Evaluation of Parametric and Nonparametric Statistical Methods in Genomic Prediction

Reka Howard
Iowa State University
Evaluation of parametric and nonparametric statistical methods in genomic prediction

by

Reka Howard

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

Major: Statistics and Plant Breeding

Program of Study Committee:
Alicia Carriquiry, Co-major Professor
William Beavis, Co-major Professor
Rohan Fernando
Paul Scott
Alyson Wilson

Iowa State University
Ames, Iowa
2016

Copyright © Reka Howard, 2016. All rights reserved.
DEDICATION

This dissertation is gratefully dedicated to my loving husband Seth for supporting me and believing in me, to my wonderful sons Bence and Sethy who are my treasures, and to my sister Andus for inspiring me. I would also like to thank my friends and family for their loving guidance during the time of this work.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>xvi</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xvii</td>
</tr>
<tr>
<td>CHAPTER 1. GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2. PARAMETRIC AND NONPARAMETRIC STATISTICAL METHODS FOR GENOMIC SELECTION OF TRAITS WITH ADDITIVE AND EPISTATIC GENETIC ARCHITECTURES</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Abstract</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Introduction</td>
<td>6</td>
</tr>
<tr>
<td>2.2.1 Parametric Methods in Genome-Wide Selection</td>
<td>10</td>
</tr>
<tr>
<td>2.2.2 Nonparametric Methods in Genome-Wide Selection</td>
<td>20</td>
</tr>
<tr>
<td>2.3 Materials and Methods</td>
<td>29</td>
</tr>
<tr>
<td>2.3.1 Least Squares Regression</td>
<td>32</td>
</tr>
<tr>
<td>2.3.2 Ridge Regression</td>
<td>32</td>
</tr>
<tr>
<td>2.3.3 Bayesian Ridge Regression:</td>
<td>33</td>
</tr>
<tr>
<td>2.3.4 BLUP</td>
<td>33</td>
</tr>
<tr>
<td>2.3.5 LASSO</td>
<td>33</td>
</tr>
<tr>
<td>2.3.6 Bayesian LASSO</td>
<td>34</td>
</tr>
<tr>
<td>2.3.7 Bayesian Alphabet</td>
<td>34</td>
</tr>
<tr>
<td>2.3.8 Nadaraya-Watson Estimator</td>
<td>34</td>
</tr>
<tr>
<td>2.3.9 Reproducing Kernel Hilbert Space</td>
<td>35</td>
</tr>
</tbody>
</table>
CHAPTER 3. RESPONSE SURFACE METHODOLOGY IN GENOMIC SELECTION

3.1 Abstract .............................................................. 60
3.2 Introduction .......................................................... 61
3.3 Materials and Methods .............................................. 62
  3.3.1 Response Surfaces and Approximations ..................... 62
  3.3.2 Experimental Design and Estimation ......................... 66
  3.3.3 Moving on the Estimated Response Surface .................. 68
  3.3.4 Scales and Other Considerations .............................. 70
  3.3.5 Simulated Data ................................................ 71
3.4 Results ............................................................... 73
  3.4.1 Implementing Response Surface Methodology for Evaluating Factors Affecting Genomic Selection ......................... 73
3.5 Discussion ............................................................ 78

CHAPTER 4. PREDICTING PROGENY’S PHENOTYPE USING PARENTAL PHENOTYPIC AND GENOTYPIC INFORMATION .......................... 84

4.1 Abstract .............................................................. 84
4.2 Introduction .......................................................... 85
4.3 Materials and Methods .............................................. 87
4.4 Results and Discussion .............................................. 91

CHAPTER 5. SUMMARY AND DISCUSSION .................................. 112

APPENDIX PARAMETRIC AND NONPARAMETRIC STATISTICAL METHODS FOR SIMULATED BC POPULATION ......................... 114
LIST OF TABLES

Table 2.1 Specification of the simulated $F_2$ population. The table contains information about the genetic architecture and the heritability. . . . . . . . 30

Table 2.2 The parameter specifications for Bayes A, Bayes B, Bayes C, and Bayes $C_{\pi}$ used in GenSel. The table contains the number of iterations used for chain length, and burn-in period, the genotypic variance, the residual variance, the degrees of freedom for residual variance, the degrees of freedom for marker variance, and the probability corresponding to having a 0 effect marker. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

Table 2.3 Bandwidth values used for each of the four combinations of genetic architectures and heritabilities for the Nadaraya-Watson prediction. . . 35

Table 2.4 The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 213 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 493 replicates, and for the rest, the calculations are based on 500 replicates. 41
Table 2.5 The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.30$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 184 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 498 replicates, and for the rest, the calculations are based on 500 replicates.

Table 2.6 The mean and standard error of the mean squared error values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations are based on 500 replicates.

Table 2.7 The mean and standard error of the mean squared error values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.30$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations are based on 500 replicates.

Table 3.1 The table shows the treatment combinations and the factorial effects for a $2^3$ design. The table contains the $+$ and $-$ signs for the factorial effect levels. Taking only the treatment combinations where the $ABC$ factorial effect is $+$ or $-$ will give a half $2^{3-1}$ fractional factorial design.
Table 3.2 Specification of the two levels of the factors include \( n \), number of segregating progeny, \( m \), marker number, QTL number, proportion of genetic variance due to epistasis and narrow sense heritability. Epistasis 0 means that all of the genetic variance is additive variance, 0.5 epistasis means that half of the genetic variance is additive and the other half is epistatic.

Table 3.3 Mean accuracy of BLUP, mean accuracy of SVM, and the response (difference of mean accuracy of SVM and mean accuracy of BLUP) for 16 treatment combinations.

Table 3.4 The levels of factors, the mean of the levels of the factors and half of the difference between the levels of factors.

Table 3.5 Base, increment and the coordinates of the steepest ascent for the number of individuals, number of markers, number of QTL, proportion of epistasis and the degree of heritability for the initial 16 experimental runs.

Table 3.6 Mean accuracy of BLUP, mean accuracy of SVM, and the response (difference of mean accuracy of SVM and mean accuracy of BLUP) for the additional treatment combinations for the additional runs.

Table A.1 Specification of the simulated BC population. The table contains information about the genetic architecture and the heritability.

Table A.2 Bandwidth values used for each of the four combinations of genetic architectures, heritabilities and population types for the Nadaraya-Watson prediction.
Table A.3 The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the BC population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 183 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 494 replicates, and for the rest, the calculations are based on 500 replicates.

Table A.4 The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the BC population with heritability $h^2 = 0.30$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 201 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 494 replicates, and for the rest, the calculations are based on 500 replicates.

Table A.5 The mean and standard error of the mean squared error values for the parametric and the nonparametric methods for the BC population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations are based on 500 replicates.
Table A.6  The mean and standard error of the mean squared error values for the
parametric and the nonparametric methods for the BC population with
heritability $h^2 = 0.30$. The table contains the mean and standard error
of the prediction accuracy values for both the additive and the epistatic
cases. The first 10 methods are parametric, and the last 4 are nonpara-
metric. The calculations are based on 500 replicates. 


LIST OF FIGURES

Figure 2.1 The influence of the bandwidth in kernel density estimation. From left to right the first plot shows simulated data from a mixture of two normal distributions. The second, third and fourth plots show the Gaussian kernel density estimates using bandwidth values $h = 0.1$, $h = 2$, and $h = 10$. ................................................................. 21

Figure 2.2 Loss functions used for SVM regression. The first panel shows the absolute loss function. The second panel is the square loss function, and the last panel is the $\epsilon$-insensitive loss function. ......................... 26

Figure 2.3 A tree-layer feed-forward neural network with $K$ input layer units, $L$ hidden layer units, and $M$ output layer units. ..................... 28

Figure 2.4 The histogram of the simulated phenotypic values. The histograms represent the distribution of the phenotypic values for the $F_2$ population. 31

Figure 2.5 The boxplots of accuracy of prediction for the $F_2$ population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. .................... 44

Figure 2.6 The boxplots of accuracy of prediction for the $F_2$ population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. .................... 44
Figure 2.7 The boxplots of accuracy of prediction for the $F_2$ population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.8 The boxplots of accuracy of prediction for the $F_2$ population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.9 The boxplots of mean squared error for the $F_2$ population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.10 The boxplots of mean squared error for the $F_2$ population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.11 The boxplots of mean squared error for the $F_2$ population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.12 The boxplots of mean squared error for the $F_2$ population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.13 Plots of the parametric to nonparametric accuracy and MSE ratios. The left side of the plots shows the additive cases, and the right side of the plots shows the epistatic cases.

Figure 3.1 Response surface of yield in relation to temperature and drought.
Figure 3.2  Contour plot (level curves) of the response surface of yield. 64

Figure 4.1  The proportion of polymorphic markers in the 45 crosses constructed by 10 genetically diverse SoyNAM parents. 93

Figure 4.2  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications. 94

Figure 4.3  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications. 95

Figure 4.4  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications. 96

Figure 4.5  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications. 97

Figure 4.6  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications. 98

Figure 4.7  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications. 99
Figure 4.8  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications. 100

Figure 4.9  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications. 101

Figure 4.10  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications. 102

Figure 4.11  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications. 103

Figure 4.12  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications. 104

Figure 4.13  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications. 105

Figure 4.14  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications. 106
Figure 4.15 The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications. 107

Figure 4.16 The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications. 108

Figure 4.17 The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications. 109

Figure A.1 The histogram of the simulated phenotypic values. The histograms represent the distribution of the phenotypic values for the BC population. 115

Figure A.2 The boxplots of accuracy of prediction for the BC population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. 119

Figure A.3 The boxplots of accuracy of prediction for the BC population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. 119

Figure A.4 The boxplots of accuracy of prediction for the BC population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. 120
Figure A.5  The boxplots of accuracy of prediction for the BC population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. . . . . . . . . . . . . . 120

Figure A.6  The boxplots of mean squared error for the BC population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. . . . . . . . . . . . . . 121

Figure A.7  The boxplots of mean squared error for the BC population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. . . . . . . . . . . . . . 121

Figure A.8  The boxplots of mean squared error for the BC population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. . . . . . . . . . . . . . 122

Figure A.9  The boxplots of mean squared error for the BC population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods. . . . . . . . . . . . . . 122
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere gratitude to those who helped me with various aspects of conducting this research and my graduate studies. First, my major advisors Dr. Alicia Carriquiry and Dr. William Beavis, for their guidance, encouragement, understanding, patience, and for being great mentors. Without their support and insights I would have never been able to write this dissertation. No words can truly describe my consideration for them so I will just simply say thank you for everything.

I would also like to thank my committee members Dr. Rohan Fernando, Dr. Paul Scott, and Dr. Alyson Wilson. Throughout my years of study, I have been fortunate to interact with many graduate students, and faculty. I thank all of them for their kindness and support. However, special thanks are due to Dr. Minsun Riddles who showed me the beauty of R, and who provided comfort and a lot of fun throughout this dissertation.

I would like to thank my family, especially my sister, mother-in-law, father-in-law, Gail, sisters-in-law, brothers-in-law, and uncle Mike for their constant support and unconditional love. My love and prayers go to them. Throughout my years of study, my husband Seth has been by my side in good times and in difficult times. His love and care have made my work and life much easier and enjoyable. His courage, love, and dedication have always inspired me. To him I say: I love you. Last but not least I would like to thank my little boys for being such incredible children, and making life so wonderful that I could not have imagined before having them.
ABSTRACT

The availability of high-density markers resulted an increased interest in the use of markers for phenotype prediction in plant breeding. Genomic Prediction is a technique that uses marker and phenotypic information of individuals to build a model that enables plant breeders to predict the phenotypic value of individuals with only genotypic scores. In recent years there have been a large number of parametric and nonparametric statistical methods developed for purposes of genomic prediction.

Initially we review parametric methods including Least Squares Regression, Ridge Regression, Bayesian Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Bayesian LASSO, best linear unbiased prediction (BLUP), Bayes A, Bayes B, Bayes C, and Bayes C\(\pi\), and nonparametric methods including Nadaraya-Watson Estimator, Reproducing Kernel Hilbert Space, Support Vector Machine Regression, and Neural Networks. We also contrast the methods based on accuracy and mean squared error (MSE) using simulated genetic architectures consisting of completely additive or two-way epistatic interactions in populations derived from crosses of inbred lines where the genetic architecture contributes low (0.3) and high (0.7) proportions of the total simulated phenotypic variability.

