference breed; \( \mathbf{Q}_i \) is an \( n \times 3 \) incidence matrix relating the genotypic values at locus \( i \) to the corresponding individuals, with each row of \( \mathbf{Q}_i \).
being one of the vectors \([1 \ 0 \ 0], [0 \ 1 \ 0], \text{or} [0 \ 0 \ 1]\); \(\delta_i\) is an \(3 \times 1\) vector that contains the genotypic effects at locus \(i\): \([a_i \ d_i \ -a_i']\) \([9]\) in the reference breed. The parameters of this model are: \(\eta\), the genotypic effects \(a_i\) and \(d_i\), and gene frequency \(p_i\), for locus \(i = 1, \ldots, N\).

In matrix notation, the vector \(y\) of phenotypic values of \(n\) individuals can be written as a function of the genotypic values as follows

\[
y = X\beta + Zu + e, \quad (4.2)
\]

where \(X\) is the incidence matrix relating the vector \(\beta\) of fixed effects to \(y\); \(Z\) is the incidence matrix relating \(u\) to \(y\); \(u\) is the vector of genotypic values from 4.1 and \(e\) is the vector of residuals \(\sim N(0, I\sigma_e^2)\).

Consider first the situation where \(u\) is modeled using a large number of loci each with a small effect. Under such a model, the distribution of genotypic values is approximately multivariate normal. As a result, we can assume that \(u\) and \(y\) are approximately multivariate normal,

\[
\begin{bmatrix} u \\ y \end{bmatrix} \sim N \begin{pmatrix} \mu_u \\ \mu_y \end{pmatrix}, \begin{bmatrix} G & C \\ C' & V \end{bmatrix} \quad (4.3)
\]

where \(\mu_u\) is the vector of genotypic means; \(\mu_y = X\beta\); \(G\) is the genotypic variance covariance matrix; \(C = GZ'\) is the covariance matrix between \(u\) and \(y'\); \(V = ZGZ' + I\sigma_e^2\) is the variance covariance matrix of \(y\). Under multivariate normality the conditional mean is also the BLP and can be written as

\[
E(u \mid y) = \mu_u + CV^{-1}(y - \mu_y), \quad (4.4)
\]

Note that BLP is a function of the first and second moments of the genotypic values and the phenotypes. The theory for modeling genetic means is well known for both purebred and crossbred populations \([4, 6]\). The theory for modelling the genetic covariances is also known for both purebred \([14, 16]\) and crossbred \([23]\) populations. However, the covariance theory for
crossbred populations is more complex. For example, in a non-inbred, unselected, purebred population, if we ignore linkage and if only additive and dominance gene action are considered, the genetic variance covariance matrix can be written as

\[ G = A \sigma_a^2 + D \sigma_d^2 \]  

(4.5)

where \( A \) is the additive relationship matrix, \( \sigma_a^2 \) is the additive variance, \( D \) is the dominance relationship matrix and \( \sigma_d^2 \) is the dominance variance. However, for example, in a two breed situation where inbreeding is present the genetic variance covariance matrix becomes,

\[ G = \sum_{q=1}^{25} C_q \theta_q \]  

(4.6)

where \( \theta_q \) is the dispersion parameter corresponding to one of 25 breed-specific identity states and \( C_q \) is the matrix of coefficients for \( \theta_q \). Recursive formulae are available to compute the elements of \( C_q \) [23]. In the absence of inbreeding, the number of dispersion parameters is reduced from 25 to 12. Thus, for small pedigrees given known parameters, BLP’s can be obtained for both purebred and crossbred populations. For large pedigrees, under non additive inheritance, BLP’s cannot be obtained for either purebred or crossbred populations because efficient algorithms to invert \( G \) are not available.

Consider now the situation where it is modeled using a limited number of loci. In this situation, BP can be calculated by summing over all possible genotype configurations as follows

\[
E(u \mid y) = 1\eta + \sum_{g} u_g \Pr(g \mid y),
\]

(4.7)

where \( u_g \) is the vector of of genotypic values that corresponds to the genotype configuration \( g \), and

\[
\Pr(g \mid y) = \frac{\Pr(g, y)}{\Pr(y)} \propto \Pr(y \mid g) \Pr(g).
\]

(4.8)

where \( \Pr(y \mid g) \) represents the conditional probability of the phenotypes given genotype con-
configuration $g$ and $\Pr(g)$ represents the probability of the genotype configuration $g$. Under a finite locus model, efficient methods to calculate these probabilities are available [1, 7]. From equation (5.1), it can be seen that the key aspect of this type of genetic evaluation is the correct and efficient computation of the sum over all possible genotype configurations. This sum can be calculated exactly using the Elston-Stewart algorithm. This algorithm, however, is computationally feasible only for simple pedigrees and models with up to about three loci. For complex pedigrees and models with more than three loci, MCMC methods hold most promise for the efficient calculation of the desired sum [29]. In this paper, BP evaluations were calculated using the Elston-Stewart algorithm whenever it was computationally feasible. When the use of the Elston-Stewart algorithm was not feasible, BP evaluations were obtained by using an MCMC method called ESIP. ESIP combines the Elston Stewart algorithm with iterative peeling to generate joint samples from the entire pedigree one locus at a time [10, 29]. In a previous study [29] we have investigated the performance of ESIP when used for genetic evaluation by BP. From the results of that study, it was determined that 50,000 samples from ESIP are sufficient to estimate the BP accurately.

The first and second moments needed for genetic evaluation by BLP, were calculated from the gene frequencies and genotypic effects of the FLML used to simulate the data. In contrast, for genetic evaluation by BP, the gene frequencies and genotypic effects of the FLMS were chosen, as described below, such that they yielded the same genotypic mean and the same additive and dominance variances as the FLML that was used for simulation. For convenience, we define an $N_1$ locus model to be “equivalent” to an $N_2$ locus model ($N_2 > N_1$) if the genotypic means, the additive variances and the dominance variances of the two models are identical.

Consider first the case of purebred data. Further consider the simple situation when the gene frequency and the additive effect at all loci of a given model are equal. For this case, we discuss below how to assign values to the gene frequencies and the genotypic effects for the FLMS with $N_1$ loci and the FLML with $N_2$ loci so that they are “equivalent”.

