# Estimation of Variance Components in Finite Polygenic Models and Complex Pedigrees

by

Katharine Gage Lahti

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

IN

DAIRY SCIENCE

APPROVED:

_____

Ina Hoeschele, Chair

_____ _____

David Notter Eric Hallerman

_____

William E. Vinson, Department Head

June 5, 1998

Blacksburg, Virginia

**Keywords**: finite polygenic model, nonadditive genetic variance, Bayesian methods

# ESTIMATION OF VARIANCE COMPONENTS IN FINITE POLYGENIC MODELS AND COMPLEX PEDIGREES

by

Katharine Gage Lahti

Ina Hoeschele, Chair

Dairy Science

(ABSTRACT)

Various models of the genetic architecture of quantitative traits have been considered to provide the basis for increased genetic progress. The finite polygenic model (FPM), which contains a finite number of unlinked polygenic loci, is proposed as an improvement to the infinitesimal model (IM) for estimating both additive and dominance variance for a wide range of genetic models. Analysis under an additive five-loci FPM by either a deterministic Maximum Likelihood (DML) or a Markov chain Monte Carlo (MCMC) Bayesian method (BGS) produced accurate estimates of narrow-sense heritability (0.48 to 0.50 with true values of $h^2 = 0.50$) for phenotypic data from a five-generation, 6300-member pedigree simulated without selection under either an IM, FPMs containing five or forty loci with equal homozygote difference, or a FPM with eighteen loci of diminishing homozygote difference. However, reducing the analysis to a three- or four-loci FPM resulted in some biased estimates of heritability (0.53 to 0.55 across all genetic models for the 3-loci BGS analysis and 0.47 to 0.48 for the 40-loci FPM and the infinitesimal model for both the 3- and 4-loci DML analyses). The practice

of cutting marriage and inbreeding loops utilized by the DML method expectedly produced overestimates of additive genetic variance (55.4 to 66.6 with a true value of $\sigma_a^2$ = 50.0 across all four genetic models) for the same pedigree structure under selection, while the BGS method was mostly unaffected by selection, except for slight overestimates of additive variance (55.0 and 58.8) when analyzing the 40-loci FPM and the infinitesimal model, the two models with the largest numbers of loci. Changes to the BGS method to accommodate estimation of dominance variance by sampling genotypes at individual loci are explored. Analyzing the additive data sets with the BGS method, assuming a five-loci FPM including both additive and dominance effects, resulted in accurate estimates of additive genetic variance (50.8 to 52.2 for true $\sigma_a^2$ = 50.0) and no significant dominance variance (3.7 to 3.9) being detected where none existed. The FPM has the potential to produce accurate estimates of dominance variance for large, complex pedigrees containing inbreeding, whereas the IM suffers severe limitations under inbreeding. Inclusion of dominance effects into the genetic evaluations of livestock, with the potential increase in accuracy of additive breeding values and added ability to exploit specific combining abilities, is the ultimate goal.

# ACKNOWLEDGMENTS

I would like to thank everyone who made it possible for me to study at Virginia Tech and complete my degree—especially my major professor, Dr. Ina Hoeschele, and my committee members, Drs. David Notter, William Vinson, and Eric Hallerman. I also wish to thank Fengxing Du for his programming assistance and helpful discussions regarding dominance estimation.

I am extremely grateful for the support and wise advise I received along the way from Dr. Dennis Banks, Dr. Fernando Grignola, and all my fellow graduate students, especially Beth Grove Dransfield, Lori Smith, and Amin Ahmadzadeh.

My parents, Allan and Carole Gage, and my husband, Richard Lahti, receive highest encomiums for their understanding, encouragement, and patience.

# TABLE OF CONTENTS

# LIST OF TABLES

# LITERATURE REVIEW

Traits of economic importance in livestock breeding generally are assumed to be quantitative and polygenic in nature, consisting of a measurable phenotype influenced by genes at many loci and by environmental factors. The genetic component of the phenotype can be partitioned into additive contributions from individual alleles and non-additive contributions due to specific combinations or interactions of alleles at one or more loci. Non-additive gene action is defined as dominance when the interaction involves the alleles at one locus and as epistasis when the interaction is between genotypes at different loci. This introduction will concentrate on the dominance form of non-additive gene action, the need for an improved method to estimate its variance, and its potential for use in genetic evaluations. The primary focus of this thesis, the finite polygenic model (FPM), will also be introduced. When utilized as an alternative to the infinitesimal model, the FPM, used in conjunction with a Bayesian analysis implemented by Markov chain Monte Carlo (MCMC) sampling algorithms, may prove to be an improved method of estimating dominance variance.

## Dominance Variance

The components mentioned previously, into which the phenotype can be partitioned, are expressed as deviations from the population mean. Dominance deviation *d* is the deviation of the genotypic value of the heterozygous genotype at a biallelic locus

from the average genotypic value of the homozygous genotypes at that same locus. The dominance effect, dependent on the gene frequency of the population and *d*, for a biallelic locus consisting of alleles $A_1$ with frequency *p* and $A_2$ with frequency *q*, equals $-2q^2d$, $2pqd$, and $-2p^2d$ for genotypes $A_1A_1$, $A_1A_2$, and $A_2A_2$, respectively. Dominance variance at one locus is the sum of the squared dominance effects multiplied by the genotypic frequencies of $p^2$, $2pq$, and $q^2$, equal to $(2pqd)^2$. Dominance variance for multiple loci is then the sum of this result across all loci

$$\sigma_d^2 = \sum_{loci} (2pqd)^2 \qquad (1)$$

(Falconer & Mackay, 1996).

## Including Dominance Variance in Genetic Evaluations

The goal of genetic improvement programs is to improve the phenotype by enhancing genetic merit through implementation of selection programs based on well-designed mating systems. Genetic evaluation of livestock species currently is conducted using additive genetic models and Best Linear Unbiased Prediction (BLUP) methodology. Non-additive genetic variation in the form of dominance is ignored for a variety of reasons: the amount of dominance variation existing for most traits is assumed to be negligible as a percentage of total phenotypic variation, the number of dominance relationships in the population of interest is often small, and estimation of dominance variance under the infinitesimal gene model is difficult and computationally expensive, especially when inbreeding is present.

## Existence of Dominance Variation

There is ample evidence that dominance variation occurs in traits of economic importance in livestock species, including production, growth and reproduction traits. In Holstein dairy cattle, Allaire and Henderson (1965) found that dominance variation explained 50.0% of phenotypic variation in milk yield, although there appeared to be no dominance component to fat yield. Tempelman and Burnside (1990a, 1991) reported just the opposite, finding that 25.0% of fat yield variance and no significant proportion of milk yield variance were due to dominance. They also looked at nine conformation traits in Canadian Holsteins and found dominance genetic variation to be important for final score, capacity, and mammary system (Tempelman & Burnside, 1990b). Thomas *et al.* (1985) studied type traits in Jerseys and found dominance components twice those of additive components for dairy character, rump width, fore udder, rear udder height, rear udder width, and suspensory ligament. Hoeschele (1991) examined the female fertility traits of days open and service period in Holsteins and found small variance components (less than 3.4% of phenotypic variance), dependent on upper bounds, for both additive and non-additive effects, although dominance variance was generally higher than additive variance. Van Raden *et al*. (1992) estimated much smaller dominance variances for milk and fat yield than previous studies, 3.5% and 3.3% of phenotypic variance, respectively.

Fewer studies have been performed on meat animal species. Wei and van der Werf (1993) found an appreciable amount of dominance variance (10% to 20% of phenotypic variance) for egg number, but lesser amounts for egg weight and specific gravity in three lines of White Leghorn chickens. Rodriguez-Almeida *et al*. (1995) estimated dominance variances between 11% and 28% for height and weight of three synthetic lines of beef cattle at birth and weaning, while Mizstal *et al*. (1998) found that 9.9% of variance of post-weaning gain in Limousin beef cattle was accounted for by dominance. In a population of purebred American Yorkshire swine, Culbertson *et al*. (1998) found sizable estimates of dominance variance for number born alive, 21-day litter weight, days to 104.5 kg, and backfat thickness.

Estimates of additive and dominance variance as a percent of phenotypic variance, along with standard errors when available, are presented in Table 1 for dairy cattle and in Table 2 for meat animal species. The various methods used to obtain these estimates will be described later. Despite discrepancies in estimation of the exact amount of dominance variance and large standard errors in some of the studies, there is reason to believe that dominance variance contributes significantly to the variation of many economically important traits in livestock. Therefore, improved methods of estimating dominance variance should be investigated to facilitate the inclusion of dominance effects in genetic evaluations.