Based on these preliminary results we introduce Response Surface Methodology (RSM) as a systematic strategy for investigating Genomic Prediction methods as an efficient approach to investigating a wide range of the design variables. We illustrate RSM with a simulated example where the response we optimize is the difference between prediction accuracies of a parametric method and a nonparametric method. We examine how the number of individuals, markers, QTL, and different percentage of epistasis and heritability maximize the estimated differences in accuracies. We found the the greatest impact on estimates of accuracy and MSE was due to genetic architecture of the population and the heritability of the trait. When epistasis and
heritability are highest, the advantage of using a nonparametric method versus a parametric prediction method is greatest.

Finally, we simulate data for a structured population consisting of multiple families parental generation’s phenotypic and genotypic information to predict the progeny’s phenotypes. Simulations utilized high density molecular genotypic scores from a sample of soybean varieties adapted to maturity zone 3 to establish the structured breeding population. In the simulation we consider low and high heritability, two different genetic architectures, and the training data contain either all of the parents or only a subset of the parents with the highest phenotypic values. We define a different metric to evaluate genomic prediction techniques, where we compare simulated progeny having the highest phenotypic values with predicted progeny having the highest phenotypic values based on their parental phenotypic and genotypic values. We found that if the genetic architecture is additive then the parametric and nonparametric methods perform similarly according to the new metric. When epistasis is present, the nonparametric method had a higher percentage of identical parents than the parametric method.
CHAPTER 1. GENERAL INTRODUCTION

Plant Breeding is a discipline that contributes to improving people’s lives through genetic improvement of traits that are needed to sustainably provide food security to all people. One way to efficiently accomplish genetic improvement is through accurate prediction of the phenotypic trait values using molecular marker information. Genomic prediction (GP) became an important research topic in statistical genetics beginning about 2001 (Meuwissen et al., 2001). The general strategy is to utilize genetic and phenotypic information in a training set to model the association between the phenotype and the genotype. The model is subsequently used to predict the phenotypic value of individuals for which only the genotypic marker information is available.

For purposes of predicting phenotypes in Plant Breeding, both parametric and nonparametric statistical models have been developed. The question is whether a parametric or a nonparametric method should be preferred for phenotype prediction, and how these statistical methods are influenced by the phenotypic variation explained by the genotypic variation, the number of individuals, number of markers and quantitative trait loci (QTL), and the genetic architecture. With emergence of ‘omic’ technologies there has been considerable evidence gene-by-gene interaction (epistasis) represents a major contributor to phenotypic expression (Flint and Mackay, 2009). Examining how epistasis affects the GP techniques is crucial. In our first study, we evaluated and compared 10 parametric and 4 nonparametric statistical methods developed for GP. Our goal was to examine how genetic architecture and heritability influence prediction accuracy through use of correlations between predicted and actual values and use of the mean squared error (MSE). Comparisons were based on simulated data sets of phenotypes and genotypes for plant populations derived from a cross of two inbred lines. For half of the data sets genetic architecture of the phenotypes were simulated to have independent additive effects, and for the
other half the genetic architecture was due to dependence among alleles at independent pairs of loci, i.e., epistasis. Within the data sets, the genetic architecture explained either 30% or 70% of the phenotypic variability, and the rest was due to non-genetic sources. We found that the greatest impact on estimates of accuracy and mean squared error was due to genetic architecture. Parametric methods were unable to predict phenotypic values when the underlying genetic architecture was based entirely upon epistatic causative alleles. Parametric methods performed slightly better than nonparametric methods for additive genetic architectures. Distinctions among parametric methods for additive genetic architectures were incremental. The proportion of phenotypic variability (i.e. heritability) had the second greatest impact on estimates of accuracy and mean squared error.

Our first study was important to validate the notion that genetic architecture can have an enormous impact on prediction accuracy. It also helped practicing plant breeders understand the distinctions between parametric and nonparametric statistical methods. However, it is important to recognize that the methods were evaluated under a limited set of factors that could influence differences among GP methods. Our first project suggested a need to evaluate a much larger combination of factor levels to better understand the behavior of parametric and nonparametric methods in terms of prediction accuracy and precision.

In our second research project we examined how changing the combinations of factors might affect the accuracy of prediction. Also, we wanted to find the combinations of factors that would maximize differences of prediction accuracies among prediction methods. To address these goals, we utilized Response Surface Methodology (RSM) and approximated the relationship between a set of factors that influence GP and the Pearson correlation coefficient between the true phenotype and the predicted phenotype. We found that the difference of estimated accuracies between nonparametric and parametric methods is maximized when epistasis and heritability were maximized, while the number of individuals, number of markers and the num-
ber of influential markers had little influence on the performance differential. Importantly, we demonstrated that development and assessment of statistical genetic methods can be pursued in a rigorous rather than an ad hoc manner.

For the first two research projects comparisons were based on simulated progeny from a single cross of completely unrelated inbred lines. In our third study, we simulated phenotypic and genotypic information for a more realistic breeding population structure in soybeans. We used real soybean genotypic data to simulate data that mimic the start of a breeding population with existing cultivars. The use of real genotypic data also enabled us to simulate recombination using experimental estimates of recombination. We developed a metric that is different from the routinely used accuracy or MSE to evaluate GP methods. Because plant breeders are interested in progeny from the tails of distributions, we were interested in predicting which parents are most likely to produce the most extreme desirable phenotypes, and how genetic architecture and heritability influence the prediction. The results agreed with our previous findings that when epistasis is present the nonparametric prediction method outperforms the parametric method. For additive cases there is no significant difference between the parametric and the nonparametric method in terms of predicting the parents that should be crossed to produce superior cultivars.

The details of the research projects can be found in the subsequent chapters. In Chapter 2 we compare the parametric and nonparametric statistical methods used in GP using simulated \( F_2 \) data. The results were very similar when we compared the methods using simulated \( BC \) data, so we put all of the results for the \( BC \) data into the Appendix. Chapter 3 is devoted to presenting the use of RSM in GP using simulated \( BC \) population, and finding the optimum conditions when we maximize the difference between the parametric Best Linear Unbiased Prediction and the nonparametric Support Vector Machine in terms of prediction accuracy. Chapter 4 focuses on simulating a family structure instead of progeny from a single cross, and a metric is evaluated for GP that focuses on predicting which parents we need to cross to produce superior progeny instead of correlation or MSE. Each of the chapters 2 through 4 is structured as a journal article. Thus, each chapter contains its own list of literary citations. Chapter 5 summarizes the general conclusions of this dissertation.
Bibliography


CHAPTER 2. PARAMETRIC AND NONPARAMETRIC STATISTICAL METHODS FOR GENOMIC SELECTION OF TRAITS WITH ADDITIVE AND EPISTATIC GENETIC ARCHITECTURES

2.1 Abstract

Parametric and nonparametric methods have been developed for purposes of predicting phenotypes. These methods are based on retrospective analyses of empirical data consisting of genotypic and phenotypic scores. Recent reports have indicated that parametric methods are unable to predict phenotypes of traits with known epistatic genetic architectures. Herein we review parametric methods including Least Squares Regression, Ridge Regression, Bayesian Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO), Bayesian LASSO, best linear unbiased prediction (BLUP), Bayes A, Bayes B, Bayes C, and Bayes Cπ. We also review nonparametric methods including Nadaraya-Watson Estimator, Reproducing Kernel Hilbert Space, Support Vector Machine Regression, and Neural Networks. We assess the relative merits of these 14 methods in terms of accuracy and mean squared error (MSE) using simulated genetic architectures consisting of completely additive or two-way epistatic interactions in an $F_2$ population derived from crosses of inbred lines. Each simulated genetic architecture explained either 30% or 70% of the phenotypic variability. The greatest impact on estimates of accuracy and MSE was due to genetic architecture. Parametric methods were unable to predict phenotypic values when the underlying genetic architecture was based entirely upon epistasis. Parametric methods were slightly better than nonparametric methods for additive genetic architectures. Distinctions among parametric methods for additive genetic architectures were incremental. Heritability, i.e., proportion of phenotypic variability, had the second greatest impact on estimates of accuracy and MSE.
2.2 Introduction

Complex quantitative traits are measured on a continuous scale and are controlled by a network of many genes, by the environment and by genetic by environment interactions. Most traits of economical interest in Agriculture (e.g., grain yield) are measured on continuous scales, i.e., they are quantitative. Understanding the complexity of these traits and accounting for the effects that are contributed by these genes and their interactions is non-trivial.

The gene by gene interaction or epistasis is an important research topic in quantitative genetics. Epistasis can be modeled in different ways (Cordell 2002). Physiological epistasis is the difference in the phenotype when the genotype at a locus is influenced by the genotype at another locus or loci. Fisher (1918) defined epistasis as the deviation of the genotypic value from the contribution of the sum of additive effects at all functional loci in the genome. Fisher’s definition of epistasis is also known as statistical epistasis and has been used to quantify deviations from independence (Wilson 2004). Epistasis has an important role in accounting for the genetic variation for quantitative traits, and excluding it from the prediction equations for simplicity can result in poor predictions of genetic gain (Cooper et al. 2002).

Most simulation studies of genomic selection (GS) methods (Meuwissen et al. 2001) have considered genetic architectures in which the number and relative magnitudes of QTL have varied. To our knowledge no studies of GS methods have considered epistatic genetic architectures, although Gianolola (2006) predicted non-parametric methods would be better suited for epistatic genetic architectures. While theoretic models predict a significant role for epistasis in speciation (Dobzhansky 1937, Mayr 1942), adaptation (Lewontin 1974, Wade 2000) and canalization (Waddington 1949, Rice 1998), there is little empirical evidence from biometric studies of significant epistatic contributions to genetic variability. Biometric approaches, however average across epistatic genotypic values at individual loci and contribute primarily to additive genetic variance (Cockerham 1954, Cheverud and Routman 1995). With development of low cost high throughput marker technologies, it has become possible to estimate epistatic interactions based on genotypic values for all possible pairwise genotypes in Genome Wide Association Studies, although searches for higher order interactions are still limited by experimental and compu-
tional resources (Moore and Williams 2009). These studies are beginning to reveal that epistasis is not the exception, rather the most prevalent form of genetic architecture for quantitative traits (Flint and MacKay 2009, Huang et al. 2012). Nonetheless, it was hypothesized that GS should provide accurate predictions because epistatic gene action will be translated primarily into additive genetic variance (Crow 2010). Thus, for purposes of this study, we decided to evaluate GS methods for an extreme case of epistasis with ten pairs of loci each consisting of two alleles at equal frequencies and modeled using the principle of orthogonality (Goodnight 2000).

The development of DNA markers in the 1980’s was an important step in the process of identifying DNA segments that are statistically associated with quantitative traits, i.e., quantitative trait loci (QTL) mapping and for marker assisted selection. In marker assisted selection (MAS), markers and phenotypic information are used to guide indirect selection of a trait of interest. This approach is considered an improved and more efficient method for selection in plant breeding relative to phenotype-pedigree based approaches (Mohan et al. 1997). Extensive resources have been devoted to develop QTL mapping methodology as a component of MAS (Young 1996; Melchinger et al. 1998). Marker assisted backcrossing (MABC) is one of the simplest examples of MAS. In MABC, genomic regions defined by markers closely linked to QTL are identified. These genomic regions are then introgressed into the elite lines through backcrossing (Bernardo 2010). In MABC a plant with a desired gene - called a donor parent - is crossed with an elite or breeding line, called a recurrent parent. The goal is to introgress the desired gene into the genome of the recurrent parent (Visscher et al. 1996). Developing varieties can also involve accumulating multiple desired genes into a recurrent parent. The marker assisted process for alleles at multiple loci is called gene pyramiding. MAS is widely used in gene pyramiding because the use of molecular markers gives the advantage of selecting the desired plants without extensive phenotyping. With traditional phenotyping it is often impossible to distinguish among plants with all desirable alleles and the plants with some of the desirable alleles (Huang et al. 1997).

MAS has been shown to be efficient and effective for traits that are associated with one or a few major genes with large effect, but does not perform as well when it is used for selection of
polygenic traits (Bernardo, 2008). QTL detection also results in some false negative and false positive rates, and further QTL mapping does not guarantee that estimates of genetic effects are correct (Beavis 1994). Also, for MAS to be useful, the interaction between the QTL and the genetic background has to be minimal, so the QTL has the same effect in different genetic backgrounds (Bernardo 2010 p. 223). The genetic background of an organism refers to all of its alleles at all loci that can interact with the locus where the QTL is located (Yoshiki et al. 2006).

The parametric models and statistical methods introduced for QTL mapping and MAS do not address genetic improvement for quantitative traits that are influenced by a large number of genes with small effects. Some of the statistical challenges arising in MAS include the specification of threshold for multiple testing, the “large p, small n” problem (which refers to the situation when the number of predictors, p (marker data points) greatly exceeds the number of individuals, n, that have been evaluated in the study), difficulty of interpretation of effects due to collinearity among the explanatory/predictor variables, model assumptions that cannot be satisfied, and non-additivity among genetic effects.