Note that, for an arbitrary model with an even number of loci $N$, the genotypic mean ($\eta$),
additive variance ($\sigma_a^2$) and dominance variance ($\sigma_d^2$) can be written as

$$
\eta = 2na(p - q) + 2npqd_1 + 2npqd_2
$$

$$
\sigma_a^2 = 2npq[a + d_1(q - p)]^2 + 2npq[a + d_2(q - p)]^2 \quad (4.9)
$$

$$
\sigma_d^2 = n(2pqd_1)^2 + n(2pqd_2)^2.
$$

where $n = \frac{N}{2}$; $a$ is the genotypic effect of one of the homozygotes at the $N$ loci; $p$ is the frequency of the first allele at the $N$ loci; $q = 1 - p$; $d_1$ is the genotypic effect of the heterozygote at half of the $N$ loci and $d_2$ the genotypic effect of the heterozygote at the other half of the $N$ loci. We simplify further by setting the inbreeding depression ($ID = 2npqd_1 + 2npqd_2$) equal to zero. As a result, $d_1$ is equal to $-d_2$. Note that in this case, the inbreeding depression is zero while the dominance variance is nonzero. After some algebra, making use of the fact that $q = 1 - p$ and $d_1 = -d_2$, the system of equations (4.9) yields

$$
p = \frac{\eta + 2na}{4na}
$$

$$
0 = 16a^4n^4 - a^2(8n^2\eta^2 + 16n^3\sigma_a^2) + \eta^4 + 8n\sigma_d^2\eta^2 + 4n\eta^2\sigma_a^2 \quad (4.10)
$$

$$
d_1 = \frac{\sqrt{\sigma_d^2}}{2p(1 - p)\sqrt{2n}}.
$$

The second equation in the (4.10) can be solved for $a$ in terms of $n, \eta, \sigma_a^2$ and $\sigma_d^2$. Next, by substituting the value obtained for $a$ in the first equation we can obtain $p$ in terms of $n, \eta, \sigma_a^2$ and $\sigma_d^2$, and then by substituting $p$ in the third equation we can obtain $d_1$ in terms of $n, \eta, \sigma_a^2$ and $\sigma_d^2$. Thus, the gene frequencies and genotypic effects of a model with $N$ loci are completely determined by the genotypic mean and the additive and the dominance variance.

Now consider the two models of interest, a FLMS with $N_1$ loci, and a FLML with $N_2$ loci. Under the assumptions described above, the gene frequencies and genotypic effects for each of the two models can be obtained by solving the system of equations given in (4.10) with $n = \frac{N_1}{2}$ and $n = \frac{N_2}{2}$ respectively, given the assigned values for $\eta, \sigma_a^2$ and $\sigma_d^2$. For models with an uneven number of loci, the heterozygote at the extra locus is assumed to have a genotypic
effect equal to zero.

Consider now the situation of crossbred data. For the purpose of this paper, crossbred data are simulated by adding \( k \) extra loci to the purebred FLML. Thus, crossbred data are simulated with a FLML with \( N_2 + k \) loci, where the \( N_2 \) loci have the same gene frequency in all breeds and the \( k \) loci have different gene frequencies for different breeds. The values for the gene frequencies and genotypic effects for a FLMS with \( N_1 + k \) loci are determined, so that it is "equivalent" to the FLML with \( N_2 + k \) loci, as follows. First, the FLMS and the FLML are made "equivalent" with respect to \( N_1 \) and \( N_2 \) loci under a purebred setting. Next, the same gene frequencies and genotypic effects are used for the \( k \) extra loci in both models.

For the case of purebred data, three hypothetical pedigrees were used to investigate the effect of the number of loci on the accuracy of genetic evaluations by BP. The first hypothetical pedigree, shown in Figure 4.1, has 14 individuals, no loops and will be referred to as the simple pedigree. The second pedigree, shown in Figure 4.4 in the Appendix, was obtained by extending the first pedigree for five more generations. This pedigree of 44 individuals has eight generations, no loops and will be referred to as the extended pedigree. The third pedigree, shown in Figure 4.5 in the Appendix, is a highly inbred pedigree with many loops. This pedigree of 34 individuals has eight generations, several loops generated by repeated half sib matings and will be referred to as the inbred pedigree.

Figure 4.1 Simple purebred pedigree. Genetic evaluations were obtained for individuals marked by *.
Purebred data were simulated using a FLML with 100 loci. At each of the 100 loci, the gene frequency was $p^* = 0.5$ and the additive effect was $a^* = 0.2828$. Of the 100 loci, at each of 50, the dominance effect was $d^* = 0.2828$, and at each of the remaining 50, the dominance effect was $d^* = -0.2828$. These values yield $\eta = 0$, $\sigma_a^2 = 4$ and $\sigma_d^2 = 2$. Two values were used for the error variance: $\sigma_a^2 = 4$ and $\sigma_d^2 = 34$, which combined with the genetic parameters yield two levels of heritability. In order to examine the effect of pedigree structure, missing data, and genetic parameters on the accuracy of genetic evaluations for various FLMS, nine situations were simulated for the purebred case (Table 4.1). The first four situations cover all possible combinations of two heritabilities (0.1 and 0.4) and two types of non-inbred pedigrees (simple and extended). This design allows us to examine the main effects of heritability and pedigree size as well as the interactions between these two factors. Situations 3, 4, 5, 6, 7, 8 cover all possible combinations of two heritabilities (0.1 and 0.4) and three patterns of missing data: all individuals have phenotypic data; all individuals in the first two generations have missing data (10 individuals); all sires in the pedigree have missing data (19 individuals). This design allows us to examine the main effects of heritability and missing data as well as the possible interactions between these two factors. Situation 9, which differs from situations 1 and 3 only in the pedigree type, is considered to examine the effect of the presence of inbreeding.

Table 4.1 Situations simulated for the purebred case for three different pedigrees. $h_n^2$ denotes the narrow sense heritability; $h_b^2$ denotes the broad sense heritability:

<table>
<thead>
<tr>
<th>Situation</th>
<th>Pedigree</th>
<th>No. missing</th>
<th>$h_n^2$</th>
<th>$h_b^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>simple</td>
<td>0</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>simple</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>extended</td>
<td>0</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>extended</td>
<td>0</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>extended</td>
<td>10</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>6</td>
<td>extended</td>
<td>10</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>extended</td>
<td>19</td>
<td>0.1</td>
<td>0.15</td>
</tr>
<tr>
<td>8</td>
<td>extended</td>
<td>19</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>9</td>
<td>inbred</td>
<td>0</td>
<td>0.1</td>
<td>0.15</td>
</tr>
</tbody>
</table>
The parameters of the FLMS used to calculate BP's for the data generated according to the nine situations described above, are given in Table 4.2. Note that FLM($N_1$) denotes the FLMS with $N_1$ loci, and that each of the FLMS in Table 4.2 yields $\eta = 0$, $\sigma_a^2 = 4$ and $\sigma_d^2 = 2$.