**Table 1. Estimates of additive and dominance variances as a percent of phenotypic variance (with standard errors) from different studies of Holstein dairy cattle production, conformation, and fertility traits.**

| Reference | Method of Estimation | Model | Trait | % $\sigma_a^2$ | % $\sigma_d^2$ |
|---|---|---|---|---|---|
| Allaire & Henderson, 1965 | Henderson's Method III | Sire-MGS[1] | Milk yield | 11.0 (6.0) | 50.0 (26.0) |
| | | | Fat yield | 23.0 (10.0) | -18.0 (35.0) |
| Tempelman & Burnside, 1990a | REML | Hierarchical | Milk yield | 40.0 (3.0) | 6.0 (8.0) |
| | | | Fat yield | 32.0 (3.0) | 24.0 (8.0) |
| Tempelman & Burnside 1990b | REML | Hierarchical | Final Score | 15.0 (2.2) | 15.0 (10.1) |
| | | | General Appearance | 13.6 (2.1) | 12.1 (10.0) |
| | | | Dairy character | 21.2 (2.4) | 2.7 (8.5) |
| | | | Capacity | 29.2 (2.8) | 16.1 (8.1) |
| | | | Rump | 21.9 (2.7) | 6.7 (9.0) |
| | | | Feet & legs | 9.7 (1.8) | 6.6 (10.8) |
| | | | Mammary system | 14.4 (2.2) | 12.9 (9.9) |
| | | | Fore udder | 13.4 (2.0) | 10.9 (9.3) |
| | | | Rear udder | 13.3 (2.2) | 7.6 (10.3) |

[1] MGS = Maternal grandsire

**Table 1. Continued**

| Reference | Method of Estimation | Model | Trait | % $\sigma_a^2$ | % $\sigma_d^2$ |
|---|---|---|---|---|---|
| Tempelman & Burnside, 1991 | REML | Animal | Milk yield (by region) | 34.0 - 49.0 (4.0-6.0) | 8.0 - 31.0 (15.0 - 19.0) |
| | | | Fat yield (by region) | 34.0 - 48.0 (4.0 - 5.0) | 8.0 - 49.0 (15.0 - 20.0) |
| Hoeschele, 1991 | Tilde-hat REML approximation | Sire-MGS[1] | Days open | 2.0 (1.0) | 2.3 (4.0) |
| Van Raden *et al*. 1992 | Tilde-hat REML approximation | Sire-MGS[1] | Milk yield | | 3.5 (4.2) |
| | | | Fat yield | | 3.3 (4.2) |
| Misztal *et al*. 1998 | Method R | Animal | Milk yield | 43.5 (0.7) | 5.7 (0.4) |
| | | | Fat yield | 42.6 (0.7) | 7.0 (1.2) |
| | | | Protein yield | 40.6 (0.2) | 4.9 (0.8) |
| | | | Stature | 45.3 (0.3) | 6.9 (1.2) |
| | | | Strength | 27.8 (0.5) | 8.0 (0.7) |
| | | | Body depth | 34.5 (0.3) | 9.8 (0.7) |
| | | | Dairy form | 23.4 (0.4) | 5.3 (1.0) |
| | | | Fore udder attachment | 24.3 (0.5) | 4.7 (0.7) |

[1] MGS = Maternal grandsire

**Table 2. Estimates of additive and dominance variances as a percent of phenotypic variance (with standard errors) from different studies of meat animal production traits.**

| Reference | Method of Estimation | Model | Species | Trait | % $\sigma_a^2$ | % $\sigma_d^2$ |
|---|---|---|---|---|---|---|
| Wei & van der Werf, 1993 | REML | Animal | 3 White Leghorn lines | Egg number | 12.0 - 52.0 (3.1 - 5.8) | 10.0 - 20.0 (4.6 - 6.0) |
| | | | | Egg weight | 38.0 - 63.0 (3.9 - 5.6) | 1.0 - 13.0 (3.8 - 5.3) |
| | | | | Egg specific gravity | 31.0 - 40.0 (3.7 - 4.8) | 1.0 - 13.0 (4.2 - 5.4) |
| Rodriguez-Almeida *et al*. 1995 | REML | Animal | 3 synthetic beef lines | Birth weight | 53.0 | 18.0 |
| | | | | Birth hip height | 42.0 | 26.0 |
| | | | | 205-day weight | 27.0 | 28.0 |
| | | | | 205-day hip height | 35.0 | 11.0 |
| Misztal *et al*. 1998 | Method R | Animal | Limousin beef cattle | Postweaning gain | 21.0 (1.1) | 9.9 (1.6) |
| Culbertson, *et al*. 1998 | | | Yorkshire Swine | Number born alive | 8.8 (0.5) | 2.2 (0.7) |
| | | | | 21-day litter weight | 8.1 (1.1) | 6.3 (0.9) |
| | | | | Days to 104.5 kg | 33.2 (0.4) | 10.3 (1.5) |
| | | | | Backfat thickness | 43.6 (0.9) | 4.8 (0.7) |

## Dominance Relationships

Another obstacle to the estimation of dominance variance and the inclusion of dominance effects in genetic evaluations of livestock species is the low frequency of dominance relationships present under traditional animal breeding practices. Mizstal *et al*. (1998) report that accurate estimates of dominance variances under the animal model require data sets with at least 30,000 to 100,000 animals for populations with many full-sibs, and even larger data sets for cattle populations, which contain comparatively fewer full-sibs. The three dominance relationships in livestock with the largest dominance contributions to covariances are shown in Table 3.

**Table 3. Contributions of dominance variance to covariances of close relatives with no inbreeding present.**

| Relative | Dominance contribution |
|---|---|
| Identical twin/clone | 1.0 |
| Full sibs | 0.25 |
| Three-quarter sibs | 0.0625 |

The lack of dominance relationships in livestock populations is due to low fecundity and other reproductive limitations, especially in cattle. In the dairy industry, improvements in this area are constantly being realized through increased use of artificial insemination (AI) to increase numbers of ¾-sibs, embryo transfer (ET) to increase

numbers of full-sibs, and embryo splitting and cloning technologies to increase numbers of identical (monozygotic) twins or clones. Even though the numbers of full-sibs, twins, and clones may be small compared to the entire population, such animals are likely to be the elite individuals of the population, and therefore, have a disproportionately greater influence on production at the population level. The structures of poultry, swine, and sheep populations include a larger number of dominance relationships than cattle populations, and may provide opportunities for dominance models to provide added value for genetic prediction and mating pair allocation.

## Estimating Dominance Variance

The accurate estimation of dominance variance in a computationally efficient manner with a model of analysis that includes maximum information from relatives comprises a distinct challenge for the large, complex, inbred pedigrees that occur in the livestock populations undergoing genetic evaluation. Methods used in previous studies of dominance variance estimation include Henderson's Method III (Henderson, 1953), Restricted Maximum Likelihood (REML) (Patterson and Thompson, 1971), and the tilde-hat approximation to REML (Van Raden and Jung, 1988). Because of several shortcomings to these methods, however, recent interest has been focused primarily on Method R (Reverter *et al*. 1994) and Bayesian methods using the Markov Chain Monte Carlo (MCMC) Gibbs sampler algorithm (Geman and Geman, 1984). Models of analysis have also evolved over the years, and the size of data sets and the amount of genetic information that can be utilized has increased. Early studies employed certain groups of

relatives or hierarchical models, later ones used the sire/maternal-grandsire (MGS) model, and the most recent analyze with the animal model.

Allaire and Henderson (1965) fit a two-way classification of sire, MGS, and sire x MGS effects, and estimated variances of these components by equating sums of squares to their expectations. Additive and non-additive genetic variances were then calculated from these components.

The next methods to be used were REML and approximate REML methods, which require the inverse of the dominance relationship matrix **D**. Prior to Hoeschele and Van Raden (1991) developing a rapid procedure to invert **D**, analyzing large data sets with an animal model was not feasible. Tempelman and Burnside (1990a,b) estimated dominance variance for production and type traits with REML using a hierarchical model assuming no epistatic, maternal, or cytoplasmic effects, which resulted in selection bias upward due to the common environment of full sisters within herds. Later they repeated their estimation of dominance variance for production traits using an animal model and derivative-free REML, but with very small data sets (Tempelman and Burnside, 1991). Although approximate under inbreeding, the inversion procedure of Hoeschele and Van Raden (1991) allowed dominance variance to be estimated with an animal model for a data set containing over 400,000 animals using successive over-relaxation (Van Raden *et al*. 1992), but at a processing cost of approximately 40 times that of an additive-only evaluation. Larger data sets required the use of a sire model (Van Raden *et al*. 1992; Hoeschele, 1990), which considers only about one quarter of the dominance information, because ignoring maternal granddams causes full-sibs to be treated as ¾-sibs and

eliminates three out of the four possible ¾-sib relationships, resulting in potential biases. Estimation was by the tilde-hat approximate REML method, which was found to be incompatible with the individual animal model, and also is not resistant to selection bias because its quadratics do not account for relationship matrices.

Misztal (1997) presented a procedure to estimate variance components for dominance models in larger data sets (up to 3 million first lactation records for stature of Holstein cattle) using Method R, which regresses current predictions on previous subsample predictions until the regression coefficient is equal to one, implying an accurate estimate of the variance ratio. This requires repeated solving of the mixed model equations (MME) instead of inversion of the MME coefficient matrix, resulting in an evaluation less than twice as computationally expensive as the additive model (Misztal, 1997).