With advanced molecular techniques that provide dense marker maps it is possible to overcome some shortcomings of MAS. Meuwissen et al. (2001) proposed predicting the genotypic value for individuals using all marker information simultaneously. Their proposed method and the subsequent derivative methods have been referred to as genomic selection (GS). They modeled the associations between the markers and a phenotype focusing on calculating a breeding value for an individual (which can be calculated as the sum of the average effect of the alleles for the individual’s genotype) instead of identifying significant marker-trait associations. In their approach they estimated the effect of each QTL, and then used the sum of all estimates to calculate a genotypic value for the individual.

In GS individuals with both phenotypic and marker information (called the training set) are used to model the association between the phenotype and the genotype. The model is used to predict the phenotypic value of individuals for which only the marker information is available (called the validation set or testing set). In GS all available markers are included in the model, not just those above a significant threshold, thus eliminating the problem of multiple testing.
There also is effort underway to find ways to model epistasis, the gene by gene interaction. In the presence of epistasis the effect of one locus changes the effect of another locus on the phenotype. Usually several loci are involved which means that multiway interactions may need to be modeled. Because the volume of marker data points available is huge, the number of epistatic interactions can be overwhelming, and computationally intractable to estimate with parametric methods (Moore et al. 2009).

More recently, Gianola et al. (2006) proposed non-parametric methods capable of accounting for complex epistatic models without explicitly modeling them.

Herein, we review some existing statistical methods used in GS. First, we discuss the parametric methods in more detail, and then focus on the nonparametric and semi-parametric methods. Among parametric methods we review linear least squares regression, penalized Ridge regression, Bayes Ridge regression, LASSO and Bayes LASSO methods, Best Linear Unbiased Prediction (BLUP) and some Bayesian alternatives used in GS (Bayes A, Bayes B, Bayes C and Bayes Cπ). We also explain the nonparametric kernel regression using the Nadaraya-Watson estimator (NWE) and the semi-parametric Reproducing Kernel Hilbert Space (RKHS) regression. Finally, we describe Support Vector Machine (SVM) regression and Neural Networks (NN) applications to GS. De los Campos et al. (2013) give an overview of some of the parametric methods used in GS, and Gianola et al. (2010) provide information about some of the nonparametric models used in GS. Heslot et al. (2012) compared some parametric and nonparametric GS methods. However, they did not consider epistatic genetic architectures in their simulated data. Daetwyler et al. (2010) discussed the impact of the genetic architecture in GS, but they defined genetic architecture by the effective population size and the number of QTL.

We use simulated data to compare the performance of the parametric models with the nonparametric procedures for predicting the genetic value for individuals in a $F_2$ and a backcross (BC) populations. We simulate the $F_2$ and the BC populations with low and high heritabilities, and compared the two extreme genetic architectures. One architecture has only additive genetic effects from alleles at 30 loci, and the other has only 2-way epistatic genetic effects among 30 loci. The performance of the methods is illustrated by comparing the accuracy of prediction
which we define by the correlation between the true phenotypic value and the predicted phenotypic value and the MSE. We demonstrate the advantage of some nonparametric methods for the epistatic genetic architecture. Since the results for the $F_2$ and a backcross (BC) populations were similar, we only illustrate the $F_2$ population in the manuscript. In the supporting information we provide accuracy and MSE values for a simulated BC population with low and highheritabilities, and with two extreme genetic architectures.

2.2.1 Parametric Methods in Genome-Wide Selection

2.2.1.1 Linear Least-Squares Regression Model

In GS the main goal is to predict the individual’s breeding value by modeling the relationship between the individual’s genotype and phenotype. One of the simplest models is

$$y_i = \mu + \sum_{j=1}^{p} X_{ij} m_j + e_i,$$  \hspace{1cm} (2.1)

where $i = 1,...,n$ individual, $j = 1,...,p$ marker position/segment and $y_i$ is the phenotypic value for individual $i$, $\mu$ is the overall mean, $X_{ij}$ is an element of the incidence matrix corresponding to marker $j$ and individual $i$, $m_j$ is a random effect associated with marker $j$, and $e_i$ is a random residual. Typically, the residual term, $e$ is chosen to have a normal distribution with mean of 0 and variance of $\sigma_e^2$. The model for the data vector $y$ can be written as

$$y_{n \times 1} = \mu_{n \times 1} + X_{n \times p} m_{p \times 1} + e_{n \times 1}.$$  \hspace{1cm} (2.2)

To estimate $(\mu, m)$ we can use least squares to minimize the sum of squared vertical distance between the observed response and the estimated response, which can be represented as $|y - Xm|^2$ (where $|$ denotes the norm of a vector). The estimate of $m$ is obtained by solving the linear equations $X'Xm = X'y$. Then it is estimated as $\hat{m} = (X'X)^{-1}X'y$. For more details about linear models the reader can refer to Linear Models in Statistics by G. Bruce Schaalje and Alvin C. Rencher (2000) or Linear Models with R by Julian J. Faraway (2006). The elements of the design matrix $X$ depend on the number of different alleles present. For example, individuals having marker genotypes $AA$, $Aa$, $aa$ have elements coded as $-1$, 0, and 1 in $X_{ij}$ respectively.
One obvious problem with linear regression is that usually the number of markers (explanatory variables) available is much greater than the number of individuals with phenotypic information (response variables), which means that \( p \) is much greater than \( n \), and it is impossible to do the estimation. Using a subset of the markers can be an alternative (using a variable selection method like the forward, backward or stepwise selection procedure) (George 2000), but it can still perform poorly if the relative ratio of the number of markers and the number of individuals is large or multi-collinearity, e.g., linkage disequilibrium (LD) exists among the markers.

Meuwissen et al. (2001) used a modification of LS regression for GS. First, they performed a least square regression analysis on each segment separately using the model \( \mathbf{y} = \mathbf{\mu} + \mathbf{X}_j \mathbf{m}_j + \mathbf{e} \) where \( \mathbf{y} \) is the vector of the phenotypic information, \( \mathbf{\mu} \) is the overall mean vector, \( \mathbf{X}_j \) is the \( j^{th} \) column of the design matrix corresponding to the \( j^{th} \) segment, \( \mathbf{m}_j \) is the genetic effect associated with the \( j^{th} \) segment, and \( \mathbf{e} \) is the vector of the error terms. By plotting the log likelihood of this model, segments with significant effects were found. The segments with significant effect (QTLs) were used for simultaneous estimation by the model \( \mathbf{y} = \mathbf{\mu} + \sum_{j=1}^{q} \mathbf{X}_j \mathbf{m}_j + \mathbf{e} \) where \( q \) is the number of QTLs. By this approach they eliminated the problem of having more predictor (explanatory/independent) variables than regressands (response/dependent variables), but it does not fully take advantage of the whole marker information since only markers with a significant effect are included into the final model. To overcome some of the drawbacks of the linear regression approach, other methods for GS have been introduced.

### 2.2.1.2 Ridge Regression

In marker data it is very likely that multi-collinearity exists. As discussed in the previous section, multicollinearity can negatively affect the performance of variable selection methods. Further, LS equations are inefficient when the determinant of the matrix \( \mathbf{X}'\mathbf{X} \) is close to zero due to column dependencies. Using a penalized regression model (Ridge regression of Hoerl and Kennard 1970a, 1970b) can be a solution to this problem. The goal is to derive an estimator of \( \mathbf{m} \) with smaller variance than the least squares estimator. There is a price to pay in that the ridge regression estimator of \( \mathbf{m} \) is biased; the increase in bias is more than compensated by
the decrease in variance, which result in an estimator $\hat{m}_R$ with smallest mean squared error (MSE). Another advantage of ridge regression is that it can be used when a large amount of marker information is available, so it can overcome the “$p > n$” problem.

Ridge regression adds an extra term to the likelihood function to shrink the regression coefficients by an amount depending on the variance of the covariates. It removes the problem of the columns of the design matrix being dependent on each other, and hence the $X'X$ matrix will be non-singular. Instead of minimizing the sum of squared residuals, Ridge regression minimizes the penalized sum of squares $|y - Xm|^2 + \lambda^2 m'm$ where $\lambda$ is the penalty parameter, and the estimate of the regression coefficient is given by $\hat{m} = (X'X + \lambda I)^{-1}X'y$ where $I$ is a $p \times p$ identity matrix. The penalty parameter $\lambda$ can be calculated by several different methods, for example by plotting $\hat{m}$ as a function of $\lambda$, and choosing the smallest $\lambda$ that results in a stable estimate of $\hat{m}$. Another way to choose $\lambda$ is by an automated procedure proposed by Hoerl et al. (1975). They claimed that a reasonable choice of $\lambda$ is given by $\lambda = \frac{r s^2}{\hat{m}'\hat{m}}$ where $r$ is the number of parameters in the model not counting the intercept, $s^2$ is the residual mean square obtained by linear least squares estimation, and $\hat{m}$ is the vector of least squares estimates of regression coefficients.

Meuwissen et al. (2001) implemented ridge regression in GS by assuming that the marker effects ($m_j$’s $j = 1...p$) were random, and they were drawn from a normal distribution with $Var(m_j) = \sigma_m^2$ where $\sigma_m^2 = \sigma_a^2/n_k \sigma_a^2$ represents additive genetic variance expressed among individuals and $n_k$ is the number of marker loci (Habier et al. 2007).

It can be shown that ridge regression is a special case of the Best Linear Unbiased Prediction (BLUP) (Ruppert et al. 2003), which we will demonstrate after introducing BLUP. Thus the mixed linear model can be implemented. Within the mixed model context the restricted maximum likelihood (REML) estimation is a good choice for finding a reasonable value for the penalty parameter and estimating the variance components (Henderson 1988). Piepho (2009) discusses some models that feature the Ridge regression in terms of mixed models, and uses REML for variance and penalty parameter estimation in GS.

Ridge regression can also be viewed from a Bayesian perspective. In this case we assume that the parameter vector $m$ is random. We can account for the belief that the estimator of $m$
has a small variance by a choice of a prior distribution. In particular, we can suppose that \( \mathbf{m} \sim \mathcal{N}(\mathbf{0}, \Sigma_\beta) \) where \( \Sigma_\beta \) is a known covariance matrix (de Boer et al. 2005). Given that the likelihood of \( y_i \) (\( i = 1, 2, \ldots n \), where \( n \) is the number of individuals) has a normal distribution with mean \( \sum_{j=1}^{p} x_{ij} m_j \) and variance \( \sigma^2 \), the Bayesian estimator of \( \mathbf{m} \) is the mean of the posterior distribution, and it is given by \( \hat{\mathbf{m}}_{BRR} = (\sigma^2 \Sigma^{-1} + X'X)^{-1} X'y \) (Judge et al. 1985 p. 286). Comparing \( \hat{\mathbf{m}}_{BRR} \) to \( \hat{\mathbf{m}}_{RR} \) we can see that they are identical if \( \Sigma^{-1} = \frac{1}{\sigma^2} I \).

Pérez et al. (2010) discussed the application of Bayesian Ridge regression in genomic selection. They assumed that the marker effects are independent and identically distributed (iid), and have a normal prior distribution with mean 0 and variance \( \sigma^2 \), where \( p(m|\sigma^2) = \prod_{i=1}^{n} \mathcal{N}(m_j|0, \sigma^2) \). Then the mean of the posterior distribution \( \hat{\mathbf{m}}_{BRR} \) is equivalent to \( \hat{\mathbf{m}}_{RR} \) if \( \lambda = \frac{\sigma^2_{\beta}}{\sigma^2} \).

### 2.2.1.3 Best Linear Unbiased Prediction

The Best Linear Unbiased Prediction (BLUP) theory and the mixed model formulation was first discussed by Henderson (1949), and it was influential for selection purposes in animal breeding (Henderson 1959). BLUP is a statistical procedure, and it is useful in situations when the data available are unbalanced (for example in different locations the number of individuals is not the same), and it can accommodate family information (Bernardo 2010). Since Henderson’s first work in BLUP, the theory has been widely expanded (Henderson 1959, Henderson 1963, Henderson 1975, Harville 1975). Since the 1990’s BLUP has not only been used in animal breeding applications (Henderson 1984) but also in plant breeding (Bernardo 1994).