Table 4.2 Parameters for the FLMS used to analyze purebred data. The second column contains the number of loci in the respective FLMS; $a$ denotes the additive effect at all loci; $d_1$ the dominance effect at half of the loci; $d_2$ the dominance effect at the other half of the loci; $p$ the gene frequency at each locus. * Here the dominance effect of the third locus was set to 0.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>No. loci</th>
<th>$a$</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>-2</td>
<td>0.5</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>3*</td>
<td>1.63</td>
<td>2</td>
<td>-2</td>
<td>0.5</td>
</tr>
<tr>
<td>FLM(4)</td>
<td>4</td>
<td>1.4142</td>
<td>1.4142</td>
<td>-1.4142</td>
<td>0.5</td>
</tr>
<tr>
<td>FLM(6)</td>
<td>6</td>
<td>1.1547</td>
<td>1.1547</td>
<td>-1.1547</td>
<td>0.5</td>
</tr>
</tbody>
</table>

For the case of crossbred data, two hypothetical pedigrees were used to investigate the effect of the number of loci on the accuracy of genetic evaluations by BP. The first pedigree is shown in Figure 4.2. This corresponds to a two-breed pedigree with 14 individuals, and no loops. The second pedigree is also a two-breed pedigree obtained by extending the first pedigree for five more generations. This extension is done in the same way as in the purebred case, but starting with generation three, sires from alternate breeds are used in alternate generations. Thus, an extended two-breed pedigree with 44 individuals and no loops was generated.

Two-breed data were simulated using a FLML with $100 + 1$ loci. The gene frequency and genotypic effects for the first 100 loci in both breeds were assigned the same values as the ones used for the purebred case. For breed $A$, the extra locus had a gene frequency $p_A = 0.9$, while for breed $B$ the extra locus had a gene frequency $p_B = 0.1$. The genotypic effects for the extra locus in both breeds were: $a = 2$ and $d_1 = 0$. These values yield $\eta_A = 1.6$, $\eta_B = -1.6$, $\sigma_{a_A}^2 = \sigma_{d_B}^2 = 4.72$ and $\sigma_{d_A}^2 = \sigma_{d_B}^2 = 2$. Two values were used for the error variance: $\sigma_e^2 = 5.08$ or $\sigma_e^2 = 40.48$, which combined with the genetic parameters yield two levels of heritability. In order to examine the effect of pedigree structure and genetic parameters on the accuracy of
genetic evaluations obtained with various FLMS, four situations were simulated for the two-breed case (Table 4.3). No missing data were present in these four situations. The design of the simulation allows us to examine the main effects of heritability and pedigree size as well as the interactions between these two factors. Also, it allows us to compare the effect of the number of loci on genetic evaluations by BP in crossbred versus purebred situations.

Table 4.3 Situations simulated for the two-breed case for two pedigrees.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Pedigree</th>
<th>$h_n^2$</th>
<th>$h_b^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>simple</td>
<td>0.1</td>
<td>0.142</td>
</tr>
<tr>
<td>2</td>
<td>simple</td>
<td>0.4</td>
<td>0.57</td>
</tr>
<tr>
<td>3</td>
<td>extended</td>
<td>0.1</td>
<td>0.142</td>
</tr>
<tr>
<td>4</td>
<td>extended</td>
<td>0.4</td>
<td>0.57</td>
</tr>
</tbody>
</table>

In the following, $\text{FLM}(N_1,k)$ denotes the FLMS with $N_1 + k$ loci, where $N_1$ are the loci that have the same gene frequencies in both breeds and $k$ are the loci that have different gene frequencies in the two breeds. For the crossbred model $\text{FLM}(N_1,k)$, the gene frequencies and genotypic effects for the $N_1$ loci are identical to those in purebred model $\text{FLM}(N_1)$. For the extra $k = 1$ locus, the gene frequencies and genotypic effects used in the simulation were
assigned.

For each of the pure bred and cross bred situations considered, 100 replicates of the data were generated and genetic evaluations by BLP and BP were calculated for each of four animals in the last generation. For each of these data sets, BP evaluations were obtained under one or more FLMS. For traits with low heritability, BP evaluations were obtained under FLM(2) and FLM(3) for purebred data and FLM(2,1) for crossbred data. For traits with high heritability, BP evaluations were obtained under FLM(2), FLM(3), FLM(4) and FLM(6) for purebred data and FLM(2,1), FLM(3,1) and FLM(4,1) for crossbred data. In each replicate, for each of the four animals, the absolute difference between BLP and BP evaluations was calculated and then scaled by the genetic standard deviation. Thus, 400 scaled absolute differences were obtained for each analysis. Because full sibs have the same evaluations, however, only 200 of these values are unique. For each of the FLMS, the maximum, mean and standard deviation of these 200 scaled differences were used to quantify the accuracy of genetic evaluations by BP.

4.3 Results

Prior to discussing any results, it is important to note that the Elston-Stewart algorithm was used to calculate BP evaluations only for the simple pedigree under FLM(2) and FLM(3) and for the extended pedigree under FLM(2). For all other situations, ESIP was used to calculate BP evaluations. While the Elston-Stewart algorithm yields the exact value of the BP evaluations, ESIP yields estimates of the BP evaluations. Table 4.4 shows the difference between the summary statistics of evaluations obtained by the Elston-Stewart algorithm and evaluations obtained by ESIP, under FLM(3) for situation 7 of the purebred case.

To obtain the exact evaluations presented in Table 4.4, the Elston-Stewart algorithm program was run for six days and 17 hours. On the same computer, the ESIP evaluations were obtained in 13 hours. Note that, the evaluations presented Table 4.4 are the summary of 100 replicates. For each replicate, one hour and 37 minutes was needed to obtain exact evaluations using the Elston-Stewart algorithm. The time needed by ESIP to generate the 50,000 samples to be used to estimate BP evaluations for one replicate was about eight minutes. The BP
Table 4.4 Summary statistics of 200 values from 100 replicates of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(3) for purebred situation 7, calculated using the Elston-Stewart algorithm and ESIP. For ESIP, each evaluation was obtained by averaging 50,000 samples.