Another method used recently in preliminary animal model estimation of dominance variance, and developed further in this thesis, is that of Bayesian inference implemented with MCMC sampling-based algorithms such as the Gibbs sampler (Goddard, 1998; Pong-Wong *et al*. 1998). Using the Gibbs sampler, means or modes of the joint posterior distribution of fixed effects, random effects, and variance components, given the data, can be calculated from sequential samples of the parameters from fully-conditional distributions. This method is being used in conjunction with assumptions of a finite polygenic model (FPM), as opposed to the traditional infinitesimal model assumed by the previous studies. A more complete discussion of the finite polygenic model follows. This MCMC Bayesian method assuming a FPM is not complicated by selection

or inbreeding as are many of the previously mentioned methods, and the elements of the mode of the joint posterior density of variance components when flat (non-informative) priors are used are equal to REML estimates. In preliminary studies, Pong-Wong *et al*. (1998) underestimated additive genetic variance $\sigma_a^2$ and error variance $\sigma_e^2$, and overestimated dominance genetic variance $\sigma_d^2$ in 50 replicates of a 425-member, five-generation pedigree, with the analysis assuming a constant dominance deviation across loci. However, estimates became more accurate as the number of loci in the analysis model increased. For a similar pedigree structure, also using an MCMC Bayesian analysis, Goddard (1998) held $\sigma_e^2$ constant and produced overestimates of both $\sigma_a^2$ and $\sigma_d^2$.

## Improvement in Genetic Evaluations and Mating Systems

Potential gains from including dominance in genetic evaluations include increased accuracy of prediction of additive breeding values and the added ability to use specific combining abilities for choosing mating pairs. Because current genetic evaluations ignore all non-additive genetic effects, evaluations lose accuracy as a function of the variance of nonadditive effects and the number of animals with dominance relationships (Misztal 1997).

Looking at traits from Tables 1 and 2, Misztal *et al*. (1998) found that omitting dominance from genetic evaluations caused biased estimates of additive breeding values

for full-sib families and particularly dams of embryo-transfer animals in cattle. Changes of additive breeding values in proven sires were small. They predict that with Method R estimation, "genetic gains from using dominance information will not be large, but they will outweigh expenditures to derive the information." Culbertson *et al*. (1998) also examined changes in prediction of additive breeding values of the swine traits found in Table 2 when inbreeding and dominance effects were included in the genetic prediction model. Although the correlation between breeding values with and without dominance and inbreeding effects was always greater than 0.99, changes for individual animals were noticeable for all traits except backfat thickness, and changes in prediction due to dominance genetic effects were found to increase as the amount of dominance information available for a given family increased.

The inclusion of dominance in the model provides marginal improvement in the evaluation of additive effects and allows selected matings between individuals to make use of specific combining abilities. DeStefano and Hoeschele (1992) proposed three different mating strategies which yielded significant genetic gains from 1.5% to 8% of the phenotypic standard deviation, which may even be an underestimate due to their use of the sire model. To develop a complete mating system, it would be ideal to have estimates of all possible parental combinations in the population. Varona and Mizstal (1998) describe a relatively inexpensive computing procedure to estimate any parental combination from the information provided by the mixed model analysis, even for very large data sets. Using two recursive steps, for simulated dairy cattle populations with random mating of top cows and bulls and no large full-sib families, the genetic

improvement was 5.8% to 11% of the parental combination standard deviation. This is a lower bound for real dairy cattle populations which include full-sibs from ET and larger numbers of ¾-sibs.

## Finite Polygenic Model

After examining the justification for including dominance effects in genetic evaluations and some of the methods used to estimate it, assumptions about the genetic model and model of analysis must also be considered. A quantitative trait that is influenced by a large number of genes, each with a small effect, is said to exhibit polygenic inheritance. The genotypic value of a polygenic trait can be modeled as a continuous distribution (usually normal) by assuming an infinite number of loci, as in the infinitesimal model, or as a discrete distribution by assuming a finite number *k* of unlinked polygenic loci, as in the finite polygenic model (FPM). The traditional infinitesimal model may not be the most realistic model for the distribution of gene effects. A model where few genes with a large effect and several genes with small effects control a quantitative trait may be closer to the real nature of the distribution of gene effects. Also, although the infinitesimal model has appealing statistical properties, it often leads to computationally inefficient analyses, not only for the estimation of dominance variance under inbreeding, but also for the mixed model of inheritance with a normally distributed polygenic component, the threshold model with a normally

distributed underlying genetic variable, and multi-breed models with normally distributed genotypic values (Stricker and Fernando, 1998).

Finite polygenic models have been used by Thompson and Skolnick (1977), who proposed a FPM assuming different numbers of polygenic loci for estimating the heritability of longevity in complex human pedigrees, and by Fernando *et al*. (1994) and Stricker *et al*. (1995a), who were able to accurately estimate major gene effects and polygenic heritability for simulated pedigrees under the mixed model of inheritance, with the latter study also including a linked marker locus. In Fernando *et al*. (1994), the likelihood under a five-loci FPM with $p = 0.5$ was closer to the exact likelihood computed for 500 2-generation pedigrees of 6000 total individuals than were the likelihoods under the infinitesimal model from either an approximate likelihood analysis by the method of Hasstedt (1991) or an analysis under a regressive model (Bonney 1984, 1992). The FPM linkage analysis by Stricker *et al*. (1995a), using the Segregation and Linkage Analysis for Pedigrees (SALP) computer package (Stricker *et al*. 1994), of five 4021-member pedigrees with a major QTL, a marker locus, and 40 residual QTL gave estimates of homozygote difference, allele frequency at both the major and marker loci, recombination fraction, and major locus, polygenic, and residual variances closer to simulated values than the approximate analysis under the infinitesimal model by the approach of Hasstedt (1982) using the Pedigree Analysis Package, Rev. 3 (PAP) by Hasstedt (1989). The additive FPM results in a computationally more simple likelihood than the infinitesimal model, in particular under the mixed model of inheritance containing both a major quantitative trait locus (QTL) and a polygenic component. The

difference between the FPMs in these investigations is in the sampling of offspring alleles from parents. Thompson and Skolnick (1977) sample *n* alleles without replacement from the 2*n* alleles of each parent, which does not follow laws of Mendelian inheritance. Both Fernando *et al*. (1994) and Stricker *et al*. (1995a) sample alleles according to Mendelian inheritance, which decrees that the offspring receive one allele at each locus from each parent.

Despite the different assumptions regarding transmission of alleles from parents to offspring, both of these FPMs share common properties which contrast with the infinitesimal model (IM). In both FPMs, the Mendelian property of the genotype of an individual being conditionally independent of sibs and ancestors given the genotypes of its parents does not hold, but must be assumed to compute the likelihoods. However, because of the assumption of an infinite number of loci, this property does hold true for the infinitesimal model. Another difference is that unlike the IM, in which the genotypic distribution is always symmetric, the FPM results in asymmetric genotypic distributions for allele frequencies other than 0.5.

The assumption of the infinitesimal model induces major theoretical and computational complications in the effort to estimate dominance variance, as witnessed by previous investigations. This is especially true in the presence of inbreeding, where the additional genetic parameters of dominance variance and covariance between additive and dominance effects in a completely inbred population with allele frequencies equal to those in the base population, and the effect of inbreeding depression must be added to the model (Gillois 1964; Harris 1964; DeBoer and Hoeschele 1993). These complications to

the genetic covariance structure due to dominance and inbreeding under the infinitesimal model motivated investigation of an alternative model, the FPM.

## Aim of this thesis

The main objective of this thesis is to address the issue of the finite polygenic model as an alternative to the infinitesimal model, especially with regard to its ability to provide improved estimates of additive variance under a wide range of genetic models.

First the ability of the FPM to estimate the narrow-sense heritability in the absence of a major gene for data simulated under various additive FPMs and an infinitesimal model, either with or without selection, was investigated. Analysis was performed by additive FPMs with varying numbers of loci and two methods: a deterministic Maximum Likelihood (DML) method implemented by the Segregation and Linkage Analysis for Pedigrees package (SALP) by Stricker *et al.* (1994), and a Bayesian approach implemented via Gibbs sampling (BGS) in a Fortran 90 program written by Gage-Lahti. The DML analysis, because of its deterministic nature, is forced to cut inbreeding and marriage loops in the pedigrees, while the sampling-based algorithm is not. Hence, the need to investigate presence versus absence of selection. Comparisons of the accuracy of the two methods' heritability estimates are then drawn for various combinations of simulation models and analysis models in the absence or presence of selection.

The initial section of this thesis addresses the issue of additive genetic models only, with analyses by a FPM as described in Fernando *et al*. (1994), where genotypes

with the same number of “positive” alleles have the same genotypic value and are treated as a single genotype called polygenic number. The next goal was to extend the focus of the investigation to dominance models. The BGS Fortran 90 program was therefore modified by Hoeschele and Du to sample genotypes at individual loci rather than sample polygenic numbers and to fit separate homozygote differences and dominance deviations in the FPM for each biallelic locus, in order to accomplish this objective.

Based on evidence of the existence of dominance variance in traits of economic importance, its potential to increase accuracy of additive breeding values and accommodate mate selection schemes in genetic evaluations, and the lack of an accurate, efficient method to estimate dominance variance, this investigation into the ability of additive and nonadditive finite polygenic models to estimate additive and dominance variance under a wide range of genetic models can be a potentially important contribution to the eventual estimation of nonadditive genetic variances like dominance and additive-by-additive epistasis in large, complex pedigrees with inbreeding, and the eventual inclusion of dominance effects in genetic evaluations of livestock.