BLUP was proposed as a tool in GS by Meuwissen et al. (2001). The random effects model can be written in the form

\[
\mathbf{y} = \mu + \sum_{j=1}^{p} \mathbf{Z}_j \mathbf{m}_j + \mathbf{e},
\]

where \( \mathbf{y} \) is the \((n \times 1)\) phenotypic data vector, \( \mu \) is the \((n \times 1)\) overall mean vector, \( \mathbf{Z}_j \) is the \( j^{th} \) column of the design matrix, \( m_j \) is the genetic effect associated with the \( j^{th} \) marker, and \( p \) is the number of markers. The intercept, \( \mu \) is fixed, and \( m_j \)'s are the random effects with \( E(m_j) = 0 \), \( Var(m_j) = \sigma^2_{m_j} \), \( Var(\mathbf{e}) = \sigma^2 \mathbf{I} \), and \( Cov(\mathbf{m}, \mathbf{e}) = \mathbf{0} \). In the statistical literature
the vector of random effects is usually denoted by \( \mathbf{u} \) instead of \( \mathbf{m} \). If other covariates are available, we replace the intercept \( \mu \) by \( \mathbf{X} \beta \) to include all the fixed effects. Then, we can write

\[
\mathbf{y} = \mathbf{X} \beta + \mathbf{Z} \mathbf{m} + \mathbf{e},
\]

where \( \beta \) is a \( p_1 \times 1 \) vector of unknown fixed effects where usually the first element is the population mean, and \( \mathbf{X} \) is the incidence matrix which relates \( \mathbf{y} \) to \( \beta \). The above equation is generally called a mixed model (or mixed effects model). The vector \( \beta \) is estimated by the best linear unbiased estimator (BLUE). In the biological literature the term BLUP is occasionally used loosely, and refers to both BLUE and BLUP. BLUP is the predictor of the random effects. It is a linear function of the data vector \( \mathbf{y} \). Within the linear functions of the data, it is unbiased which means that the expected value of the prediction is the same as the population parameter, and it can be formulated as \( E(\hat{\mathbf{m}}) = E(\mathbf{m}) \). In addition within the unbiased linear predictors, it is the best in the sense of minimizing the MSE. BLUE is similar to BLUP in that it is a linear function of the data \( \mathbf{y} \), it is unbiased among the linear estimators and it is best in the sense that it minimizes the MSE.

Henderson (1953) proposed that the BLUE and BLUP of \((\beta, \mathbf{m})\) be obtained by maximizing the joint likelihood of \((\mathbf{y}, \mathbf{m})\) given by

\[
L(\mathbf{y}, \mathbf{m}) = f(\mathbf{y}|\mathbf{m})f(\mathbf{m})
\]

\[
= \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}} \left[ -\frac{1}{2} (\mathbf{y} - \mathbf{X} \beta - \mathbf{Z} \mathbf{m})' \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X} \beta - \mathbf{Z} \mathbf{m}) \right] \times \frac{1}{(2\pi)^{p/2}|\mathbf{G}|^{1/2}} \left[ -\frac{1}{2} \mathbf{m}' \mathbf{G}^{-1} \mathbf{m} \right].
\]

By maximizing the likelihood \( L(\mathbf{y}, \mathbf{m}) \) with respect to \( \beta, \mathbf{m} \) and equating it to zero, we obtain a set of linear equations (known as Henderson’s mixed model equations (MME)):

\[
\begin{pmatrix}
\mathbf{X}' \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}' \mathbf{R}^{-1} \mathbf{Z} \\
\mathbf{Z}' \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}' \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1}
\end{pmatrix}
\begin{pmatrix}
\hat{\beta} \\
\hat{\mathbf{m}}
\end{pmatrix}
=
\begin{pmatrix}
\mathbf{X}' \mathbf{R}^{-1} \mathbf{y} \\
\mathbf{Z}' \mathbf{R}^{-1} \mathbf{y}
\end{pmatrix},
\]

where \( R = \text{Var}(\mathbf{e}) \) and \( G = \text{Var}(\mathbf{m}) \). The solution to the MME is the BLUE of \( \beta \) and the BLUP of \( \mathbf{m} \). Henderson’s derivation assumes that \( \mathbf{m} \) and \( \mathbf{e} \) are normally distributed, and maximizes the joint likelihood of \((\mathbf{y}, \mathbf{m})\) over the unknowns \( \beta \) and \( \mathbf{m} \). Maximizing the likelihood
implies an optimization criterion of \((y - X\beta - Zm)'R^{-1}(y - X\beta - Zm) + m'G^{-1}m\), and it can be viewed as the “Ridge regression formulation" of the BLUP (Ruppert et al. 2003).

We have assumed that \(R\) and \(G\) are known covariance matrices. In general they are unknown and need to be estimated together with \(\beta, m\). The Restricted Maximum Likelihood (REML) approach to estimate the variance components maximizes the “restricted” likelihood associated with a specific set of linear combinations of the data. The restricted likelihood depends only on the variance components. REML produces unbiased estimates of the variance parameters \(R\) and \(G\). More information about variance estimation using REML can be found in Corbeil and Searle (1976), Harville (1977), and McGilchrist (1993). There are other ways to derive the BLUP solution for \(m\). Robinson (1991) showed that the BLUE solution to \(\beta\) can be written as \(\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y\), and the the BLUP solution to \(m\) can be written as \(\hat{m} = GZV^{-1}(y - X\hat{\beta})\).

### 2.2.1.4 LASSO Method

To overcome the limitations of linear LS we can use the Least Absolute Shrinkage and Selection Operator (LASSO) for GS. LASSO was first introduced by Tibshirani (1996), and Usai et al. (2009) first implemented it in GS using cross-validation. We can write the model for individual \(i\) as

\[
y_i = \sum_{j=1}^{p} X_{ij}m_j + e_i,
\]

where \(i = 1...n\) individual, \(j = 1...p\) marker position and \(y_i\) is the phenotypic value for individual \(i\), \(X_{ij}\) is an element of the incidence matrix corresponding to individual \(i\) and marker \(j\), \(m_j\) is the marker effect for marker \(j\) and \(e_i\) is the random residual. The LASSO estimate of the marker effect is obtained by minimizing the residual sum of squares \(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}m_j)^2\) subject to the constraint of the sum of the absolute value of the marker effects being less than a constant \(s\), \(s \geq 0\), and we can write it as \(\sum_{j=1}^{p} |m_j| \leq s\). This constraint shrinks some of the marker effects, and sets some of them to zero. One of the major differences between LASSO and ridge regression is that in LASSO as we increase the penalty, more marker effects will shrink to zero and in ridge regression all parameters will be reduced but still remain non-zero.

The LASSO estimator of the regression coefficients \(m_j's\) can be found by an algorithm that was
first described by Tibshirani (1996), and used computational ideas from Lawson and Hansen (1974). First, we assume that the elements of the incidence matrix are standardized such that \( \sum_{i=1}^{n} X_{ij} = 0 \) and \( \sum_{i=1}^{n} X_{ij}^2 = n \). Then, the algorithm describes a quadratic programming problem with \( 2^p \) linear constrains, corresponding to the different signs for the regression coefficients \( m_j \). For example if \( p = 3 \) then we have

\[
\begin{align*}
    m_1 + m_2 + m_3 &\leq s \\
    m_1 + m_2 - m_3 &\leq s \\
    m_1 - m_2 + m_3 &\leq s \\
    m_1 - m_2 - m_3 &\leq s \\
    -m_1 + m_2 + m_3 &\leq s \\
    -m_1 + m_2 - m_3 &\leq s \\
    -m_1 - m_2 + m_3 &\leq s \\
    -m_1 - m_2 - m_3 &\leq s.
\end{align*}
\]

Let \( f(m) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} m_j X_{ij})^2 \) and for \( k = 1 \ldots 2^p \) let \( \gamma_k \) be a vector of indicator variables \( 1, 0, -1 \) depending on the signs of the regression coefficients corresponding to the \( k^{th} \) inequality. Also, let \( E = \{i : \gamma_i' m = s\} \) and \( S = \{i : \gamma_i' m < s\} \). \( G_E = [\gamma_1, \gamma_2, \ldots \gamma_{2^p}]' \).

The steps of the algorithm finding the LASSO estimator can be written as

1. Let \( E = \{i_0\} \) where \( i_0 \) corresponds to the LS estimate of \( m \) and \( \gamma_{i_0} = \text{sign}(\hat{m}_{LS}) \).

2. Find \( \hat{m} \) such that \( f(m) \) is minimized subject to \( G_E m \leq s1 \).

3. If \( \sum_{j=1}^{p} |m_j| \leq s, \) done.
   
   If \( \sum_{j=1}^{p} |m_j| > s \), \( E = \{i_0, i\} \) such that \( \gamma_i = \text{sign}(\hat{m}) \). Repeat Step 2 and 3.

The algorithm described above is computationally intensive. Efron et al. (2004) proposed a new model selection algorithm called Least Angle Regression (LARS) that can be used in combination with LASSO estimation. LARS is similar to the traditional forward selection method. It starts with all the coefficients (marker effects) at zero. First, the marker which has the highest correlation with the phenotypic values is added into the model. The next marker
added has to have a correlation with the residual which is at least as large. The third marker entered into the model is equiangular with the first two markers already in the model. At each iteration a new marker is added, and the algorithm is accomplished in $p$ iterations where $p$ is the number of the available markers. However, for LASSO the LARS procedure is modified. Since the LASSO has a constraint, the LARS procedure has to apply a restriction, so this model selection method is more closely related to the stepwise selection method. For a detailed description of LARS and the LARS-LASSO relationship, the reader can refer to Efron et al. (2004).

One other important question is how to find the upper bound of the sum of the absolute value of the marker effects, $s$. Finding the best value for $s$ can be viewed as the selection of the size of the best subset of markers (Usai et al. 2009). Usai et al. used Kohavi’s (1995) cross-validation approach with random sub-sampling replication. In every replication, the data are randomly divided into a training set and a validation set. The training set is used to estimate the marker effects using the LARS algorithm for the LASSO method. The estimated marker effects were used to calculate the genomic breeding values (GEBV) for the individuals in the validation set, and then the correlation coefficients between the GEBV and the true phenotypic value were reported. The LARS iterations were carried forward until the maximum correlation was reached.

### 2.2.1.5 The Bayesian Alphabet

Meuwissen et al. (2001) proposed two hierarchical Bayesian models for GS denoted by Bayes A and Bayes B. In both methods the data and the variances of the marker positions need to be modeled. For individual $i$ we can write

$$y_i = \mu + \sum_{j=1}^{p} X_{ij} m_j + e_i,$$

where $i = 1...n$ individual, $j = 1...p$ marker position/segment and $y_i$ is the phenotypic value for individual $i$, $\mu$ is the $n \times 1$ dimensional overall mean vector, $X_{ij}$ is an element of an incidence matrix for marker $j$ and individual $i$, $m_j$ is a random effect for marker $j$, and $e_i$ is a random residual. (In general the model can be written as $y = \mu + \sum_{j=1}^{p} X_j m_j + e$.)
Inferences about model parameters are based on the posterior distribution. By Bayes’ Theorem, the posterior is obtained by combining the prior distribution and the likelihood function. For detailed information about Bayesian methods the reader can refer to Kruschke (2010) or Gelman et al. (2003).

The difference between Bayes A and Bayes B lies in the way in which we model the variances of parameters. In both methods each marker position has its own variance. The Bayes A approach applies the same prior distribution for all of the variances of the marker positions. The scaled inverted (or inverse) chi-squared probability distribution $\chi^{-2}(\nu, S^2)$ can be used with degrees of freedom $\nu$ and scale parameter $S^2$ as the prior distribution. This is a convenient choice because it is a conjugate prior so the posterior distribution is in the same family of distributions as the prior distribution. The posterior distribution is also a scaled inverse chi square distribution $\chi^{-2}(\nu + n_j, S^2 + m^T_j m_j)$ where $n_j$ is the number of haplotype effects at marker position $j$.

The Bayes B approach seems more realistic for GS than Bayes A. Bayes B assumes that not all markers contribute to the genetic variation. It has a prior density on the variance that is a mixture. It has a high probability mass at $\sigma_{m_j} = 0$, and an inverted chi-square distribution when $\sigma_{m_j} > 0$. It can be summarized as $\sigma_{m_j} = 0$ with prob = $\pi$ and $\sigma_{m_j} \sim \chi^{-2}(\nu, S)$ with prob = $(1 - \pi)$.

For the Bayes B method if $m^T m > 0$, one cannot sample $\sigma^2_{jj} = 0$. So we can sample $m_j$ and $\sigma^2_{jj}$ simultaneously by $p(\sigma^2_{m_j}, m_j|y^*) = p(\sigma^2_{m_j}|y^*)p(m_j|\sigma^2_{m_j}, y^*)$ where $y^*$ is the data that is corrected for the mean and for all genetic effects except $m_j$.

To sample from the distribution $p(\sigma^2_{m_j}|y^*)$, we can use the Metropolis-Hastings algorithm in the following way:

1. Sample $\sigma^2_{m(new)}$ from the prior distribution of $p(\sigma^2_{m_j})$
2. $\sigma^2_{m_j} = \sigma^2_{m(new)}$ with probability of $\text{Min} \left[ \frac{p(y^*|\sigma^2_{m(new)})}{p(y^*|\sigma^2_{m_j})} ; 1 \right]$.