<table>
<thead>
<tr>
<th>Method</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elston-Stewart</td>
<td>$1.1 \times 10^{-2}$</td>
<td>$1.7 \times 10^{-3}$</td>
<td>$1.8 \times 10^{-3}$</td>
</tr>
<tr>
<td>ESIP</td>
<td>$2.1 \times 10^{-2}$</td>
<td>$5.5 \times 10^{-3}$</td>
<td>$4.1 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

evaluations estimated based on 50,000 samples generated by ESIP, however, were very close to the exact values obtained by the Elston-Stewart algorithm. The computing time needed by the Elston-Stewart algorithm restricts its use to simple models of the type mentioned above. In terms of accuracy, the differences between exact evaluations and ESIP evaluations were small. The existence of these differences, however, needs to be taken into account when making inferences about the accuracy reached by different FLMS.

4.3.1 Purebred case

For situation 1, Table 4.5 summarizes the level of accuracy reached by BP evaluations under FLM(2) and FLM(3). Note that a high level of accuracy was already reached under FLM(2). FLM(3) further increased the accuracy of the evaluations. In this situation, a two locus model yielded acceptable evaluations for data generated with a FLML with 100 loci.

Table 4.5 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2) and FLM(3), for purebred situation 1. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$8.6 \times 10^{-2}$</td>
<td>$8.5 \times 10^{-3}$</td>
<td>$1.2 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$6.0 \times 10^{-2}$</td>
<td>$5.4 \times 10^{-3}$</td>
<td>$7.8 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

For situation 2, Table 4.6 summarizes the level of accuracy reached by BP under FLM(2), FLM(3), FLM(4) and FLM(6). For this situation, which differs from situation 1 only in the
magnitude of the heritability, a higher number of loci is needed to reach a good level of accuracy.

Table 4.6  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 2. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$4.6 \times 10^{-1}$</td>
<td>$1.0 \times 10^{-1}$</td>
<td>$9.5 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$3.3 \times 10^{-1}$</td>
<td>$5.0 \times 10^{-2}$</td>
<td>$5.7 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(4)</td>
<td>$3.1 \times 10^{-1}$</td>
<td>$3.5 \times 10^{-2}$</td>
<td>$4.3 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(6)</td>
<td>$1.8 \times 10^{-1}$</td>
<td>$2.1 \times 10^{-2}$</td>
<td>$2.7 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

For situation 3, Table 4.7 summarizes the level of accuracy reached by BP evaluations under FLM(2) and FLM(3). For this situation, which differs from situation 1 only in the pedigree type, the level of accuracy reached with the two models is higher than the level reached in situation 1. The increase in accuracy was moderate.

Table 4.7  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2) and FLM(3), for purebred situation 3. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$9.2 \times 10^{-2}$</td>
<td>$7.6 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$1.2 \times 10^{-2}$</td>
<td>$4.2 \times 10^{-3}$</td>
<td>$3.0 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

For situation 4, Table 4.8 summarizes the level of accuracy reached by BP under FLM(2), FLM(3), FLM(4) and FLM(6). For this situation, which differs from situation 2 only in the pedigree type, the level of accuracy reached is moderately higher than the level reached in situation 2 for all models.

The results obtained for the first four situations allow us to assess the effect of heritability and pedigree size on the number of loci needed for accurate BP evaluations. For a lowly heritable trait, modeled with a FLML with 100 loci, FLMS with two to three loci yield acceptable
Table 4.8  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 4. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$4.4 \times 10^{-1}$</td>
<td>$7.3 \times 10^{-2}$</td>
<td>$7.9 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$3.1 \times 10^{-1}$</td>
<td>$3.8 \times 10^{-2}$</td>
<td>$4.6 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(4)</td>
<td>$2.6 \times 10^{-1}$</td>
<td>$2.7 \times 10^{-2}$</td>
<td>$3.4 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(6)</td>
<td>$1.7 \times 10^{-1}$</td>
<td>$1.8 \times 10^{-2}$</td>
<td>$2.3 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

BP evaluations. For traits with high heritability, modeled with a FLML with 100 loci, FLMS with six loci yield acceptable BP evaluations. The accuracy of the BP evaluations was improved when the size of the pedigree increased, regardless of the heritability of the trait. The number of loci needed for accurate evaluations did not increase when a larger pedigrees was considered.

For situation 5, Table 4.9 summarizes the level of accuracy reached by BP evaluations under FLM(2) and FLM(3). For this situation, which differs from situation 3 only in that the individuals in the first two generations have missing phenotypic data, the level of accuracy reached was lower than for situation 3. However, the reduction in accuracy was moderate.

Table 4.9  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), for purebred situation 5. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$8.8 \times 10^{-2}$</td>
<td>$7.6 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$5.3 \times 10^{-2}$</td>
<td>$7.2 \times 10^{-3}$</td>
<td>$7.5 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

For situation 6, Table 4.10 summarizes the level of accuracy reached by BP under FLM(2), FLM(3), FLM(4) and FLM(6). For this situation, which differs from situation 4 only in that the individuals in the first two generations have missing phenotypic data, the level of accuracy reached is almost identical to the level reached in situation 4.
Table 4.10  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 6. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$4.4 \times 10^{-1}$</td>
<td>$7.4 \times 10^{-2}$</td>
<td>$7.9 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$3.1 \times 10^{-1}$</td>
<td>$3.8 \times 10^{-2}$</td>
<td>$4.5 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(4)</td>
<td>$2.6 \times 10^{-1}$</td>
<td>$2.7 \times 10^{-2}$</td>
<td>$3.4 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(6)</td>
<td>$1.7 \times 10^{-1}$</td>
<td>$1.8 \times 10^{-2}$</td>
<td>$2.3 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

For situation 7, Table 4.11 summarizes the level of accuracy reached by BP evaluations under FLM(2), FLM(3). For this situation, which differs from situation 3 and 5 only in that now all sires in the pedigree have missing phenotypic data, the level of accuracy reached was higher than the level reached for both situation 3 and 5.

Table 4.11  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), for purebred situation 7. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$1.8 \times 10^{-2}$</td>
<td>$2.6 \times 10^{-3}$</td>
<td>$3.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$1.1 \times 10^{-2}$</td>
<td>$1.7 \times 10^{-3}$</td>
<td>$1.8 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

For situation 8, Table 4.12 summarizes the level of accuracy reached by BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6). For this situation, which differs from situation 4 and 6 only in that now all sires in the pedigree have missing phenotypic data, the level of accuracy reached was higher than the level reached for both situation 3 and 5.