# Estimation of Variance Components in Finite Polygenic Models and Complex Pedigrees

# SUMMARY

The finite polygenic model (FPM) contains a finite number of unlinked polygenic loci. In this study, phenotypic data simulated under either the infinitesimal model or under various FPMs containing five or forty loci with equal homozygote difference or eighteen loci with diminishing homozygote difference and different levels of dominance deviation were analyzed with three-, four-, or five-loci FPMs. Data simulated under additive gene action were analyzed with additive FPMs using both a deterministic Maximum Likelihood (DML) method and a Markov chain Monte Carlo (MCMC) Bayesian approach implemented with a Gibbs sampler (BGS). With the exception of the three-loci BGS analysis across all genetic models and the three- and four-loci DML analysis for the infinitesimal and the forty-equal-polygenic-loci models, accurate estimates of heritability (0.48 to 0.50 with true values of $h^2 = 0.50$) and variance components (48.1 to 50.9 for true values of $\sigma_a^2 = 50.0$ and 49.8 to 51.9 for true values of $\sigma_e^2 = 50.0$) were obtained. When both analysis methods were applied to equivalent data sets on populations that had undergone selection, the DML method produced clearly more biased estimates of additive genetic variation (59.2 to 66.6) and heritability (0.53 to 0.57) across all four genetic models due to its use of pedigree loop cutting, relative to the BGS analysis, which produced overestimates for additive genetic variation (58.8 and 55.0) and heritability (0.53 and 0.52) for the infinitesimal and 40-loci FPM genetic models only. Analyzing these additive data sets with an altered BGS method, assuming a

five-loci FPM including both additive and dominance effects, resulted in accurate estimates of additive genetic variance (50.8 to 52.2 for true $\sigma_a^2$ = 50.0) and no significant dominance variance (3.7 to 3.9) being detected where none existed. Preliminary results from analyzing data simulated with both additive and dominance effects using the BGS analysis, although not shown in this thesis, are encouraging.

# INTRODUCTION

A finite polygenic model (FPM) was first proposed by Thompson and Skolnick (1977) for estimating the heritability of longevity in complex human pedigrees. In contrast to the infinitesimal model, in which a trait is assumed to be influenced by a normally-distributed polygenic component consisting of an infinite number of additive genes of small effect, a FPM has a finite number *k* of unlinked polygenic loci. The additive FPM results in a computationally more simple likelihood than the infinitesimal model, in particular under the mixed model of inheritance containing both a major quantitative trait locus (QTL) and a polygenic component (Fernando *et al.* 1994).

The mixed models of inheritance of Fernando *et al.* (1994) and Stricker *et al.* (1995a), which contained finite polygenic loci along with a segregating major QTL, appeared to accurately estimate major gene effects and polygenic heritability. In this contribution, the ability of the FPM to estimate the narrow-sense heritability is investigated in more detail and in the absence of a major gene. Data simulated under various additive FPMs and an infinitesimal model, all either with or without selection, were analyzed with additive FPMs assuming different numbers of polygenic loci as suggested by Thompson and Skolnick (1977). The analysis was performed using two methods: a deterministic Maximum Likelihood method implemented by the Segregation and Linkage Analysis for Pedigrees package (SALP) by Stricker *et al.* (1994), and a

Bayesian approach implemented via Gibbs sampling in a Fortran 90 program written by Gage-Lahti.

Accurate estimation of dominance variation under the infinitesimal model is difficult for theoretical and computational reasons, especially when inbreeding is present. Under inbreeding, with no epistasis, there are additional genetic parameters other than the additive and dominance variance in the base population. These extra parameters include dominance variance and covariance between additive and dominance effects in a completely inbred population with allele frequencies equal to those in the base population, and the effect of inbreeding depression (Gillois 1964; Harris 1964; DeBoer and Hoeschele 1993). These complications to the genetic covariance structure due to dominance and inbreeding under the infinitesimal model motivated investigation of an alternative model, the FPM. The BGS analysis was therefore modified by Hoeschele and Du to sample genotypes at individual loci rather than sample polygenic numbers, and to fit separate homozygote differences and dominance deviations in the FPM for each biallelic locus, ultimately allowing estimation of both the additive and non-additive variances across loci, although only additive variance estimation is examined in this thesis.

# METHODOLOGY

## Additive finite polygenic model with polygenic numbers

In the additive FPM, as outlined in Fernando *et al.* (1994) but without the major gene, there are two alleles at each polygenic locus with equal, fixed gene frequency of $p$ = 0.5 and equal but unknown homozygote difference $a$. The polygenic number $v$ is the total number of favorable alleles possessed by an individual across all loci, which is symmetrically distributed at a gene frequency of 0.5. There are $2k+1$ possible polygenic numbers with values from 0 to $2k$. The $v$ of an individual is assumed to be conditionally independent of the $v$ of any ancestor or sibling given the $v$'s of the parents, although under Mendelian inheritance only approximate conditional independence holds. Occasionally the polygenic numbers of full-sibs are not conditionally independent given the polygenic numbers of their parents, as shown in a counter-example in Fernando *et al.* (1994). The transition probability that a non-base (non- founder) individual will have polygenic number $v_i$ given the polygenic numbers of its sire ($v_{s_i}$) and dam ($v_{d_i}$) is denoted $f(v_i|v_{s_i}, v_{d_i})$, and is calculated recursively according to the formula in Appendix C of Fernando *et al.* (1994). The probability $f(v_i)$ that a base (founder) individual will have polygenic number $v_i$ is equal to the population frequency of $v_i$. Conditional on **v**, the linear model for phenotype (assuming single records) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{v} - \mathbf{1}k)a + \mathbf{e}, \qquad (2)$$

where **y** is a vector of phenotypes, **X** is a known incidence matrix relating observations in **y** to fixed effects in $\boldsymbol{\beta}$, $\boldsymbol{\beta}$ is a vector of fixed effects, **v** is a vector of polygenic numbers, *k* is the number of biallelic loci, *a* is the homozygote difference which is constant across loci, and **e** is the vector of error terms.

**(a) Deterministic ML**

For the deterministic ML method, the likelihood of the data given the parameters is

$$f(\mathbf{y}|\boldsymbol{\beta},a,\sigma_e^2) = \sum_{\mathbf{v}}\left[\prod_{i=1}^{n_b} f(v_i)\, f(y_i|v_i,\boldsymbol{\beta},a,\sigma_e^2) \prod_{i=n_b+1}^{n} f(v_i|v_{s_i},v_{d_i})\, f(y_i|v_i,\boldsymbol{\beta},a,\sigma_e^2)\right], \qquad (3)$$

where $n_b$ is the number of base animals in the population, $n$ is the total number of animals, $f(y_i|v_i,\boldsymbol{\beta},a,\sigma_e^2)$ is the penetrance function of *i*'s phenotype given *i*'s polygenic number and the parameters, and $f(v_i|v_{s_i},v_{d_i})$ and $f(v_i)$ are the transition probabilities mentioned previously. The summation is over all possible vectors **v** of polygenic numbers. The likelihood is maximized with respect to $\boldsymbol{\beta}$, *a*, and $\sigma_e^2$.

To maximize the likelihood in (3), the deterministic ML method implemented in SALP uses the Downhill Simplex algorithm of Nelder and Mead (1965) and cuts pedigree loops according to the method of Stricker *et al.* (1995b), in which an additional founder with an identical phenotype is inserted at the site of the cutting, resulting in an approximate likelihood. After loops are cut, the likelihood of the data is computed using the recursive algorithm of Fernando *et al.* (1993):

$$f(\mathbf{y}|\boldsymbol{\beta},a,\sigma_e^2) = \prod_{pedigrees} \sum_{v_i} \left[ \alpha_i(v_i)\, f(y_i|\boldsymbol{\beta},a,\sigma_e^2,v_i) \prod_{j \in S_i} p_{i,j}(v_i) \right]. \quad (4)$$

A single individual *i* per pedigree is selected and the anterior probability $\alpha_i(v_i)$ of individual *i* having polygenic number $v_i$ is multiplied by the penetrance function $f(y_i|\boldsymbol{\beta},a,\sigma_e^2,v_i)$ and the product over the set of mates $S_i$ of the posterior probability of *i* through mate *j*. A pedigree contains all individuals that are connected in any way to *i* through *i*'s parents, full-sibs, mates, and offspring. If a population consists of several disjoint pedigrees, a product is taken across pedigrees to give the likelihood. An anterior probability incorporates information from ancestors through parents and full-sibs, while a posterior probability includes descendant information through mates and offspring. This algorithm requires that all members of the pedigree other than *i* be *either* anterior or posterior to *i*; therefore, pedigree loops must be cut, as they result in some members of the loop being *both* anterior and posterior to *i*. After SALP converged the first time, it was restarted with those estimates to ensure that convergence to the global maximum had occurred.