Using simulated data it was shown that the Bayesian methods perform better in terms of prediction accuracy than the linear least squares regression, the Ridge regression and the BLUP method (Meuwissen et al. 2001, Habier et al. 2009, Habier et al. 2010). However, as Gianola et al. (2009) pointed out the choice of the degrees of freedom and the scale parameters of the
scaled inverse chi square distribution can influence the outcome. Improved Bayesian methods were developed by Habier et al. (2011) to deal with the weakness of Bayes A and Bayes B. Bayes C uses a common variance for all SNPs, and for Bayes D the scale parameter of the scaled inverse chi square distribution is estimated instead of specified by the user. Bayes Cπ and Bayes Dπ (Habier et al. 2011) are the modification of Bayes C and Bayes D where the probability of having a zero effect SNP, π is estimated.

### 2.2.1.6 Bayesian LASSO

Park and Casella (2008) introduced the Bayesian LASSO method for estimating the regression coefficients. They used an idea from Tibshirani (1996) to connect the LASSO method with the Bayesian analysis. Tibshirani (1996) noticed that the LASSO estimates of the regression coefficients can be viewed as posterior mode estimates assuming that the regression coefficients have double exponential prior distributions. The Bayesian LASSO is also used in GS (e.g. de los Campos et al. 2009, de los Campos et al. 2010, Long et al. 2011) using the hierarchical model with the likelihood function

\[
f(y|\mu, X, m, \sigma^2) \sim N(\mu + Xm, \sigma^2 I),
\]

where \( y \) is the \( n \times 1 \) data vector, \( \mu \) is the overall mean vector, \( m \) is a vector of the marker effects, \( X \) is the design matrix that connects \( m \) to \( y \). \( N(\mu + Xm, \sigma^2 I) \) denotes the normal density with mean \( \mu + Xm \) and variance \( \sigma^2 I \) where \( I \) is an \( n \times n \) identity matrix. The prior distribution on the marker effects \( m_j \)s \( j = 1...p \) can be written as \( p(m_j|\tau^2_j) \sim N(0, \tau^2_j) \), and the prior distribution on \( \tau_j \) is \( p(\tau_j|\lambda) \sim Exp(\lambda) \) where \( Exp(\lambda) \) denotes the exponential distribution with rate parameter \( \lambda \).

Park et al. (2008) and de los Campos et al. (2009) presented the full conditional distributions that were used to sample via the Gibbs sampler. De los Campos et al. (2009) expanded the model, and assigned a prior distribution to \( \lambda^2 \). The prior has a Gamma distribution with shape parameter \( \alpha_1 \) and scale parameter \( \alpha_2 \), and it can be written as \( p(\lambda^2) \sim \Gamma(\alpha_1, \alpha_2) \). \( \lambda \) has two interpretations. In the Bayesian formulation it is the rate parameter which controls the shape of the prior distribution of the \( \tau^2_j \)s. In the LASSO setting \( \lambda \) controls the penalty for minimizing
2.2.2 Nonparametric Methods in Genome-Wide Selection

In this section, we review some of the non-parametric estimation methods that have been proposed for the case where the form of the relationship between a response variable and a set of predictors is unknown. A popular approach, at least in terms of usage, is based on the kernel method proposed by Silverman (1986) in the context of density estimation. In that context, the goal is to estimate the unknown density using a smooth curve. The kernel method is the most commonly used nonparametric estimation procedure.

The kernel density estimator \( \hat{f}(x) \) can be written in the form

\[
\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right),
\]

where \( n \) is the number of observations, \( K \) is the kernel function which satisfies the condition \( \int K(x) dx = 1 \), \( h \) is positive real-valued smoothing parameter (also called window width or bandwidth), \( x \) is the focal point and \( X_i \) is the \( p \times 1 \) dimensional vector of dummy covariates for observation \( i \). We can calculate \( \hat{f}(x) \) at several focal points \( x \), and the observations that are closer to the focal point will get a higher weight in the calculation, so the kernel function \( K \left( \frac{x - X_i}{h} \right) \) gives bigger weight to observations closer to the focal point. The kernel function \( K \) is usually chosen to be a symmetric unimodal density, so the kernel density estimator \( \hat{f}(x) \) is also a density. A commonly used kernel function is the Gaussian kernel given by

\[
K \left( \frac{x_i - x}{h} \right) = \frac{1}{(2\pi)^{p/2}} \exp \left[ -\frac{1}{2} \left( \frac{x_i - x}{h} \right)' \left( \frac{x_i - x}{h} \right) \right].
\]

In this expression, observations with \( x_i \) coordinates closer to the focal point \( x \) are weighted more strongly in the computation of the fitted value \( \hat{E}(y|x) \). The window width provides information about the range of observations that are included. Figure 2.1. shows how the kernel
density estimation changes with different bandwidth values. Using simulated data from a mixture of two normal distributions, the second, third and fourth panels show how the estimation changes with the change of the bandwidth value.

Figure 2.1 The influence of the bandwidth in kernel density estimation. From left to right the first plot shows simulated data from a mixture of two normal distributions. The second, third and fourth plots show the Gaussian kernel density estimates using bandwidth values $h = 0.1$, $h = 2$, and $h = 10$.

When $h = 0.1$, the data have strong influence in the density estimate, resulting in little bias, and large variability among estimates. It is called an undersmoothed estimate. As we increase the bandwidth value, the estimates become smoother. When $h = 10$, the spread is too big, and even the bimodal feature of the data disappears, which implies that the estimate is oversmoothed. Setting the bandwidth too large results in a large bias with little variance.

2.2.2.1 Nadaraya-Watson Estimator

In the context of genomic selection, Gianola et al. (2006) considered the regression function

$$ y_i = g(x_i) + e_i, \quad (2.8) $$

$i = 1, 2, ..., n$ where $y_i$ phenotypic measurement on individual $i$, $x_i$ is a $p \times 1$ vector of dummy SNP covariates observed on individual $i$, $g(.)$ is some unknown function relating genotypes
to phenotypes, $g(x_i) = \mathbb{E}(y_i|x_i)$, and $e_i$ is a random residual effect for individual $i$ where $e_i \sim (0, \sigma^2)$, and is independent of $x_i$. The conditional expectation function can be written in the form

$$g(x) = \frac{\int y p(x,y) dy}{p(x)}.$$ 

A nonparametric kernel estimator (Silverman, 1986) can be used to obtain an estimate of $p(x)$. The estimator has the form

$$\hat{p}(x) = \frac{1}{nh^p} \sum_{i=1}^n K \left( \frac{x_i - x}{h} \right), \quad \int_{-\infty}^{\infty} \hat{p}(x) dx = 1,$$

where $x_i$ is the observed $p$-dimensional SNP genotype of individual $i$, $i = 1, 2, ..., n$. Similarly,

$$\hat{p}(x, y) = \frac{1}{nh^{p+1}} \sum_{i=1}^n K \left( \frac{y_i - y}{h} \right) K \left( \frac{x_i - x}{h} \right).$$

Using these expressions, Nadaraya (1964) and Watson (1964) showed that the conditional expectation function can be written as

$$\hat{E}(y|x) = \hat{g}(x)$$

$$= \frac{\int y \hat{p}(x,y) dy}{\hat{p}(x)}$$

$$= \frac{1}{nh^p} \sum_{i=1}^n y_i K \left( \frac{x_i - x}{h} \right)$$

$$= \frac{1}{nh^p} \sum_{i=1}^n K \left( \frac{x_i - x}{h} \right) \sum_{i=1}^n y_i w_i(x), \quad w_i(x) = \frac{K \left( \frac{x_i - x}{h} \right)}{\sum_{i=1}^n K \left( \frac{x_i - x}{h} \right)}.$$

The estimator is just a weighted sum of the observations $y_i$, $i = 1...n$ and is called the Nadaraya-Watson estimator.

The selection of the bandwidth, $h$, value is challenging. Hardle (1990) discussed several approaches to select $h$, including the leave-one-out cross validation, penalizing functions and plug-in methods. Gianola et al. (2006) used the leave-one-out cross-validation approach to
select the bandwidth. In this approach, first exclude the $i^{th}$ observation $(y_i, x_i)$, and fit the model to the other $n - 1$ observations. Using the marker information, predict $\hat{g}(x_i|h)$. This is repeated for all $n$ observations. The cross validation (CV) criterion is (Clark 1975):

$$CV(h) = \frac{\sum_{i=1}^{n} [y_i - \hat{g}(x_i|h)]^2}{n}.$$ 

The CV estimate of $h$ is the value of $h$ that minimizes $CV(h)$.

### 2.2.2.2 Reproducing Kernel Hilbert Space

Gianola et al. (2006) proposed a semi parametric kernel mixed model approach, where they combined the nice features of a nonparametric model (described above) with a mixed model framework. The model can be written as

$$y_i = w_i'\beta + z_i'u + g(x_i) + e_i,$$  \hspace{1cm} (2.9)

where $i = 1, 2, ..., n$, $\beta$ is a vector of fixed unknown effects (e.g. physical location of an individual), $u$ is a $q \times 1$ vector of additive genetic effects, $w_i'$ and $z_i'$ are known incidence vectors, $g(x_i)$ is an unknown function of the SNP data and the vector of residuals, $e$ is assumed to have a $N(0, \sigma^2_e)$ distribution. The vector containing additive genetic effects, $u$ is distributed as $N(0, A \sigma^2_u)$ where $\sigma^2_u$ is the additive genetic variance and $A$ is the additive relationship matrix. The authors suggested two different methods for estimation in this model. The first strategy, denoted “Mixed Model Analysis”, consists in a two-step approach, where a “corrected” data vector $y_i - g(x_i) = w_i'\beta + z'u + e_i$ in the second step of the analysis. A Bayesian approach can also be used where one can draw samples from the pseudo posterior distribution $[\beta, u, \sigma^2_u, \sigma^2_e | y^*]$, and then form semi parametric draws of the total genetic value.

The other method they suggested is the “Random $g(.)$ function” approach where it is assumed that $\beta, u$ are known. In this case, $g(x_i)$ can be estimated as

$$\hat{g}(x|\beta, u, y, h) = \hat{E}(y_i - w_i'\beta - z_i'u|x) = \sum_{k=1}^{n} w_k(x)(y_k - w_k'\beta - z_k'u),$$

and draws of $\beta^{(j)}, u^{(j)}$ can then be obtained from the distribution $[\beta, u, \sigma^2_u, \sigma^2_e | y^*, h]$. Finally, Gianola et al. (2006) discuss estimation in the Reproducing Kernel Hilbert Space (RKHS)
mixed model. The setup is similar to the mixed model approach, but estimation of model parameters is carried out using a penalized sum of squares approach. As before, the model can be written as

$$y_i = w_i' \beta + z_i' u + g(x_i) + e_i,$$  \hspace{1cm} (2.10)$$

where $i = 1, 2, ..., n$. The penalized sum of squares is given by

$$SS(g(x), h) = \sum_{i=1}^{n} [y_i - w_i' \beta - z_i' u - g(x_i)]^2 + h||g(x)||,$$

where the penalty $||g(x)||$ is a function of the second derivatives of $g(x)$. The goal is to find $g(x)$ that minimizes the penalized SS. Wahba (1999) showed that the minimizer can be written as

$$g(.) = \alpha_0 + \sum_{j=1}^{n} \alpha_j K(., x_j),$$

where $K(., .)$ is the reproducing kernel.

### 2.2.2.3 Support Vector Machine Regression

Support Vector Machine (SVM) was proposed by Vapnik and discussed by Cortes and Vapnik (1995). SVM is a supervised learning technique, which was originally developed as a classifier. A training data set is used to develop a maximum margin classifier that produces the largest possible separation between two classes of observations. In the linearly separable case, if observations $(x_i) \in \mathbb{R}^p$, the separator is a hyper-plane in $\mathbb{R}^{p-1}$.

Since fitting a regression model essentially consists in finding an optimal projection of the observations on a lower-dimensional hyper-plane, the idea can be used to estimate the unknown regression function subject to restrictions. The reader can refer to Hastie et al. (2009), Steinwart and Christmann (2008), and Christianini and Shawe-Taylor (2000) for a review of SVM. SVM regression was adopted by Maenhout et al. (2007) and Long et al. (2011) for GS in plant breeding. A nice feature of SVM regression in plant breeding applications is that the
relationship between the marker genotypes and the phenotypes can be modeled with a linear or nonlinear mapping function which takes samples from a predictor space to an abstract, multidimensional feature space (Hastie et al. 2009).