The results obtained for situations 3, 4, 5, 6, 7 and 8 allow us to assess the effect of heritability and missing data on the number of loci needed for accurate BP evaluations. The effect of heritability in these six situations was the same as the effect observed for the first four situations. Thus the same conclusion holds true, that is: highly heritable traits need to be evaluated using FLMS with a larger number of loci than lowly heritable traits. The two miss-
Table 4.12 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6), for purebred situation 8. The scale used was $\sqrt{6}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2)</td>
<td>$3.8 \times 10^{-1}$</td>
<td>$4.1 \times 10^{-2}$</td>
<td>$3.9 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3)</td>
<td>$2.2 \times 10^{-1}$</td>
<td>$1.3 \times 10^{-2}$</td>
<td>$1.9 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(4)</td>
<td>$1.6 \times 10^{-1}$</td>
<td>$1.0 \times 10^{-2}$</td>
<td>$1.3 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(6)</td>
<td>$7.4 \times 10^{-2}$</td>
<td>$7.4 \times 10^{-3}$</td>
<td>$7.9 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

Among all situations considered, the mean of the scaled absolute difference between BLP...
evaluations and BP evaluations was the largest for situation 2. The four FLMS considered in this situation, FLM(2), FLM(3), FLM(4), and FLM(6), yielded means of 0.1, 0.05, 0.035, and 0.021 (Table 4.6). Figure 4.3 provides a graphical display of the differences in accuracy between the four FLMS considered in situation 2. For FLM(2), FLM(3), and FLM(4), a large proportion of the absolute differences between BLP evaluations and BP evaluations are larger than 0.1 genetic standard deviations. For FLM(6), however, 195 out of 200 absolute differences were smaller than 0.1 genetic standard deviations. Thus, genetic evaluations obtained by BP under FLM(6) can be considered good approximations of genetic evaluations obtained by BLP. Using the same reasoning, for each of the nine situations considered, the first FLMS that yielded a mean accuracy below 0.025 was considered to have a sufficiently large number of loci for accurate BP evaluations. Table 4.14 summarizes the results of the analyses of all nine purebred situations considered. Thus, for all situations considered, the desired level of accuracy was reached with FLMS with two to six loci.

Table 4.14 The value of $N_1$ and the corresponding mean scaled absolute difference between BLP and BP evaluations under $FLM(N_1)$ for the nine situations considered in the purebred case.

<table>
<thead>
<tr>
<th>FLM</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_1$</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Mean</td>
<td>0.0085</td>
<td>0.021</td>
<td>0.0076</td>
<td>0.018</td>
<td>0.0076</td>
<td>0.018</td>
<td>0.0026</td>
<td>0.013</td>
<td>0.019</td>
</tr>
</tbody>
</table>

4.3.2 Crossbred case

For situation 1 of the two-breed case, Table 4.15 summarizes the level of accuracy reached by BP evaluations under FLM(2,1). Note that an acceptable level of accuracy is reached under this model. Thus, in this situation, a three locus model yields acceptably accurate evaluations for data generated with a FLML with 100 + 1 loci.

For situation 2 of the two-breed case, Table 4.16 summarizes the level of accuracy reached by BP evaluations under FLM(2,1), FLM(3,1) and FLM(4,1). For this situation, which differs
Figure 4.3  Box plots of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2), FLM(3), FLM(4) and FLM(6) for situation 2.

Table 4.15  Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2,1), for two-breed situation 1. The scale used was $\sqrt{6.72}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2,1)</td>
<td>$6.0 \times 10^{-2}$</td>
<td>$6.4 \times 10^{-3}$</td>
<td>$8.6 \times 10^{-3}$</td>
</tr>
</tbody>
</table>
from situation 1 only in the magnitude of the heritability, a higher number of loci is needed to reach a good level of accuracy.

Table 4.16 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2,1), FLM(3,1) and FLM(4,1), for two-breed situation 2. The scale used was $\sqrt{6.72}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2,1)</td>
<td>$3.9 \times 10^{-1}$</td>
<td>$5.7 \times 10^{-2}$</td>
<td>$5.9 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3,1)</td>
<td>$2.6 \times 10^{-1}$</td>
<td>$3.2 \times 10^{-2}$</td>
<td>$3.4 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(4,1)</td>
<td>$2.1 \times 10^{-1}$</td>
<td>$2.3 \times 10^{-2}$</td>
<td>$2.5 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

For situation 3 of the two-breed case, Table 4.17 summarizes the level of accuracy reached by BP evaluations under FLM(2,1). This situation, differs from situation 1 only in the pedigree type. A lower level of accuracy is reached in this case as opposed to situation 1. However, in this situation the BP evaluations were calculated by ESIP as opposed to situation 1 where BP evaluations were calculated by the Elston-Stewart algorithm.

Table 4.17 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2,1) for two-breed situation 3. The scale used was $\sqrt{6.72}$.

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2,1)</td>
<td>$6.5 \times 10^{-2}$</td>
<td>$8.8 \times 10^{-3}$</td>
<td>$9.5 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

For situation 4, Table 4.18 summarizes the level of accuracy reached by BP evaluations under FLM(2,1), FLM(3,1) and FLM(4,1). For this situation, which differs from situation 2 only in the pedigree type, a similar accuracy level is observed. In this situation again ESIP was used but the extension of the pedigree seems to compensate the loss in accuracy when compared to the Elston-Stewart algorithm.

If we use the same level of accuracy used to summarize the purebred case we can conclude that, for the two-breed data considered, FLMS with three to five loci yield accurate evaluations for traits modeled with FLML with $100 + 1$ loci.
Table 4.18 Summary statistics of 200 values of the scaled absolute difference between BLP evaluations and BP evaluations under FLM(2,1), FLM(3,1) and FLM(4,1), for two-breed situation 4. The scale used was $\sqrt{6.72}$.  

<table>
<thead>
<tr>
<th>FLMS</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLM(2,1)</td>
<td>$3.4 \times 10^{-1}$</td>
<td>$5.3 \times 10^{-2}$</td>
<td>$5.3 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(3,1)</td>
<td>$2.1 \times 10^{-1}$</td>
<td>$3.2 \times 10^{-2}$</td>
<td>$3.4 \times 10^{-2}$</td>
</tr>
<tr>
<td>FLM(4,1)</td>
<td>$1.6 \times 10^{-1}$</td>
<td>$2.4 \times 10^{-2}$</td>
<td>$2.5 \times 10^{-2}$</td>
</tr>
</tbody>
</table>

4.4 Discussion

This paper investigates the effect of the number of loci on genetic evaluations under finite locus models. Purebred and multibreed data were simulated with finite locus models with a large number of loci. Genetic evaluations were obtained by BLP. The parameters needed for BLP evaluation were calculated using the gene frequencies and genotypic effects of the simulation model. Genetic evaluations were also obtained by BP under several FLMS. The parameters for these FLMS were derived such that they yield the same genotypic mean, additive and dominance variance as the simulation model. Genetic evaluations obtained by BLP were used as reference values to investigate the accuracy of BP evaluations calculated under several “equivalent” FLMS.