**(b) Bayesian Gibbs sampler method**

For the Bayesian method, parameter estimates consisted of the means of the posterior distribution of the parameters estimated as averages of samples generated using the Gibbs sampler. The joint posterior density of the parameters and polygenic numbers is proportional to the probability density of the data given the polygenic numbers and parameters multiplied by the priors of the parameters and *v*'s as shown in

$$f(\boldsymbol{\beta}, a, \sigma_e^2, \mathbf{v} \mid \mathbf{y}) \propto \prod_{i=1}^{n} f(y_i \mid v_i, \boldsymbol{\beta}, a, \sigma_e^2) \prod_{i=1}^{n_b} f(v_i) \prod_{i=n_b+1}^{n} f(v_i \mid v_{s_i}, v_{d_i}) f(\boldsymbol{\beta}) f(a) f(\sigma_e^2). \tag{5}$$

Flat (uninformative) priors were assumed for $\boldsymbol{\beta}$, $a$, and $\sigma_e^2$ to indicate a lack of prior knowledge about the parameters and allow the data to have primary influence, as in non-Bayesian analyses. The effects in $\boldsymbol{\beta}$ and $a$ were sampled from fully conditional, normal distributions and the error variance was sampled from an inverse-$\chi^2$ distribution derived as in Wang *et al*. (1993). The polygenic number of individual $i$ was sampled from a multinomial distribution conditional on the phenotype of $i$, the parameters, $\boldsymbol{\beta}$, $a$, $\sigma_e^2$, and the polygenic numbers of the sire ($v_{s_i}$), dam ($v_{d_i}$), mates ($v_{m_i}$), and offspring ($v_{o_i}$) of $i$. The sampling distribution of $v_i$ has probability density

$$f(v_i \mid \mathbf{y}, \boldsymbol{\beta}, a, \sigma_e^2, \mathbf{v}) \propto f(y_i \mid v_i, \boldsymbol{\beta}, a, \sigma_e^2) f(v_i \mid v_{s_i}, v_{d_i}) \prod_{o_i} f(v_{o_i} \mid v_i, v_{m_i}), \tag{6}$$

which is proportional to the probability density of $i$'s phenotype given $i$'s polygenic number and the parameters, multiplied by the transition probability of $i$ inheriting polygenic number $v_i$ given the polygenic numbers of its sire and dam, and further multiplied by the product over offspring of an offspring's transition probability given the polygenic numbers of its parents (individual $i$ and one of $i$'s mates). If individual $i$ is a base animal, its transition probability is replaced by the equilibrium frequency of $v_i$ in the base population. Despite large numbers of offspring for some individuals, the univariate sampling distribution for the polygenic number in (6) allows fast mixing of the sampler, unlike the slow mixing which occurs due to large numbers of offspring for the sampling

of genotypes (Janss *et al.* 1995). The 2*k*+1 polygenic numbers are more finely discretized than genotypes at individual loci with many different genotype combinations having the same polygenic number, causing more rapid mixing. Additive genetic variance ($\sigma_a^2$) was calculated as $\frac{pka^2}{2(1-p)}$ and narrow-sense heritability ($h^2$) for an additive model as $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$ in each Gibbs cycle using that cycle's samples of *a* and $\sigma_e^2$, with allele frequency $p = 0.5$.

**Nonadditive finite polygenic model with genotypes at individual loci**

To generalize the FPM to also incorporate dominance variation, the model becomes

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}_\mathbf{G}\,\boldsymbol{a} + \mathbf{W}_\mathbf{G}\,\boldsymbol{d} + \mathbf{e}, \quad (7)$$

where **y**, **X**, **β** and **e** are defined as before, ***a*** is a vector of homozygote differences at *k* biallelic loci, **d** is a vector of dominance deviations at *k* biallelic loci, $\mathbf{Z}_\mathbf{G}$ is a known design matrix with *k* columns containing coefficients of -1, 0, and 1 corresponding to the three genotypes at a biallelic locus, and $\mathbf{W}_\mathbf{G}$ is a known design matrix with *k* columns containing coefficients of 1 for the heterozygous genotype and 0 for the homozygous genotypes. Both $\mathbf{Z}_\mathbf{G}$ and $\mathbf{W}_\mathbf{G}$ are based on knowing **G,** a matrix of genotypes of all individuals at all *k* biallelic loci. The unknowns now include the parameters in ***a*** and **d**, $\sigma_e^2$, and the *nk* genotypes in **G**, instead of the *n* polygenic numbers in **v**. The joint posterior of all unknowns is now

$$f(\boldsymbol{\beta}, \boldsymbol{a}, \mathbf{d}, \boldsymbol{\sigma}_e^2, \mathbf{G} \mid \mathbf{y}) \propto \prod_{i=1}^{n} f(y_i \mid \boldsymbol{\beta}, \boldsymbol{a}, \mathbf{d}, \boldsymbol{\sigma}_e^2, \mathbf{G}_i) \prod_{i=1}^{n_b} \prod_{j=1}^{k} f(g_{ij}) \prod_{i=n_b+1}^{n} \prod_{j=1}^{k} f(g_{ij} \mid g_{s_i j}, g_{d_i j}) \cdot$$

$$f(\boldsymbol{\beta}) f(a) f(d) f(\sigma_e^2), \qquad (8)$$

where $\mathbf{G}_i$ is row *i* of **G** and $\mathbf{G}_i = \{g_{ij}\}$, with $g_{ij}$ being the genotype of individual *i* at locus *j*. Parameters in ***a*** and **d** are sampled as before, from fully-conditional normal distributions, but are not sampled jointly. Once again, flat priors are assumed for $\boldsymbol{\beta}$, *a*, *d*, and $\sigma_e^2$. Genotypes of individuals at single loci are also sampled from standard, fully-conditional distributions, except that the genotypes of a parent and its final offspring at a given locus are sampled jointly as in Janss *et al*. (1995).

**Data and analysis**

The simulated population structure consisted of *n* = 6300 individuals over one base generation and three discrete offspring generations. Every generation, 50 males and 250 females were randomly chosen, with each male randomly mated to 5 females. Females produced eight-offspring litters divided equally in half by sex, giving each sire forty progeny. Unintentional inbreeding and the mating structure resulted in 945 cuts of inbreeding and marriage loops under the DML analysis for the pedigree of the unselected population. The phenotypic trait was assumed to have been recorded on every individual, and the pedigree structure was replicated twenty times with different phenotypes assigned. Under selection, the size and structure of the pedigree was maintained. In each generation, the 50 males and 250 females with the highest breeding values as estimated by Best Linear Unbiased Prediction (BLUP) with the animal model program

JAA (Misztal, 1989) were selected to be parents. This scheme results in a selection intensity of five percent for males and twenty-five percent for females. Unlike the unselected pedigree structure which was identical across replicates, the selected pedigrees are different across replicates, due to different animals being selected, and there were on average 985 cuts made to pedigree loops.

Phenotypic data were simulated under various additive models with true values of $\sigma_e^2 = 50$ and $\sigma_a^2 = 50$ for residual and additive genetic variance (across loci), respectively, and a mean ($\mu$) equal to zero as the only element of $\beta$. Model 1 was the infinitesimal model, while Models 2,3, and 4 were FPMs of either 40 loci of equal effect, 18 loci of diminishing effect, or 5 loci of equal effect, respectively, with all loci being biallelic and $p = 0.5$ (see Table 4). The eighteen loci in Model 3 included one locus with variance 25, two loci with variance 5, five loci with variance 2, and ten loci with variance 0.5. The polygenic component of the phenotype for the infinitesimal model was generated in the usual manner by multiplying a standard normal deviate (IMSL STAT/Library: FORTRAN, 1991) by the additive genetic standard deviation, while for FPMs assuming $p = 0.5$, the sum of the genotypic values across loci comprised the polygenic component.

**Table 4. Four genetic models used for simulation of data and three analysis models used for each of the genetic models.**

| Simulation Models | Analysis Models |
|---|---|
| 1. Infinitesimal model | 3-loci FPM |
| 2. FPM - 40 loci of equal effect | 4-loci FPM |
| 3. FPM - 18 loci of diminishing effect | 5-loci FPM |
| 4. FPM - 5 loci of equal effect | |

For analyses under additive gene action, data sets from the four genetic models were all analyzed by both the deterministic ML and the BGS methods, all using additive FPMs of three, four, and five loci (see Table 4). Starting values of $\mu = 5$, $\sigma_e^2 = 65$, and $\sigma_a^2 = 35$ were arbitrarily chosen, and the starting homozygote difference *a* was calculated as $2\sqrt{\frac{\sigma_a^2(1-p)}{2pk}}$, resulting in *a* = 4.83, 4.18, or 3.74 for the three different analysis models. Because of biased parameter estimates from the three- and four-loci FPM analyses of the unselected data sets, as will be seen in the results, the selected data sets were analyzed by FPMs containing 5 loci only. Ten additive Model 4 data sets were analyzed with a nonadditive FPM of five loci, fitting constant or variable homozygote differences and dominance deviations across loci, using arbitrary starting values of $\mu = 0$, $\sigma_e^2 = 10$, a = 2, and d = 0.

# RESULTS

## Additive genetic FPM—unselected populations

For analysis under the additive genetic model, chains of 10,000 Gibbs cycles with 50 burn-in cycles were found to be sufficient based on the Monte Carlo standard errors averaged across ten replicates of unselected Model 4 data sets as shown in Table 5. These were calculated as suggested by Geyer (1992) using the square root of the variance of a sample mean found in Sorensen *et al.* (1995).