Suppose that we have a training sample \( S = \{(x_i, y_i), x_i \in \mathbb{R}^n, y_i \in \mathbb{R}, i = 1...n\} \), where \( x_i \) is a \( p \) dimensional vector containing the genotypic values for the \( p \) markers for individual \( i \), and \( y_i \) is the phenotypic value for individual \( i \). A model that describes the relationship between the phenotype and the genotype of an individual can be written as

\[
 f(x) = b + wx, \tag{2.11}
\]

where \( b \) is a constant and \( w \) is a vector of unknown weights. The constant \( b \) reflects the maximum error we are willing to commit when estimating the weights \( w \). We learn about the function \( f(x) \) by minimizing the expression \( \lambda \sum_{i=1}^{n} L(y_i - f(x_i)) + \frac{1}{2}||w||^2 \). \( L(.) \) denotes the loss function which measures the quality of the estimation. The regularization parameter \( \lambda \) quantifies the trade-off between the sparsity and the complexity of the model. Increasing \( \lambda \) implies a higher penalty on the error. The norm \( ||w|| \) of vector \( w \) is inversely associated with model complexity; by choosing \( w \) to minimize \( ||w|| \), we reduce model complexity.

There are many loss functions used for SVM regression. Some of the popular loss function choices include the squared loss, absolute loss and the \( \epsilon \)-insensitive loss. Here, we present these loss function formulations.

1. The squared loss function has the form \( L(y - f(x)) = (y - f(x))^2 \). It scales the loss quadratically by the size of the error. Using this loss function indicates that outliers are also weighted quadratically which requires the user to deal with the outliers before the regression analysis.

2. The absolute loss function has the form \( L(y - f(x)) = |y - f(x)| \). The absolute loss function scales the loss linearly by the size of the error eliminating the difficulty of using data sets with outliers.
3. The \( \epsilon \)-insensitive loss function has a form

\[
L(y - f(x)) = \begin{cases} 
0 & \text{if } |y - f(x)| < \epsilon \\
|y - f(x)| - \epsilon & \text{otherwise}
\end{cases},
\]

where \( \epsilon \) determines the number of support vectors used in the regression function. By definition (Vapnik 1995, Vapnik and Vashist 2009) a support vector is a vector \( x_i \) which satisfies the equation \( y_i(wx_i + b) = 1 \). Increasing \( \epsilon \) implies that fewer support vectors are used in the fitting. The \( \epsilon \)-insensitive loss function ignores the errors in the regression that have size less than \( \epsilon \). When the error is greater than \( \epsilon \), the loss is \( |y - f(x)| - \epsilon \).

Figure 2.2 illustrates the absolute loss, squared loss and \( \epsilon \)-insensitive loss functions as a function of the error \( y - f(x) \). In the remainder we focus on the \( \epsilon \)-insensitive loss function,

![Graphs of loss functions](image)

Figure 2.2 Loss functions used for SVM regression. The first panel shows the absolute loss function. The second panel is the square loss function, and the last panel is the \( \epsilon \)-insensitive loss function.

which needs a more robust representation to account for the noise in the data. We can add extra “cost” (or allow for additional uncertainty) by introducing non-negative “slack variables” \( \xi \)
constrained as follows (Long 2011): \( \xi_{1i} \geq y_i - f(x_i) - \epsilon \) where \( i = 1, \ldots, n \) is the number of training observations, and \( \xi_{2i} \geq f(x_i) - y_1 - \epsilon \) where \( i = 1, \ldots, n \). We can now re-write the objective function to be minimized as

\[
\lambda \sum_{i=1}^{n} (\xi_{1i} + \xi_{2i}) + \frac{1}{2} ||w||^2.
\]

The solution to this constrained minimization problem has the form

\[
\hat{f}(x) = \sum_{i=1}^{n} \alpha_i x_i x + b
\]

(Nocedal and Wright 1999). The solution depends on the training data through the inner product \( \langle x_i, x_j \rangle \), which is a linear function of the observations.

To take advantage of higher dimensional feature spaces we can introduce the data via non-linear functions. For example, we can replace the inner product of the data by a kernel function

\[
k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle.
\]

Some commonly used kernel functions include:

1. The linear kernel \( k(x, z) = \langle x, z \rangle \)

2. The Gaussian Radial Basis Function \( k(x, z) = \exp(-\sigma ||x - z||^2) \) where \( \sigma \) is the bandwidth parameter

3. The Laplace Radial Basis Function \( k(x, z) = \exp(-\sigma ||x - z||) \).

The solution to the minimization problem can also be written as a function of the kernel function. The resulting expression is \( \hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) + b \). The choice of the kernel function as well as of the tuning parameters \( \lambda, \epsilon, \) and \( \sigma \) is not straightforward. Because optimizing SVMs is not the focus of the manuscript, we refer the reader to Cherkassy and Ma (2004).

### 2.2.2.4 Neural Networks

Neural networks represent a nonparametric prediction procedure that captures additivity and epistasis by being able to model linear and complex non-linear functions. The original idea
of neural network came from the theory of how neurons in the human brain work and interact, and how the brain conducts computations. In the neural network every unit is analogous to a brain neuron, and the connections between them are analogous to synapses (Hastie et al. 2009). The first introduction of neural networks in the context of brain architecture was presented by Alexander Bain (1873) and William James (1890). McCulloch and Pitts (1943) developed a mathematical model for neural networks.

The basic layout of a neural network is a two-stage network with three types of layers: an input layer, a hidden layer and an output layer. This model is called the feed-forward neural network and is illustrated in Figure 2.3.

Figure 2.3  A tree-layer feed-forward neural network with $K$ input layer units, $L$ hidden layer units, and $M$ output layer units.

Figure 2.3 shows a diagram of a tree layer feed-forward neural network with $K$ input layer units, $L$ hidden layer units, and $M$ output layer units. $H_1, H_2, ..., H_L$ are called hidden layer units because they are not directly observed. When the neural network is used to estimate a regression function, there typically is only one output layer unit. The hidden layer units are functions of linear combinations of the inputs, and the output layer units are functions of the hidden layer units. The output function of a feed-forward neural network can be expressed in the following form:

$$f(I_k) = \beta_0 + \sum_{l=1}^{L} \beta_l \sigma(w_l, b_l, I_k), k = 1, 2, ..., K,$$

(2.12)

where $K$ is the number of units in the input layer, $I_k$ is the $k^{th}$ input, $\beta_0 \epsilon R^M$ is the intercept
(bias terms), \( M \) is the number of output layer units, \( L \) is the number of hidden layer units, \( \beta_l \) (\( l = 1, 2, ..., L \)) are the output layer weights connecting the \( l^{th} \) hidden layer unit to the output layer units, \( \sigma \) is the activation function modeling the connection between the hidden layer and the output layer, \( w_l \in R^K \) and \( b_l \in R \) are the unknown learning parameters of the hidden layer unit \( l \) (\( l = 1, 2, ..., L \)) connecting the \( k^{th} \) neuron in the input layer to them (Romero and Alquézar 2012).

In genomic selection typically \( I_k \) represents a vector of predictors (marker genotypes or other information) collected on individual \( k \) (\( k = 1, 2, ..., K \)) where \( K \) is the number of individuals in the analysis. The activation function \( \sigma \) is typically chosen to be the sigmoid (logistic) or the Gaussian radial basis function.


### 2.3 Materials and Methods

For the purpose of illustrating the parametric and non-parametric prediction approaches, simulated data were created by the R (R Development Core Team 2008) package QTL Bayesian Interval Mapping (“qtlbim”) (Yandell et al. 2012). R can be downloaded from http://www.r-project.org, the qtlbim package can be accessed by library(qtlbim) in R, and the description of the package can be found at http://cran.r-project.org/web/packages/qtlbim/qtlbim.pdf. The reader can refer to Yandell et al. (2007) for detailed information about the qtlbim package. There are other publications as well where the qtlbim package is used to implement statistical methods. Some examples include Yi et al. (2007), Yi and Shriner (2008), and Piao et al. (2009). For comparing methods, we used a simulated \( F_2 \) population with specifications listed in Table 2.1.

We simulated four sets of phenotypic and genotypic information for a \( F_2 \) and a BC population. The results for the BC population can be found in the supporting information section. For each sets we created 20 replicates, which yielded to a total of 80 phenotypic and 80 genotypic data sets. Within each replicate we created 25 different training-testing data sets. Half of the data sets assume only additive effects and half assume only epistatic effects without
Table 2.1  Specification of the simulated $F_2$ population. The table contains information about the genetic architecture and the heritability.

<table>
<thead>
<tr>
<th>Genetic architecture</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>additive</td>
<td>0.70</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.70</td>
</tr>
<tr>
<td>additive</td>
<td>0.30</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.30</td>
</tr>
</tbody>
</table>

any additive effects. We only evaluated the two extreme genetic architectures. Finally, for each genetic architecture we generated data with two different narrow sense heritabilities. The low heritability was determined to be 0.30, and the high heritability was 0.70. For each of the simulated combinations of population, genetic architecture and heritability the data contain phenotypic information for 1000 individuals and genotypic information for 2000 biallelic markers (the possible values coded as “A”, “H”, and “B”) for each individual. Out of the 1000 individuals, 800 were chosen randomly to be in the training set to fit the model, and 200 individuals were in the testing set to predict for. The qtlbim package uses Cockerham’s model as the underlying genetic model. The simulated genome has 10 chromosomes, each having a specified length. The 2000 markers were distributed throughout the genome in such a way that each chromosome had 200 markers and the markers were equally spaced over the chromosomes. We assumed no missing genotypic values and no missing phenotypic values. The phenotypic values are normally distributed.

For the additive model we placed 2 QTL on each of the 10 chromosomes with either positive or negative additive effect. For the additive model we assumed no epistatic interaction.

For the epistatic model we only considered two-way interactions between the QTL. The interacting QTL were at the same genomic location as the QTL for the additive model, and only neighboring QTL were associated with each other resulting 10 two-way epistatic interactions, each having either positive or negative epistatic effect on the phenotype. For the epistatic model we assumed that the QTL contributed no additive effect. The phenotypic values were
drawn from a normal distribution, and are based on the $P = G + E$ model. Figure 2.4 shows the histograms of the simulated phenotypic values for the 4 population - genetic architecture - heritability combinations.

Figure 2.4 The histogram of the simulated phenotypic values. The histograms represent the distribution of the phenotypic values for the $F_2$ population.

To compare the performance of the methods, we used cross validation, where we divided the data into training sets and testing sets. The training sets were used to fit the models, and the testing sets were used to determine the performance of the particular method. The performance of the methods was calculated by the accuracy of prediction and the MSE. We define accuracy of prediction as the correlation between the true phenotypic values and the predicted phenotypic values. We evaluated parametric methods including parametric least squares (LS) regression, Ridge regression, Bayesian Ridge regression, best linear unbiased prediction (BLUP), Least Absolute Shrinkage and Selection Operator (LASSO), Bayesian LASSO, Bayes A, Bayes B, Bayes C and Bayes C$\pi$. We also evaluated nonparametric methods including Nadaraya-Watson estimator, Reproducing Kernel Hilbert Space (RKHS) method, Support Vector Machine, and Neural Network. To implement the parametric and nonparametric meth-
ods, the statistical software R and software written in C++ provided by the Animal Science Department at Iowa State University were used. Specifications of the parameters and inputs for each method are described below.

### 2.3.1 Least Squares Regression

Since the number of markers exceed the number of the individuals in all simulated data sets, Meuwissen’s (2001) idea was adopted, where we first performed simple regression by coding each marker genotype as −1, 1. After fitting the 2000 simple linear models - one for each marker, we choose 300 of the markers with the most significant p-values. Then, these 300 markers were included into a final model, and we simultaneously used them to fit a linear model. To carry out the linear regression, the *lm* function was used that can be found in the *stats* package (R Development Core Team 2008) in R. Finally, the prediction for the testing set was done by using the marker data for the testing set and the output for the estimated marker effects provided by the *lm* function.