For the simulated purebred data, models with two to six loci yield accurate BP evaluations for traits determined by 100 loci. Lowly heritable traits can be evaluated with high accuracy by BP under models with two to three loci. As the heritability increases, a higher number of loci is needed to obtain accurate evaluations by BP. For a highly heritable trait, accurate BP evaluations were obtained under models with six loci. The number of loci needed for accurate evaluation by BP was not affected by pedigree size. The presence of inbreeding reduced the accuracy of the BP evaluations but, this reduction was small. The number of loci needed to obtain accurate BP evaluations was not affected by missing phenotypic data.

For the simulated crossbred data, models with two to five loci yield accurate BP evaluations for traits determined by $100 + 1$ loci. In all situations considered, the results obtained for
crossbred data were similar to the results obtained for purebred data.

The results presented in this paper suggest that genetic evaluation by BP under models with two to six loci yields accurate results for data generated with about 100 loci. With models of this size, MCMC methods can be used efficiently for large pedigrees with non-additive gene action. BP evaluation by MCMC under models with no more than six loci could provide a solution to the problem of genetic evaluation of traits that show non-additive inheritance even in multibreed populations.

Finite locus models with small number of loci could be very useful for the problem of parameter estimation in multibreed populations as well. Under a linear model, even in a purebred population, large amounts of data are needed to obtain estimates of non-additive effects [3, 13]. In a two breed population with inbreeding, estimates of 5 location and 25 dispersion parameters [23] are needed. Thus, even larger amounts of data would be needed in this case. This, however, might be impractical in livestock populations. By using a finite locus model with a small number of loci, the number of parameters that need to be estimated could be reduced significantly [12, 24, 28]. Thus, parameter estimation in multibreed populations would become practical.
Figure 4.4 Extended Pedigree
Figure 4.5  Pedigree with loops
Bibliography


CHAPTER 5. A STRATEGY TO IMPROVE THE COMPUTATIONAL EFFICIENCY OF MARKER ASSISTED GENETIC EVALUATION UNDER FINITE LOCUS MODELS

A paper accepted for publication in the Proceedings of the 7th WCGALP Congress

Liviu R. Totir, Rohan L. Fernando, Jack C.M. Dekkers and Soledad A. Fernández

Abstract

Marker assisted genetic evaluation is most useful for lowly heritable traits. These type of traits are also known to have non additive inheritance. BLUP methodology has computational problems under non additive inheritance. Under a finite locus model, the best predictor can be estimated by MCMC regardless of the mode of inheritance. ESIP, is an MCMC sampler that samples genotypes jointly from the entire pedigree one locus at a time conditional on the genotypes at all the other loci. For models with more than one locus, the computational efficiency, as well as the accuracy of ESIP are influenced by the number of samples generated per peeling. For lowly heritable traits, ESIP with 5 to 15 samples per peeling yielded the most accurate results in a fixed amount of time.

5.1 Introduction

BLUP methodology can be used for marker assisted genetic evaluation (MAGE) [2]. MAGE is most useful for traits that have low heritability. Lowly heritable traits are also known to
have non additive inheritance. Although BLUP is computationally efficient under additive inheritance, it is inefficient under non additive inheritance.

Regardless of the mode of inheritance, under a finite locus model, the best predictor (BP), which is the conditional mean of the genotypic value given trait and marker data, can be estimated by Markov chain Monte Carlo (MCMC) [3, 4, 6].

The single site Gibbs sampler has been widely used in genetic analyses. However, it is known that this sampler may yield unreliable results. For example, when a marker locus has more than two alleles, single site Gibbs may not result in a chain that is irreducible. Even when the generated chain is irreducible, mixing might be very slow [5]. These problems can be overcome by sampling jointly genotypes from the entire pedigree. ESIP, an MCMC sampler that combines two peeling techniques (the Elston-Stewart algorithm and iterative peeling) can be used to generate joint genotype samples from the entire pedigree [1].

The computational efficiency of ESIP is determined by two distinct processes: peeling and sampling. Peeling is used to calculate genotype probabilities, which are then used in reverse peeling to sample genotypes jointly from the entire pedigree [1]. Once the pedigree has been peeled, the computing time required for sampling genotypes by reverse peeling is comparable to the computing time for single site Gibbs. The computing time required for peeling depends on the complexity of the pedigree and the number of missing genotypes. For the situation considered below, the time required for peeling was about ten times that for sampling. For a single locus model, the same genotype probabilities are reused repeatedly in reverse peeling. Thus, peeling must be done only once, and as a result the computing time required for peeling has negligible effect on the computational efficiency of ESIP. For multilocus models, in order to preserve a linear relationship between computational efficiency and the number of loci in the model, genotypes are sampled one locus at a time conditional on the current genotype configuration at the other loci. Thus, whenever the sampler moves to a new locus, genotype probabilities must be recomputed by peeling. For a given locus, after peeling, k samples can be obtained before moving to the next locus. As the size of k increases the computational efficiency of the sampler increases, but it also results in increased dependence between samples.
The effect of the number of samples (k) per peeling on the computational efficiency of ESIP is discussed in this paper.

5.2 Materials and Methods

For a simple pedigree with 14 individuals and no loops, data were simulated using a finite locus model that contained one quantitative trait locus flanked by two markers (MQTL), and 100 remaining QTL (RQTL) unlinked to the flanking markers. All QTL had a gene frequency of 0.5. The additive effect of the MQTL was 2 while the dominance effect was 0. Each RQTL had an additive effect of 0.2828, and 50 of the RQTL had a dominance effect of 0.2828, while the other 50 had a dominance effect of -0.2828. An environmental variance of 63 was used. Thus, the simulated trait had a narrow sense heritability of 0.08 and a broad sense heritability of 0.11. The marker data were simulated assuming 12 alleles at each marker locus, and a recombination rate of 0.05 between each marker and the MQTL.