**Table 5. Monte Carlo standard errors (and standard errors) averaged across ten replicates of the unselected Model 4 data sets analyzed under an additive FPM.**

| # Loci in Analysis | $\mu$ | $\sigma_e^2$ | $\sigma_a^2$ | $h^2$ |
|---|---|---|---|---|
| 3 loci | 0.01 (0.0003) | 0.08 (0.002) | 0.12 (0.003) | 0.0008 (0.00003) |
| 4 loci | 0.05 (0.0020) | 0.13 (0.009) | 0.22 (0.015) | 0.0020 (0.00060) |
| 5 loci | 0.06 (0.0040) | 0.12 (0.006) | 0.21 (0.010) | 0.0020 (0.00020) |

True values are $\mu = 0.0$, $\sigma_e^2 = 50.0$, $\sigma_a^2 = 50.0$, and $h^2 = 0.50$.

Posterior mean estimates of the parameters of interest and empirical standard errors for the unselected populations are given in Tables 6, 7, and 8 for the three-, four-, and five-loci FPMs, respectively. When analyzed under the three-loci FPM, the BGS approach significantly underestimated $\mu$ and overestimated additive genetic variance and

heritability for all four of the simulation models (see Table 6), presumably due to overdiscretization of the polygenic effects. However, by changing the analysis model to four- or five-loci FPMs, this method produced accurate estimates. Tested with a two-sided *t*-test at the 0.05 significance level, SALP significantly underestimated heritability and additive genetic variance when using the three- or four-loci models to analyze the infinitesimal model or the forty-loci FPM, but accurately estimated the parameters for all other analysis models across all other simulation models. Once again, the biased estimates of heritability probably occurred due to the assumption of too few alleles in a FPM attempting to analyze a genetic model with either forty or an infinite number of loci. The approximation to the likelihood which results from cutting pedigree loops does not seem to affect the accuracy of parameter estimation in SALP for unselected populations. In Table 8, when the 5-loci FPM was used as both the simulation and the analysis model, the homozygote difference *a* was estimated accurately with both SALP and BGS.

**Table 6. Average parameter estimates (and standard errors) obtained from twenty replicates of a 3-loci FPM analysis with either the Bayesian Gibbs Sampler (BGS) approach or SALP of data simulated without selection under genetic models with different numbers of loci (see Table 4).**

| | | Genetic Models | | | |
|---|---|---|---|---|---|
| | | Infinitesimal | 40 Equal | 18 Diminishing | 5 Equal |
| BGS | $\mu$ | -8.61 (0.16) | -8.65 (0.22) | -9.08 (0.19) | -9.07 (0.08) |
| | $\sigma_e^2$ | 52.8 (0.53) | 52.2 (0.40) | 51.6 (0.42) | 52.3 (0.57) |
| | $\sigma_a^2$ | 60.5 (0.72) | 61.1 (0.85) | 63.8 (0.85) | 62.1 (0.95) |
| | $h^2$ | 0.53 (0.004) | 0.54 (0.004) | 0.55 (0.004) | 0.54 (0.004) |
| SALP | $\mu$ | 0.09 (0.22) | 0.05 (0.19) | -0.18 (0.19) | -0.08 (0.12) |
| | $\sigma_e^2$ | 51.6 (0.67) | 51.9 (0.93) | 50.3 (0.84) | 49.9 (0.65) |
| | $\sigma_a^2$ | 47.4 (1.09) | 46.7 (1.46) | 50.9 (1.51) | 49.4 (1.31) |
| | $h^2$ | 0.48 (0.01) | 0.47 (0.01) | 0.50 (0.01) | 0.50 (0.01) |

True values are $\mu = 0.0$, $\sigma_e^2 = 50.0$, $\sigma_a^2 = 50.0$, and $h^2 = 0.50$.

**Table 7. Average parameter estimates (and standard errors) obtained from twenty replicates of a 4-loci FPM analysis with either the Bayesian Gibbs Sampler (BGS) approach or SALP of data simulated without selection under genetic models with different numbers of loci (see Table 4).**

| | | Genetic Models | | | |
|---|---|---|---|---|---|
| | | Infinitesimal | 40 Equal | 18 Diminishing | 5 Equal |
| BGS | $\mu$ | 0.05 (0.10) | 0.12 (0.12) | 0.05 (0.12) | -0.06 (0.07) |
| | $\sigma_e^2$ | 51.0 (0.48) | 50.9 (0.43) | 50.1 (0.52) | 49.8 (0.43) |
| | $\sigma_a^2$ | 48.2 (0.70) | 48.3 (0.89) | 50.1 (0.91) | 50.6 (0.83) |
| | $h^2$ | 0.49 (0.004) | 0.49 (0.01) | 0.50 (0.01) | 0.50 (0.004) |
| SALP | $\mu$ | 0.05 (0.21) | 0.13 (0.21) | -0.12 (0.20) | -0.16 (0.15) |
| | $\sigma_e^2$ | 51.5 (0.60) | 51.9 (0.73) | 50.7 (0.84) | 50.6 (0.79) |
| | $\sigma_a^2$ | 46.8 (1.07) | 46.9 (1.34) | 49.9 (1.33) | 49.2 (1.16) |
| | $h^2$ | 0.48 (0.01) | 0.47 (0.01) | 0.50 (0.01) | 0.49 (0.01) |

True values are $\mu = 0.0$, $\sigma_e^2 = 50.0$, $\sigma_a^2 = 50.0$, and $h^2 = 0.50$.

**Table 8. Average parameter estimates (and standard errors) obtained from twenty replicates of a 5-loci FPM analysis with either the Bayesian Gibbs Sampler (BGS) approach or SALP of data simulated without selection under genetic models with different numbers of loci (see Table 4).**

| | | Genetic Models | | | |
|---|---|---|---|---|---|
| | | Infinitesimal | 40 Equal | 18 Diminishing | 5 Equal |
| BGS | $\mu$ | 0.88 (0.12) | 0.08 (0.14) | -0.05 (0.12) | 0.11 (0.10) |
| | $\sigma_e^2$ | 51.7 (0.51) | 51.5 (0.40) | 50.7 (0.57) | 51.2 (0.62) |
| | $\sigma_a^2$ | 48.1 (0.67) | 48.3 (0.78) | 50.3 (0.96) | 49.1 (0.88) |
| | $h^2$ | 0.48 (0.01) | 0.48 (0.01) | 0.50 (0.01) | 0.49 (0.01) |
| | *a* | | | | 4.44 (0.04) |
| SALP | $\mu$ | 0.11 (0.20) | 0.34 (0.23) | -0.14 (0.20) | -0.18 (0.13) |
| | $\sigma_e^2$ | 50.7 (0.79) | 49.9 (0.73) | 50.8 (0.68) | 51.3 (0.74) |
| | $\sigma_a^2$ | 48.1 (1.09) | 48.8 (1.31) | 49.2 (1.22) | 48.2 (1.44) |
| | $h^2$ | 0.49 (0.01) | 0.49 (0.01) | 0.49 (0.01) | 0.48 (0.01) |
| | *a* | | | | 4.38 (0.06) |

True values are $\mu = 0.0$, $\sigma_e^2 = 50.0$, $\sigma_a^2 = 50.0$, $h^2 = 0.50$, and $a = 4.48$.

**Additive genetic FPM—selected populations**

Table 9 gives parameter estimates under selection, with the 5-loci additive FPM as the analysis model. The BGS method produced accurate parameter estimates for the two genetic models with the fewest numbers of loci, Models 3 and 4. SALP expectedly produced biased estimates for all four genetic models due to its practice of cutting inbreeding and marriage loops, and SALP also greatly overestimated the mean. There appeared to be a slight trend with both methods of increasingly overestimating genetic variance as the number of loci in the genetic model increased. Therefore, the selected data sets were reanalyzed using a 20-loci FPM for the BGS approach and a 14-loci FPM for SALP (14 was the maximum number of loci fitted by SALP). However, increasing the number of loci in the analysis model did not significantly improve the SALP estimates of either genetic variance or the mean, and surprisingly the Bayesian estimates were virtually unchanged as well (results not shown). This overestimation of additive genetic variance by SALP was also seen in Hagger and Stricker's (1998) FPM analysis of egg weights from a multi-generation selection experiment with chickens. When a loop is cut and an individual is replaced by a founder with the same phenotype, that pseudo-founder has a higher prior probability of having a "poor" genotype at each locus than the original individual. As a consequence, the overall mean may be biased upwards, and additive genetic variance increased.