### 2.3.2 Ridge Regression

For this method we used Gustavo de los Campos’ software written in R. The code implements the calculation of the Ridge regression estimation of the marker effects discussed in the Ridge regression section, and uses these estimates to carry out the prediction for the individuals in the testing set. For the procedure, all of the available phenotypic and marker information is used. In the estimation of the marker effect the penalty parameter $\lambda$ is chosen to have a value of $\frac{1-h^2}{h^2} Var(X)$ where $h^2$ is the narrow sense heritability and $Var(X)$ is the sum of the 2000 marker variances. For all of the scenarios, we used $h^2 = 0.4$. 
2.3.3 Bayesian Ridge Regression:

To fit the Bayesian Ridge regression model, the function BLR was used that can be found in the Bayesian Linear Regression (BLR) package (de los Campos and Rodriguez 2010) in R. For the specifications of the BLR function, we used a Gaussian prior for the marker effects with mean 0 and a common variance $\sigma^2_{BR}$ where $\sigma^2_{BR}$ is unknown. $\sigma^2_{BR}$ is assigned to have a scaled inverse $\chi^2$ distribution with degrees of freedom $df_{BR} = 5$ and scale parameter $S_{BR} = 0.01$. The residual variance $\sigma_E$ has a scaled inverse $\chi^2$ distribution with degrees of freedom $df_E = 4$ and scale parameter $S_E = 1$. The BLR function implements the Gibbs sampler, and the number of iterations is specified to be 20000. We used 2000 iterations for the burn-in period without any thinning. To fit the Bayesian Ridge regression, we used all available phenotypic and genotypic data.

2.3.4 BLUP

To implement Best Linear Unbiased Prediction, we used the mixed.solve function in R that can be found in the rrBLUP package (Endelman 2011). The available marker data was used as the design matrix for the random marker effects, and there was no fixed effect specified. The prediction was carried out using the marker data for the testing set and the output for the predicted marker effects provided by the mixed.solve function.

2.3.5 LASSO

To predict phenotypic values in the testing set using the Least Absolute Shrinkage and Selection Operator (LASSO) method, we used the glmnet function of the glmnet package (Friedman 2010) in R. For the initial parameter values the default setting was applied. The prediction was carried out with the tuning parameter, $\lambda$ that minimized the average cross-validation error ($cvm$).
2.3.6 Bayesian LASSO

To fit the Bayesian LASSO method, the function BLR of the Bayesian Linear Regression (BLR) package (de los Campos and Rodriguez 2010) in R was used. The regularization parameter, $\lambda$ is specified to be random, and has a Gamma prior distribution with shape parameter, $\alpha_1 = 0.53$ and rate parameter $\alpha_2 = 0.00005$. The residual variance $\sigma_E$ has a scaled inverse $\chi^2$ distribution with degrees of freedom $df_E = 4$ and scale parameter $S_E = (df - 2)(1 - h^2)Var(y)$ where we specify that $df = 4$, $h^2 = 0.5$, and $Var(y)$ is the phenotypic variance. The Gibbs sampler was applied with 20000 iterations, and 2000 iterations were in the burn-in period. The chain was not thinned.

2.3.7 Bayesian Alphabet

To implement the Bayes A, Bayes B, Bayes C, and the Bayes C$\pi$ models, software called GenSel (Version 2.12) was used. GenSel was written by Rohan Fernando and Dorian Garrick (2009) in C++, and is used for genomic selection in animal breeding populations. In GenSel Bayes A, Bayes B, Bayes C, and Bayes C$\pi$ have been implemented. We used the settings for the four methods, that are listed in Table 2.2.

2.3.8 Nadaraya-Watson Estimator

To use the Nadaraya-Watson estimator for predicting the phenotypic values in the testing set, first we formed the cross-validation criteria, and we evaluated it on a grid of values. We examined the cross validation criteria between 1 and 1000, and choose the value to be the bandwidth, $h$ that minimized the criteria. Table 2.3 shows the bandwidth values that minimized each of the 4 data combinations for the Nadaraya-Watson prediction. The code for calculating the optimal bandwidth value and for the prediction was written in R.
Table 2.2 The parameter specifications for Bayes A, Bayes B, Bayes C, and Bayes Cπ used in GenSel. The table contains the number of iterations used for chain length, and burn-in period, the genotypic variance, the residual variance, the degrees of freedom for residual variance, the degrees of freedom for marker variance, and the probability corresponding to having a 0 effect marker.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain length (number of iterations)</td>
<td>41000</td>
</tr>
<tr>
<td>Burn-in period (number of iterations)</td>
<td>1000</td>
</tr>
<tr>
<td>Genotypic variance for data with $h^2 = 0.30$</td>
<td>0.42</td>
</tr>
<tr>
<td>Genotypic variance for data with $h^2 = 0.70$</td>
<td>2.10</td>
</tr>
<tr>
<td>Residual variance for data with $h^2 = 0.30$</td>
<td>0.98</td>
</tr>
<tr>
<td>Residual variance for data with $h^2 = 0.70$</td>
<td>0.90</td>
</tr>
<tr>
<td>Degrees of freedom for residual variance</td>
<td>10</td>
</tr>
<tr>
<td>Degrees of freedom for marker variance</td>
<td>4</td>
</tr>
<tr>
<td>$\pi$</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 2.3 Bandwidth values used for each of the four combinations of genetic architectures and heritabilities for the Nadaraya-Watson prediction.

<table>
<thead>
<tr>
<th>Genetic architecture</th>
<th>Heritability</th>
<th>Bandwidth value</th>
</tr>
</thead>
<tbody>
<tr>
<td>additive</td>
<td>0.70</td>
<td>195</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.70</td>
<td>195</td>
</tr>
<tr>
<td>additive</td>
<td>0.30</td>
<td>205</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.30</td>
<td>205</td>
</tr>
</tbody>
</table>

2.3.9 Reproducing Kernel Hilbert Space

The RKHS regression based on methods and algorithms described in de los Campos (2010b), and the R implementation was developed by Gustavo de los Campos. To specify the RKHS regression, we choose the Gaussian reproducing kernel with the Euclidean distance for all of the eight combinations of genetic architectures, heritabilities and population types. We fitted the model using 3 arbitrarily chosen bandwidth values. We carried out the prediction for the testing set with each of the bandwidth values, and we averaged the 3 values of accuracy of selection, and the 3 MSE values.
2.3.10 Support Vector Machine

To implement the Support Vector Machine regression, we used the `ksvm` function of the `kernlab` package (Karatzoglou et al. 2004) in R. For the `ksvm` function we used epsilon-regression as the type, and the Radial Basis (Gaussian) kernel as the kernel function. After fitting the model, the `predict` function was used to carry out the prediction of the phenotypic values for the testing set. For the other input parameters, the default values were used.

2.3.11 Neural Network

We implemented the Neural Network model using the `brnn` function of the `brnn` package (Rodriguez and Gianola 2013) in R. This function fits a two-layer neural network. We first map the input information into some basis function. Then, the inputs of the Neural Network model are the marker-derived principal components. We specified the number of neurons to be 3 and the number of epochs to train to be 30 in the model. The other parameters were left at the default setting. For a detailed description of the application of the neural network using the R package `brnn`, the reader can refer to Pérez-Rodiguez et al. (2013).

2.4 Results and Discussion

We compared ten parametric and four nonparametric statistical GS methods. Comparisons were based on predicted accuracies of a simulated F2 progeny derived from crosses of inbred lines where genotypic variability was responsible for either 30% or 70% of the phenotypic variability. The underlying genetic architectures responsible for the genotypic variability consisted of 20 independently segregating biallelic loci that either contributed equally in an additive manner to a quantitative phenotype or through additive by additive epistatic interactions among 10 pairs of loci. Each GS method was applied to simulated 20 sets of progeny with 25 replicates for each of the four combinations of genetic architecture and heritability, which yielded 500 total replicates for each combination. Training sets were used to develop a model,
and the model was used to predict phenotypes in the testing sets. Training sets consisted of simulated phenotypes and 2000 marker genotypes for 800 random progeny while the testing sets associated with the training sets consisted of the same information for 200 progeny derived from the same cross. The accuracy of prediction was determined by calculating the correlation between the predicted phenotypic values for the 200 individuals in the testing set with the simulated phenotypic values for the same 200 individuals. The MSE values were determined by calculating the sum of the squared differences between the 200 predicted phenotypic values in the testing set and the 200 simulated phenotypic values, and then dividing the sum by 200. Tables 2.4 – 2.5 report the average prediction accuracies and standard errors (sampling variabilities) of the ten parametric and four non-parametric methods applied to the 500 replicates of the four combinations of genetic architecture and heritability. Tables 2.6 – 2.7 report the average MSE values and standard errors of the MSE values of the 14 methods applied to the 500 replicates of the four combinations of genetic architecture and heritability. Figures 2.5 – 2.8 each contain 14 boxplots of accuracy of prediction values for the 14 different methods. The boxplots show the distribution of the accuracy of prediction values for the 500 runs. Figures 2.9 – 2.12 each contain 14 boxplots of MSE values for the 14 different methods. In each figure, the first 10 boxplots are for the parametric methods, and the last 4 (shaded) are for the nonparametric methods. These boxplots show the distribution of the MSE values for the 500 runs. The first plot of Figure 2.13 shows the ratio of the accuracy averaged over the parametric methods (excluding the LS method as it is an outlier) and the accuracy averaged over the nonparametric methods, and the second plot of Figure 2.13 shows the ratio of the MSE averaged over the parametric methods (excluding the LS method) and the MSE averaged over the nonparametric methods. The left side of the plots show the ratios for the additive genetic architecture, and the right side of the plots show the ratios for the epistatic genetic architecture. These summary plots clearly show the advantage of using nonparametric methods when epistasis is present. In both heritability scenarios the parametric to nonparametric accuracy
ratio is lower for the epistatic genetic architecture than for the additive genetic architecture. The parametric to nonparametric MSE ratio is higher for the epistatic genetic architecture than for the additive genetic architecture.

Genetic architecture responsible for the genetic contribution to the phenotypes had the greatest impact on differences of accurate predictions among the GS methods. If the genetic architecture for the trait is due to additive by additive epistasis among ten pairs of independently segregating loci, parametric GS methods are unable to predict the phenotypes in the testing sets (shown in Figures 2.6, 2.8, 2.10, 2.12). In contrast, non-parametric methods, particularly the Nadaraya-Watson estimator, the RKHS, and SVM, provided predictions that are reasonably accurate, especially for traits with higher heritabilities (shown in Figures 2.6, 2.8). Our results are consistent with Gianola’s statement (2006) that nonparametric methods should be able to better predict phenotypes that are based on genetic architectures consisting of epistatic interactions. If the underlying genetic architecture is additive, parametric GS methods are slightly better than the non-parametric methods for both levels of heritability and types of segregating progeny. Both, the accuracy of prediction and the MSE results suggest the same about the models in terms of predictive performance. When additive effects are present, the LS regression performs the worst among the parametric methods, and the Nadaraya-Watson estimator performs the worst among the nonparametric methods (shown in Figures 2.5, 2.7, 2.9, and 2.11). When epistasis is present, the nonparametric Nadaraya-Watson estimator, the RKHS, and the SVM performs significantly better than the parametric methods (shown in Figures 2.6, 2.8, 2.10, and 2.12). Among the parametric methods, the LS regression has the highest accuracy of prediction values when epistasis is present. However, LS has the highest MSE values among the parametric methods as well when epistasis is present. It suggests that the LS method estimates the QTL effects from both loci involved in the epistasis more accurately than the other parametric methods in the F2 population. The parametric methods other than the LS are shrinking the QTL effects too much. Among the nonparametric methods, the NN showed poor predictive ability when epistasis is present. We know that NN is prone to over-fitting (Lawrence et al. 1997, Smith 1996) which would affect prediction ability. Most of the results are consistent with the fact that parametric approaches assume that the explanatory
variables fitted in the model are independent. When we only simulate additive effects, and no epistasis, the markers are assumed to be independent. In this case, we satisfy the parametric model assumption of having independent explanatory variables, so the parametric models have a larger predictive power than the nonparametric models. However, when we simulate epistasis, the markers are dependent, which violates the parametric model assumption. Nonparametric models can handle epistatic models without explicitly modeling the interactions.

Recently, the inability of parametric GS methods to predict has been observed in experimental data. MacKay et al. (personal communication) found parametric GS methods were unable to predict chill coma recovery, a quantitatively measured adaptive trait in Drosophila. Two-dimensional scans of the whole genome had previously revealed that the genetic architecture of this trait is composed primarily of interactions involving many loci. Thus, the simulated architectures used in our study are reasonable for many quantitative traits.

The clear distinctions of estimated accuracies and MSE values between parametric and nonparametric methods when underlying genetic architecture is epistatic suggests that data analyses consisting of a combination of parametric and non-parametric GS methods could be used as a diagnostic to reveal the prevalent genetic architecture of the trait. It is likely that the true underlying genetic architecture consists of mixtures of additive and epistatic genetic effects, so the inferential limits of applying pairs of GS methods to data from samples of breeding populations as a diagnostic needs further investigation. However, the first step is to look at the extremes in terms of genetic architecture.