The simulated data were analyzed using a finite locus model that contained one MQTL and two RQTL. The MQTL had the same parameters as the MQTL in the simulation. For the two RQTL, the parameters were derived so that they yield the same first and second moment as the 100 RQTL of the simulation. Thus, the additive effect for each RQTL was 2, and one RQTL had a dominance effect of 2 while the other had a dominance effect of -2. In the analysis, an environmental variance of 63 was used as well.

The conditional mean of the vector of genotypic values (v) given the vector of trait data (y) and the matrix of marker data (M) can be written as

\[ E(v \mid y, M) = 1\eta + \sum_G v_G \Pr(G \mid y, M), \]  

(5.1)

where 1 is an 14 × 1 vector of ones; η is a fixed effect common to all individuals; \( v_G \) is the 14 × 1 vector of genotypic values that corresponds to the genotypic configuration \( G \). A given genotypic configuration \( G \) can be represented as a 14 × 3 matrix, where each column of \( G \) is
equivalent to the $14 \times 1$ vector of genotypes for a given QTL (MQTL or RQTL). BP's were estimated by averaging joint genotype samples generated by several ESIP samplers.

We use the notation ESIP(1:k) for an ESIP sampler with k samples per peeling. Samples generated by ESIP(1:1) are independent and thus, were used to calculate reference genetic evaluations for all members of the pedigree. These reference evaluations were estimated by averaging 1,000,000 ESIP(1:1) samples.

Figure 5.1 shows the behavior of ESIP(1:1) in terms of the mean sum of squares of the difference between the reference evaluations and genetic evaluations computed at different stages of a chain of 1,000,000 samples. ESIP(1:1) reached a high level of accuracy in a short time, and then the accuracy improved slowly throughout the run. To have a meaningful

![Figure 5.1 Behavior of the mean sum of squares of the difference for ESIP(1:1)](image)
comparison between the samplers, genetic evaluations were obtained by running each sampler for the same amount of time. The time allocated to each sampler was equal to the time needed by ESIP(1:1) to generate 100,000 samples. At the end of the allocated time, the absolute difference between the reference evaluations and ESIP(1:k) were scaled by the genetic standard deviation. For each ESIP(1:k), the maximum and the mean of the scaled absolute differences, and the scaled square root of the mean squared error were used to summarize the performance of the sampler.

5.3 Results and Discussion

Table 5.1 summarizes the performance of several ESIP(1:k) samplers. As k increased, the number of samples generated in the allocated time increased as well. The rate of increase, however, decreased rapidly after k=100. In terms of accuracy, ESIP(1:5) performed best. For k's larger than 15, increased dependence between samples resulted in reduced accuracy. For example, ESIP(1:100) was approximately two times less accurate than ESIP(1:5), while ESIP(1:1000) was ten times less accurate than ESIP(1:5). ESIP(1:100000) had the highest dependence between samples and thus, it was the least accurate sampler.

<table>
<thead>
<tr>
<th>Sampler</th>
<th>No. samples included</th>
<th>Maximum</th>
<th>Mean</th>
<th>√MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESIP(1:1)</td>
<td>100,000</td>
<td>0.0059</td>
<td>0.0026</td>
<td>0.0032</td>
</tr>
<tr>
<td>ESIP(1:5)</td>
<td>361,433</td>
<td>0.0031</td>
<td>0.0014</td>
<td>0.0017</td>
</tr>
<tr>
<td>ESIP(1:10)</td>
<td>535,510</td>
<td>0.0059</td>
<td>0.0016</td>
<td>0.0021</td>
</tr>
<tr>
<td>ESIP(1:15)</td>
<td>640,000</td>
<td>0.0035</td>
<td>0.0014</td>
<td>0.0018</td>
</tr>
<tr>
<td>ESIP(1:20)</td>
<td>716,940</td>
<td>0.0048</td>
<td>0.0021</td>
<td>0.0023</td>
</tr>
<tr>
<td>ESIP(1:100)</td>
<td>950,725</td>
<td>0.0089</td>
<td>0.0027</td>
<td>0.0035</td>
</tr>
<tr>
<td>ESIP(1:1000)</td>
<td>1,025,000</td>
<td>0.0293</td>
<td>0.0107</td>
<td>0.0137</td>
</tr>
<tr>
<td>ESIP(1:10000)</td>
<td>1,041,270</td>
<td>0.0903</td>
<td>0.0247</td>
<td>0.0358</td>
</tr>
<tr>
<td>ESIP(1:100000)</td>
<td>1,041,270</td>
<td>0.2339</td>
<td>0.0998</td>
<td>0.1220</td>
</tr>
</tbody>
</table>

The finite locus model used in this analysis had only three loci, one MQTL and two RQTL. For lowly heritable traits, two RQTL provide a good approximation for the polygenic component [7]. MAGE is known to have the greatest advantage for lowly heritable traits.
When heritability is low, for models that fit a single MQTL, ESIP(1:k) with k between five and 15 will be most efficient.

For traits with high heritability, a larger number of RQTL need to be used to approximate the polygenic component [7]. As the number of RQTL in the model increases, the dependence between samples generated by ESIP(1:k) with large k will increase as well. Thus, for traits with high heritability or models with several MQTL, an ESIP(1:k) with k smaller than five might be more efficient.

Acknowledgements

Research support was provided by award no. 2002-35205-1156 of the National Research Initiative Competitive Grants Program of the USDA.

Bibliography


CHAPTER 6. GENERAL DISCUSSION

In this dissertation we investigated the use of finite locus models for genetic evaluation. The availability of genotypic data in recent years has resulted in increased interest in the use of marker assisted genetic evaluation (MAGE) in livestock species. Best linear unbiased prediction (BLUP), obtained from Henderson's mixed model equations (HMME), is currently the most widely used method for genetic evaluation. Under additive inheritance, Henderson's mixed model equations (HMME) provide an efficient way to obtain genetic evaluations by marker assisted best linear unbiased prediction (MABLUP) given pedigree relationships, trait, and marker data. For large pedigrees with many missing markers, however, it is not feasible to calculate the exact gametic variance covariance matrix required to construct HMME. Thus, approximate methods need to be used.

In Chapter 2 we have investigated the effect of using approximate gametic variance covariance matrices on response to selection by MABLUP. For simulated data sets with complete or incomplete marker data for flanking markers with two, four, or 12 alleles, the response to selection by MABLUP using two approximate methods (Method A and Method B) was compared with that obtained by MABLUP using the exact genetic variance covariance matrix estimated by ESIP. For Method A, when the linkage phase between markers was known, marker information at both flanking markers was used, and when linkage phase was not known, marker information was ignored completely. For Method B, when the linkage phase between markers was known, marker information at both flanking markers was used, but when linkage phase was not known, the marker information at one of the two flanking markers was used. For all situations considered, Method B yielded a higher response to selection than Method A but lower than the exact method. It was concluded that while Method B was a significant
improvement over Method A, a higher response to selection by MABLUP can be obtained with more accurate gametic variance covariance matrices.