**Table 9. Average parameter estimates (and standard errors) obtained from twenty replicates of a 5-loci FPM analysis with either the Bayesian Gibbs Sampler (BGS) approach or SALP of data simulated with selection under genetic models with different numbers of loci.**

| | | Genetic Models | | | |
|---|---|---|---|---|---|
| | | Infinitesimal | 40 Equal | 18 Diminishing | 5 Equal |
| BGS | $\mu$ | 0.98 (0.14) | 0.63 (0.07) | -0.16 (0.10) | 0.12 (0.14) |
| | $\sigma_e^2$ | 52.9 (0.28) | 51.8 (0.31) | 48.8 (0.30) | 49.7 (0.67) |
| | $\sigma_a^2$ | 58.8 (0.67) | 55.0 (0.68) | 48.7 (0.66) | 51.2 (0.76) |
| | $h^2$ | 0.53 (0.004) | 0.52 (0.004) | 0.50 (0.01) | 0.51 (0.01) |
| | *a* | | | | 4.52 (0.04) |
| SALP | $\mu$ | 9.09 (0.19) | 8.92 (0.14) | 8.28 (0.13) | 8.51 (0.19) |
| | $\sigma_e^2$ | 50.4 (0.47) | 51.0 (0.48) | 48.9 (0.56) | 47.3 (0.94) |
| | $\sigma_a^2$ | 66.6 (1.04) | 60.9 (0.88) | 55.4 (0.98) | 59.2 (1.05) |
| | $h^2$ | 0.57 (0.01) | 0.56 (0.02) | 0.53 (0.01) | 0.56 (0.01) |
| | *a* | | | | 4.86 (0.04) |

True values are $\mu = 0.0$, $\sigma_e^2 = 50.0$, $\sigma_a^2 = 50.0$, $h^2 = 0.50$, and $a = 4.48$.

**Nonadditive FPM**

As a preliminary step to investigate the eventual estimation of both additive and dominance variance, data sets simulated under an additive, five-loci FPM were analyzed with the BGS method extended to sample genotypes at individual loci and fit both homozygote differences and dominance deviations at individual loci. In these analyses, the *a*'s were either restricted to be constant across loci or sampled freely, and the *d*'s were either absent or sampled freely. As the estimates of the homozygote differences were quite variable across loci when not restricted to being constant (see Table 10 for average high and low *a*'s across ten data sets), accuracy of breeding value estimation was also investigated. At the assumed gene frequencies of $p = q = 0.5$, breeding values (BV's) simplify to genotypic values summed across all loci. The correlation between true BV's and estimated BV's was 0.81 for constant *a* and 0.80 for variable *a*'s, for a 5-loci genetic model simulated with constant *a*'s and analyzed with a 5-loci additive FPM (Model 4). Hence, it appears that constancy versus variability of homozygote differences across loci has little effect on accuracy of estimated genetic merits. The results in Table 10 indicate that the constant homozygote difference *a* and additive genetic variance are accurately estimated, and that when no dominance variation exists, the analysis correctly estimates a dominance variance near zero.

**Table 10. Parameter estimates (and standard errors) averaged across ten replicates of a 5-loci FPM Bayesian Gibbs sampler (BGS) analysis with various combinations of additive and dominance effects for data simulated without selection under Model 4 (see Table 4).**

| | Analysis Models | | | |
|---|---|---|---|---|
| Additive effects | One | Five | One | Five |
| Dominance effects | None | None | Five | Five |
| $\mu$ | 0.01 (0.07) | 0.04 (0.08) | 0.09 (0.08) | 0.05 (0.09) |
| average *a* | 4.56 (0.05) | 2.40 (0.41) | 4.50 (0.06) | 2.74 (0.39) |
| high *a* | _____ | 5.23 (0.19) | _____ | 5.59 (0.13) |
| low *a* | _____ | -1.29 (0.85) | _____ | -1.02 (0.73) |
| average *d* | _____ | _____ | 0.05 (0.12) | 0.08 (0.12) |
| high *d* | _____ | _____ | 1.09 (0.27) | 1.03 (0.22) |
| low *d* | _____ | _____ | -0.79 (0.15) | -0.98 (0.17) |
| $\sigma_a^2$ | 52.0 (1.15) | 52.2 (1.27) | 50.8 (1.22) | 51.1 (1.25) |
| $\sigma_d^2$ | _____ | _____ | 3.7 (0.43) | 3.9 (0.39) |
| $\sigma_e^2$ | 49.7 (0.50) | 49.6 (0.55) | 47.0 (0.71) | 46.8 (0.70) |
| $h^2$ | 0.51 (0.008) | 0.51 (0.008) | 0.52 (0.009) | 0.52 (0.009) |

True values are $\mu = 0.0$, $a = 4.48$, $d = 0.0$ for all loci, $\sigma_a^2 = 50.0$, $\sigma_d^2 = 0.0$, $\sigma_e^2 = 50.0$, and $h^2 = 0.50$.

Further estimation of dominance variance with the BGS method when dominance variance does exist is in the process of being studied by Du. Preliminary estimates of dominance variance compare favorably to those from the MCMC Bayesian program of Pong-Wong *et al*. (1998) and with REML estimates from the SAS Proc Mixed procedure (SAS System, Release 6.09, 1993) for comparable pedigree structures.

# DISCUSSION

The results obtained in this study show that the FPM gives accurate estimates of narrow-sense heritability under a wide range of additive genetic models. Analyzing under a five-loci additive FPM produces accurate estimates of heritability for data generated under infinitesimal or finite polygenic genetic models with different, e.g. larger, numbers of loci. Three- or four-loci FPMs produced biased estimates for data simulated under some genetic models. Hence, it would be prudent to fit at least 5 loci in the FPM. ML analyses implemented deterministically with loop cutting produces significantly upward-biased estimates of additive genetic variance for populations undergoing selection. Therefore, sampling based algorithms, which do not require loop-cutting, should be preferred. Some upward bias also existed when analyzing selected populations with larger numbers of loci with the BGS method, even when the analysis model included higher numbers of loci, a finding which was unexpected.

The BGS FPM analysis was generalized by sampling genotypes at individual biallelic loci rather than sampling polygenic numbers, by allowing homozygote differences to vary across loci, and by including a separate dominance deviation for each locus. When dominance deviations are fixed at zero and homozygote differences are constant across loci, and when the genotypes of all individuals at all loci are known, this FPM is equivalent to the FPM with polygenic number sampling. However, when genotypes are unknown, the sampling space of the missing data (individual genotypes or

polygenic numbers) differs between the two analyses (one polygenic number can result from several different genotypes), and this could cause some differences in the results. However, accuracy of estimates of heritability and additive genetic variance, as well as Monte Carlo standard errors, were virtually identical.

Further investigations of the FPM for parameter estimation in nonadditive genetic models are warranted, and preliminary studies by Du using this BGS method have given promising estimates of variance components. Joint sampling of mean, homozygote differences, and dominance deviations should be considered to improve high Monte Carlo standard errors due to large posterior correlations among these parameters. Alternative genotype sampling schemes should be investigated, e.g., sampling all individuals at a given locus jointly, sampling several loci jointly, or using modified transmission probabilities as in Goddard (1998). Furthermore, the effect of estimating gene frequencies at the biallelic loci, rather than fixing them at 0.5, on accuracy of estimation of genetic parameters and genetic merits has not yet been evaluated. Finally, the inclusion of additive-by-additive genetic variance in addition to dominance variance should be considered to enable more realistic analyses of actual data.

The FPM has potential for improved estimation of nonadditive genetic variance components due to the severe limitations of analyses under the infinitesimal model. At this time, there is no reliable software available which fully accounts for inbreeding in the presence of dominance variation. Development of such software would allow researchers to compare analyses under FPMs and infinitesimal genetic models for different pedigree

structures and confirm or disprove the expectation of improved accuracy of parameter estimation using FPMs and nonadditive genetic models.

# CONCLUSIONS

The finite polygenic model (FPM) has been found to give accurate estimates of narrow-sense heritability in a 6300-member, five-generation pedigree simulated under additive genetic models ranging from the infinitesimal model (IM) to FPMs containing five or forty loci with equal homozygote difference to a FPM with eighteen loci of diminishing homozygote difference. The number of loci included in the FPM for analysis was found to be important, with five loci being the minimum. These results hold for both deterministic Maximum Likelihood (DML) and Markov chain Monte Carlo (MCMC) Bayesian methods of analysis. However, the DML method was found to be more severely affected by selection due to its practice of cutting inbreeding and marriage loops in the pedigree than was the MCMC Bayesian method, which was only slightly affected by selection for large numbers of loci in the genetic model.

Using the MCMC Bayesian method, fitting a five-loci nonadditive FPM to the above additive data accurately estimated a lack of dominance variance. For data simulated under FPMs containing both additive and dominance effects, preliminary analyses with the BGS method are producing quite accurate estimates of all variance components. Further investigations of the FPM for parameter estimation in nonadditive (including dominance and eventually epistasis) genetic models by MCMC Bayesian analysis should be conducted. The FPM has potential as an alternative to the infinitesimal model for improved estimation of nonadditive genetic variance components in large, complex pedigrees complicated by the existence of inbreeding.

Although genetic evaluations of livestock currently are based only on an additive model, evidence for the existence of significant dominance genetic variance in economically important traits, ever-increasing numbers of dominance relationships in livestock populations due to advances in AI, ET, and cloning technologies, and improved hardware and software to accurately estimate dominance variance efficiently suggest that genetic evaluations incorporating both additive and dominance effects will soon be feasible. Proposed advantages to including dominance variation in livestock genetic evaluations are increased accuracy of additive breeding values and the ability to exploit nonrandom mating strategies. Uimari and Kennedy (1990) showed that after fitting a model including dominance, selecting animals only on the resulting additive effect (which is assumed to be more accurate), produces lower genetic gains than using a model without the dominance effects. Therefore, nonadditive effects should not be included in the model if they are not used in selection. Disseminating information from nonadditive evaluations will also be a challenge in the future, because every sire should technically have one genetic merit for every possible mate, creating a prohibitively large sire catalog.

This proposed method of assuming a finite polygenic model to estimate additive and eventually nonadditive variances by Markov chain Monte Carlo Bayesian analysis should contribute to overcoming the obstacle of including dominance effects in the genetic evaluations of livestock.

# REFERENCES


Allaire, F.R., & Henderson, C.R. (1965). Specific combining abilities among dairy sires. *Journal of Dairy Science* **48**, 1096-1100.

Bonney, G.E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: regressive models. *American Journal of Medical Genetics* **18**, 731-749.

Bonney, G.E. (1992). Compound regressive models for family data. *Human Heredity* **42**, 28-41.

Culbertson, M.S., Misztal, I., Mabry, J.W., Gengler, N., Bertrand, J.K. & Varona, L. (1998). Estimation and utilization of inbreeding and dominance genetic effects in genetic evaluation of purebred American swine. In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production,* Armidale, Australia, **23**, 571-574.

DeBoer, I. J. M. & Hoeschele, I. (1993). Genetic evaluation methods for populations with dominance and inbreeding. *Theoretical and Applied Genetics* **86**, 245-258.

DeStefano, A.L. & Hoeschele, I. (1992). Utilization of dominance variance through mate allocation strategies. *Journal of Dairy Science* **75**, 1680-1690.

Falconer, D.S. & Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*. 4th ed. Addison Wesley Longman Ltd., Essex, England.

Fernando, R., Stricker, C. & Elston, R. C. (1993). An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theoretical and Applied Genetics* **87**, 89-93.

Fernando, R., Stricker, C. & Elston, R. C. (1994). The finite polygenic mixed model: an alternative formulation for the mixed model of inheritance. *Theoretical and Applied Genetics* **88**, 573-580.

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

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science* **7**, 467-511.

Gillois, M. (1964). Calcul de coefficients d'identité d'après les liens de parenté. In: *Communication aux Journées d'études de la Commission de Génétique de la Fédération européenne de Zootechnie*, Lisbon, Portugal.

Goddard, M. E. (1998). Gene based models for genetic evaluation—an alternative to BLUP? In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia, **26**, 33-36.

Hagger, C. & Stricker, C. (1998). Infinitesimal and segregation analyses of a trait under selection. In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia, **26**, 37-40.

Harris, D. L. (1964). Genotypic covariances between inbred relatives. *Genetics* **50**, 1319-1348.

Hasstedt, S.J. (1982). A mixed model approximation for large pedigrees. *Comput. Biomedical Research* **15**, 295-307.

Hasstedt, S.J. (1989). PAP: Pedigree Analysis Package, Rev. 3. Department of Human Genetics, University of Utah, Salt Lake City.

Hasstedt, S.J. (1991). A variance components/major locus likelihood approximation on quantitative data. *Genetic Epidemiology* **8**, 113-125.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226-252.

Hoeschele, I. (1991). Additive and nonadditive genetic variance in female fertility of Holsteins. *Journal of Dairy Science* **74**, 1743-1752.

Hoeschele, I., & Van Raden, P.M. (1991). Rapid inversion of dominance relationship matrices for non-inbred populations by including sire by dam subclass effects. *Journal of Dairy Science* **74**, 557-569.

IMSL®, STAT/Library: FORTRAN, Subroutines for Statistical Analysis version 2. (1991). Visual Numerics, Inc., Houston, Texas.

Janss, L. L. G., Thompson, R. & VanArendonk, J. A. M. (1995). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theoretical and Applied Genetics* **91**, 1137-1147.

Miller, S.P., & Goddard, M.E. (1998). The super-breed approach to modeling additive and non-additive genetic effects between and within breeds in multi-breed evaluations.

In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia, **23**, 97-100.

Miztal, I. (1989). JAA - Mixed model program using iteration on data with support for animal model, version 2.0. Computer program package, Department of Animal Sciences, University of Illinois, Urbana-Champaign.

Misztal, I. (1997) Estimation of variance components with large-scale dominance models. *Journal of Dairy Science* **80**, 965-974.

Misztal, I., Fernando, R.L., Grossman, M., Lawlor, T.J., & Lukaszewicz, M. (1995). Dominance and epistatic effects in genetic evaluation of farm animals. *Animal Science Papers and Reports* **13**, 251-265.

Misztal, I., Varona, L., Culbertson, M.S., Gengler, N., Bertrand, J.K., Mabry, J., Lawlor, T.J., & Van Tassell, C.P. (1998). Studies on the value of incorporating effect of dominance in genetic evaluations of dairy cattle, beef cattle, and swine. In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia, **25**, 513-516.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7**, 308-313.

Patterson, H.D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545.

Pong-Wong, R., Shaw, F., Woolliams, J. A. (1998). Estimation of dominance variation using a finite-locus model. In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia, **26**, 41-44.

Reverter, A., Golden, B.L., & Bourdon, R.M. (1994). Method R variance components procedure: application on the simple breeding value model. *Journal of Animal Science* **72**, 2247-2253.

Rodriguez-Almeida, F.A., Van Vleck, L.D., Willham, R.L. & Northcutt, S. L. (1995). Estimation of non-additive genetic variances in three synthetic lines of beef cattle using an animal model. *Journal of Animal Science* **73**, 1002-1011.

SAS System, Release 6.09. (1993). SAS Institute, Inc., Cary, North Carolina.

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

Stricker, C., Fernando, R. L. & Elston, R. C. (1994). SALP - Segregation and Linkage Analysis for Pedigrees, Release 2.0. Computer program package, Swiss Federal Institute of Technology ETH, Institute of Animal Sciences, Zurich, Switzerland.

Stricker, C., Fernando, R. L. & Elston, R. C. (1995a). Linkage analysis with an alternative formulation for the mixed model of inheritance. *Genetics* **141**, 1651-1656.

Stricker, C., Fernando, R. L. & Elston, R. C. (1995b). Iteration on loops - an improvement for cutting arbitrary pedigrees with loops to calculate the likelihood and genotype probabilities. *Theoretical and Applied Genetics* **91**, 1054-1063.

Stricker, C., & Fernando, R.L. (1998). Some theoretical aspects of finite locus models. In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia, **26**, 25-32.

Tempelman, R.J. & Burnside, E.B. (1990a). Additive and nonadditive genetic variation for production traits in Canadian Holsteins. *Journal of Dairy Science* **73**, 2206-2213.

Tempelman, R.J. & Burnside, E.B. (1990b). Additive and nonadditive genetic variation for conformation traits in Canadian Holsteins. *Journal of Dairy Science* **73**, 2214-2220.

Tempelman, R.J. & Burnside, E.B. (1991). Additive and dominance genetic variation for dairy production traits under an animal model. *Journal of Animal Breeding and Genetics* **108**, 330-342.

Thomas, C.L., Vinson, W.E., Pearson, R.E., & Norman, H.D. (1985). Components of genetic variance and covariance for linear type traits in Jersey cattle. *Journal of Dairy Science* **68**, 2989-2994.

Thompson, E. A. & Skolnick, M. H. (1977). Likelihoods on complex pedigrees for quantitative traits. In: Pollack, E., Kempthorne, O. & Bailey, T. B. (eds). *Proceedings of the International Conference on Quantitative Genetics.* Iowa State University Press, Ames, Iowa, 815-818.

Uimari, P. & Kennedy, B.W. (1990). Mixed model methodology to estimate additive and dominance genetic values under complete dominance and inbreeding. In: *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production*, **XIII**, 297.

Varona, L. & Misztal, I. (1998). Prediction of parental dominance combinations for planned matings: methodology and simulation results. In: *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia, **25**, 593-596.

Van Raden, P.M. & Jung, Y.C. (1988). A general purpose approximation to REML: the tilde-hat approach. *Journal of Dairy Science* **71**, 187.

Van Raden, P.M., Lawlor, T.J., Short, T.H., & Hoeschele, I. (1992). Use of reproductive technology to estimate variances and predict effects of gene interactions. *Journal of Dairy Science* **75**, 2892-2901.

Wang, C. S., Rutledge, J. J. & Gianola, D. (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics, Selection, Evolution* **25**, 41-62.

Wei, M., & van der Werf, J.H.J. (1993). Animal model estimation of additive and dominance variances in egg production traits of poultry. *Journal of Animal Science* **71**, 57-65.

# VITA

Katharine Gage Lahti was born September 5, 1972, in Kalamazoo, Michigan. She graduated from Mattawan Consolidated High School in June 1990. She then entered the College of Agriculture and Natural Resources and the Honors College at Michigan State University and received a Bachelor of Science degree with high honors in animal science in July 1994. Since that time she has worked on her masters program in the Department of Dairy Science at Virginia Polytechnic Institute and State University. The author is currently a member of the American Dairy Science Association and an associate member of Sigma Xi.