MAS is known to be efficient especially for traits that have low heritability and non-additive gene action. BLUP methodology under non-additive gene action is not feasible for large inbred or crossbred pedigrees. It is easy to incorporate non-additive gene action in a finite locus model. Under such a model, the unobservable genotypic values can be predicted using the conditional mean (BP).

In Chapter 3 we have investigated the potential of four alternative methods to compute the BP of genotypic values for a quantitative trait based on phenotypic data using finite locus models. Simulated data were used to assess the performance of iterative peeling and of three Markov chain Monte Carlo (MCMC) methods (scalar Gibbs, blocking Gibbs, and ESIP) to calculate BP in finite locus models. Iterative peeling yielded exact results for pedigrees without loops, but for pedigrees with loops, the accuracy of the results decreased with the increase in the number of loops. Also, iterative peeling had a serious limitation due to the exponential relationship between computing time and the number of loci in the model. For MCMC methods, a linear relationship can be maintained between computing efficiency and the number of loci in the model, by sampling genotypes one locus at a time. Out of the three MCMC methods considered, ESIP performed best while scalar Gibbs performed worst. Based on the results of Chapter 3 we concluded that genetic evaluation, based on conditional means of genotypic values given the phenotypes, is feasible under a finite locus model. ESIP proved to be an efficient as well as a reliable method for computing conditional means under such a model.

Successful application of MCMC methods for genetic evaluation depends, among other factors, on the number of loci assumed in the model. In Chapter 4 we have investigated the effect of the number of loci in finite locus models for genetic evaluation by BP. Purebred and multibreed data were simulated with finite locus models with a large number of loci. Genetic evaluations were then obtained by BLP. The parameters needed for BLP evaluation were calculated using the gene frequencies and genotypic effects of the simulation model. Genetic
evaluations were also obtained by BP under several finite locus models with a small number of loci. The parameters for these models were derived such that they yield the same genotypic mean, additive and dominance variance as the simulation model. Genetic evaluations obtained by BLP were used as reference values to investigate the accuracy of BP evaluations calculated under several "equivalent" finite locus models with a small number of loci. It was concluded that genetic evaluation by BP under models with two to six loci yields accurate results for both purebred and crossbred data generated with about 100 loci. With models of this size, MCMC methods can be used efficiently for large pedigrees with non-additive gene action.

The application, in real livestock pedigrees, of MAGE by BP under finite locus models, depends on the computational efficiency of the method used to calculate the BP evaluations. In Chapter 5 we have investigated briefly a strategy to improve the computational efficiency of MAGE under finite locus models. Based on the results of Chapter 3, ESIP was chosen as the method of choice for MAGE. The computational efficiency of ESIP is determined by two distinct processes: peeling and sampling. For multilocus models, in order to preserve a linear relationship between computational efficiency and the number of loci in the model, genotypes are sampled one locus at a time conditional on the current genotype configuration at the other loci. Thus, whenever the sampler moves to a new locus, genotype probabilities must be recomputed by peeling. For a given locus, after peeling, k samples can be obtained before moving to the next locus. Simulated data for a trait determined by one marked QTL and several unmarked QTL were used to investigate the effect of the number of samples (k) per peeling on the computational efficiency of ESIP. It was concluded that a value of k between 5 and 15 yielded the most accurate results in a given amount of time.

Under additive inheritance, MABLUP is feasible for large pedigrees with many missing markers. But, the results in Chapter 2 indicate that more research is needed to make better use of the observable marker information. As MABLUP is not feasible under non additive inheritance, we have investigated the use of genetic evaluation by BP calculated using MCMC under finite locus models. Our results indicate that BP calculated using ESIP under finite locus models with two to six loci yields accurate genetic evaluations. Throughout this thesis
we have assumed that all genetic parameters were known. In theory, Bayesian methodology using MCMC methods can be used to obtain genetic evaluations by marginalizing over the unknown parameters. Much research is being undertaken at this time to improve the efficiency of MCMC methods. Improved MCMC methods would facilitate better use of marker information in MABLUP under additive inheritance and would make BP feasible under non-additive inheritance with large pedigrees for models with many marked and unmarked QTL, even when the genetic parameters are unknown.
APPENDIX. SUPPLEMENTARY RESULTS FOR CHAPTER 3.

Figure 6.1 Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 35000 samples for blocking Gibbs, or 70000 samples for scalar Gibbs, for situation 1. □ ESIP; + Blocking Gibbs; △ Scalar Gibbs.
Figure 6.2. Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 35000 samples for blocking Gibbs, or 70000 samples for scalar Gibbs, for situation 2. □ ESIP; + Blocking Gibbs; △ Scalar Gibbs.
Figure 6.3 Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5,000 samples for ESIP, 35,000 samples for blocking Gibbs, or 70,000 samples for scalar Gibbs, for situation 3. □ ESIP; + Blocking Gibbs; △ Scalar Gibbs.
Figure 6.4  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 35000 samples for blocking Gibbs, or 70000 samples for scalar Gibbs, for situation 4. □ ESIP; + Blocking Gibbs; △ Scalar Gibbs.
Figure 6.5  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5,000 samples for ESIP, 35,000 samples for blocking Gibbs, or 70,000 samples for scalar Gibbs, for situation 5. □ ESIP; + Blocking Gibbs; Δ Scalar Gibbs.
Figure 6.6 Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 65 000 samples for blocking Gibbs, or 130 000 samples for scalar Gibbs, for situation 6. □ ESIP; + Blocking Gibbs; Δ Scalar Gibbs.
Figure 6.7  Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 35000 samples for blocking Gibbs, or 70000 samples for scalar Gibbs, for situation 7. □ ESIP; + Blocking Gibbs; △ Scalar Gibbs.
Figure 6.8 Mean sum of squares of the difference of estimates of genotypic values obtained by MCMC, relative to estimates obtained by the Elston-Stewart algorithm, calculated after each 5000 samples for ESIP, 35 000 samples for blocking Gibbs, or 70 000 samples for scalar Gibbs, for situation 8. □ ESIP; + Blocking Gibbs:Δ Scalar Gibbs.