Our results also suggest that if the goal of the research is to accurately predict the genotypic value of an individual, particularly for purposes of selection, and the underlying genetic architecture of the traits are not known, then it is best to use the nonparametric Nadaraya-Watson estimator, the RKHS or the SVM. Unfortunately, these methods do not provide interpretable inferences about relative weighting that is being applied to various regions of the genome, i.e., inferences about specific allelic contributions to the trait are limited. If the goal is genetic improvement and the underlying genetic architecture is known to be additive, then parametric GS methods will provide better predictions for selection. We (Xu et al. 2011) have previously hypothesized that if all specific desirable alleles are known, then gene stacking (genome
construction) based on optimization approaches will be more effective and efficient than GS approaches. Thus, in the interest of both immediate and long term genetic improvement goals, a combination of data analyses consisting of parametric, non-parametric GS methods as well as genetic mapping (Guo et al. 2013) should be applied to data derived from plant breeding populations.

While heritability did not affect the ability to distinguish among GS methods, it did affect estimated accuracies. When heritability is high and genetic architecture is additive, predictions are more accurate than for low heritability. When genetic architecture is based on epistasis and the trait exhibits low heritability, predictions are not very accurate for almost all GS methods. Even when the heritability is 0.70, the highest mean for prediction accuracy is 0.35, which indicates that further improvement of the models is necessary. Also, further research is needed to determine the affects of more complex plant breeding population structures. Typically, plant breeding population structures consist of inbred progeny derived from multiple crosses involving related and unrelated inbreds (Guo et al. 2011, Guo et al. 2013). Thus, GS needs to accurately predict phenotypes not only among subsets of progeny from related families within generations but also among generations of related and unrelated families.

In practice plant breeders do not know the genetic architecture responsible for quantitative traits and the dynamics of selection for genetic improvement will tend to favor alleles that contribute to additive components. Genetic improvement is affected not only by the underlying genetic architecture, but also by additional types of unpredictable genetic contributions including intra-locus dominance and genotype by environment interactions. Herein we have demonstrated the superior ability of the nonparametric Nadaraya-Watson estimator, the RKHS, and the SVM methods to accurately predict phenotypes for additive by additive inter-locus interactions. We hypothesize that nonparametric GS methods also will enable more accurate predictions of individual genotypic value for traits that are affected by dominance and genotype by environment interactions.
Table 2.4  The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 213 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 493 replicates, and for the rest, the calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>$F_2$, $h^2 = 0.70$, Accuracy</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>0.56</td>
<td>0.09</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>0.80</td>
<td>0.02</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>0.80</td>
<td>0.01</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>BLUP</td>
<td>0.80</td>
<td>0.01</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.82</td>
<td>-0.01</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>Bayes LASSO</td>
<td>0.81</td>
<td>0.01</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes A</td>
<td>0.81</td>
<td>0.00</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes B</td>
<td>0.81</td>
<td>0.01</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes C</td>
<td>0.81</td>
<td>0.01</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes Cπ</td>
<td>0.83</td>
<td>0.01</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>0.67</td>
<td>0.35</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>RKHS</td>
<td>0.76</td>
<td>0.29</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.78</td>
<td>0.33</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.77</td>
<td>0.05</td>
<td>0.03</td>
<td>0.09</td>
</tr>
</tbody>
</table>

2.5 Acknowledgment

We would like to thank Rohan Fernando (Animal Science Department, Iowa State University) and Gustavo de los Campos (Biostatistics Department, University of Alabama) for their valuable input and help with the software development for this project. Also, we gratefully acknowledge the help of Dorian Garrick (Animal Science Department, Iowa State University) for providing access to the GenSel software.
Table 2.5  The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.30$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 184 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 498 replicates, and for the rest, the calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>$F_2$, $h^2 = 0.30$, Accuracy</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>0.33</td>
<td>0.09</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>0.50</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>0.50</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>BLUP</td>
<td>0.50</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.50</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes Lasso</td>
<td>0.50</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes A</td>
<td>0.50</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes B</td>
<td>0.50</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes C</td>
<td>0.50</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes Cπ</td>
<td>0.50</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>0.40</td>
<td>0.16</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>RKHS</td>
<td>0.47</td>
<td>0.11</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.47</td>
<td>0.14</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.48</td>
<td>0.00</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Table 2.6  The mean and standard error of the mean squared error values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>$F_2, h^2 = 0.70$, MSE</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>3.10</td>
<td>5.10</td>
<td>0.36</td>
<td>0.53</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>1.30</td>
<td>3.24</td>
<td>0.12</td>
<td>0.29</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>1.27</td>
<td>3.14</td>
<td>0.13</td>
<td>0.29</td>
</tr>
<tr>
<td>BLUP</td>
<td>1.26</td>
<td>3.11</td>
<td>0.12</td>
<td>0.29</td>
</tr>
<tr>
<td>LASSO</td>
<td>1.17</td>
<td>3.10</td>
<td>0.11</td>
<td>0.26</td>
</tr>
<tr>
<td>Bayes LASSO</td>
<td>1.25</td>
<td>3.10</td>
<td>0.13</td>
<td>0.26</td>
</tr>
<tr>
<td>Bayes A</td>
<td>1.25</td>
<td>3.33</td>
<td>0.12</td>
<td>0.30</td>
</tr>
<tr>
<td>Bayes B</td>
<td>1.22</td>
<td>3.31</td>
<td>0.11</td>
<td>0.30</td>
</tr>
<tr>
<td>Bayes C</td>
<td>1.24</td>
<td>3.16</td>
<td>0.11</td>
<td>0.28</td>
</tr>
<tr>
<td>Bayes C$\pi$</td>
<td>1.11</td>
<td>3.11</td>
<td>0.11</td>
<td>0.27</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>2.59</td>
<td>2.91</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td>RKHS</td>
<td>1.54</td>
<td>2.76</td>
<td>0.14</td>
<td>0.25</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>1.40</td>
<td>2.76</td>
<td>0.14</td>
<td>0.26</td>
</tr>
<tr>
<td>Neural Network</td>
<td>1.47</td>
<td>3.13</td>
<td>0.15</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Table 2.7  The mean and standard error of the mean squared error values for the parametric and the nonparametric methods for the $F_2$ population with heritability $h^2 = 0.30$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>$F_2, h^2 = 0.30$, MSE</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>1.92</td>
<td>2.32</td>
<td>0.20</td>
<td>0.26</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>1.11</td>
<td>1.48</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>1.11</td>
<td>1.46</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>BLUP</td>
<td>1.11</td>
<td>1.42</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.11</td>
<td>1.40</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Bayes Lasso</td>
<td>1.11</td>
<td>1.42</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Bayes A</td>
<td>1.10</td>
<td>1.47</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Bayes B</td>
<td>1.10</td>
<td>1.46</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Bayes C</td>
<td>1.10</td>
<td>1.42</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Bayes C$\pi$</td>
<td>1.10</td>
<td>1.40</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>1.32</td>
<td>1.38</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>RKHS</td>
<td>1.15</td>
<td>1.39</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>1.16</td>
<td>1.40</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Neural network</td>
<td>1.14</td>
<td>1.41</td>
<td>0.11</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Figure 2.5  The boxplots of accuracy of prediction for the \( F_2 \) population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.6  The boxplots of accuracy of prediction for the \( F_2 \) population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.
Figure 2.7  The boxplots of accuracy of prediction for the $F_2$ population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.8  The boxplots of accuracy of prediction for the $F_2$ population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.
Figure 2.9 The boxplots of mean squared error for the $F_2$ population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.10 The boxplots of mean squared error for the $F_2$ population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.
Figure 2.11  The boxplots of mean squared error for the $F_2$ population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure 2.12  The boxplots of mean squared error for the $F_2$ population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.
Figure 2.13  Plots of the parametric to nonparametric accuracy and MSE ratios. The left side of the plots shows the additive cases, and the right side of the plots shows the epistatic cases.
Bibliography


de Boer, P. M. C., C. M. Hafner, 2005 Ridge regression revisited. Statistica Neerlandica 59(4) : 498 – 505.


relationship information on genomic breeding values in German Holstein cattle. Genet Sel
Evol. 42 : 5.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian


Hastie, T., R. Tibshirani, and J. Friedman, 2009 The Elements of Statistical Learning: Data
Mining, Inference, and Prediction. Springer.


Henderson, C. R., 1975a Use of all relatives in intraherd prediction of breeding values and

Henderson, C. R., 1975b Best linear unbiased estimation and prediction under a selection

Henderson, C. R., 1984 Applications of Linear Models in Animal Breeding. University of
Guelph, Guelph, Ontario, Canada.

Henderson, C. R., 1988 Progress in Statistical Methods Applied to Quantitative Genetics since
ciates, Inc., Sunderland, Massachusetts.


McCullock, W., and W. Pitts, 1943 A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5(4) : 115 – 133.


CHAPTER 3. RESPONSE SURFACE METHODOLOGY IN GENOMIC SELECTION

3.1 Abstract

Genetic architecture can have a significant impact on prediction accuracies of genomic selection methods. Parametric methods are unable to predict traits composed of epistatic genetic architectures, while nonparametric methods are able to provide reasonable predictions. These differences suggest a diagnostic for revealing genetic architectures underlying traits of interest. In addition to genetic architecture, the performance of genomic selection methods can be influenced by the number of individuals in the population, the number of QTL, the relative contributions of epistatic and additive variance to total genetic variance, and the heritability of the trait. Possible values of the factors can be numerous, and the number of combinations of the factors that influence the performance of the GS methods can be large. Therefore, it is useful to introduce a methodology to find the combination of attribute levels that results in accurate predictions for a given GS method. Herein, we introduce Response Surface Methodology (RSM) as a strategy for investigating genomic selection methods. We illustrate RSM with an example where we simulate backcross populations with different numbers of individuals, markers, QTL, and different percentages of epistasis and heritability. The response we optimize is the difference between prediction accuracy using the parametric best linear unbiased prediction method and the nonparametric support vector machine method. The greatest impact on the response is due to the genetic architecture of the population and the heritability of the trait. When epistasis and heritability are highest, the advantage of using the nonparametric support vector machine method versus the parametric best linear unbiased prediction method is greatest.
3.2 Introduction

Genomic selection (GS) is an approach for improving quantitative traits through the use of genetic information provided by dense molecular markers (Meuwissen et al. 2001). To implement GS, a model that includes marker, phenotypic and possibly pedigree information is developed to predict the phenotypic values of a trait for individuals in a population. Using GS, we can improve the accuracy of prediction and selection relative to the traditional phenotypic and marker assisted selection (Lande and Thompson 1990). The model is then used to predict the phenotype for individuals for whom only marker information is available. Genomic selection can improve genetic gain through selection relative to phenotypic selection because many more individuals can be assayed for phenotypes than field plot budgets will support. There have been several parametric and nonparametric statistical methods proposed for GS, and there are numerous articles evaluating these methods for particular populations under certain conditions (de los Campos et al. 2010, Zhao et al. 2012, Howard et al. 2014). The performance of the methods depends on the attributes of the population, including size of the population, marker density, narrow sense heritability, etc. Using simulation models, these factors can be varied and their effect on prediction performance can be evaluated.

In a previous publication (Howard et al., 2014) we simulated phenotypic and genotypic information for $F_2$ and backcross populations for traits with heritabilities of 0.30 and 0.70. Half of the simulated data sets had only additive genetic effects, and the other half had only 2-way epistatic genetic effects among 30 loci. All simulated data had phenotypic values for 1000 individuals, and genotypic values for 2000 biallelic markers. Using the simulated data, we compared the performance of 10 parametric and 4 nonparametric statistical methods in terms of prediction accuracy. The accuracy measure we applied was the Pearson correlation coefficient between the simulated phenotypic value and the predicted phenotypic value. We found that the genetic architecture had the greatest impact on how the methods performed; the nonparametric methods provided higher correlations between predictions and simulated values if the genetic architectures consisted of only epistatic genetic effects. Parametric methods provided no ability to predict if the genetic architecture of the trait consisted of only epistatic effects. The
results reported in Howard et al. (2014) suggest an analytical diagnostic that could reveal the underlying unknown genetic architecture of a trait in experimental data. Comparing prediction accuracies for a given phenotype using parametric and nonparametric statistical methods could help us determine whether additive or epistatic effects dominate the genetic architecture of a given trait.

In our previous work (Howard et al. 2014) we evaluated a limited number of factor combinations using the parametric and nonparametric methods, but GS could be influenced by many factors, such as number of individuals, number of markers, number of QTL, the proportion of epistatic variance relative to the proportion of additive variance, and the contribution of each to heritability. Further, it is possible that interactions among the factors influence GS. Our goal here is to examine how changing the combination of factors might affect the accuracy of prediction. Also, we want to find the combination of attributes for which the nonparametric methods outperform the parametric methods. Herein, we introduce Response Surface Methodology (RSM) to approximate the relationship between a set of factors that influence GS and accuracy of prediction. Howard et al. (2014) found that the parametric best linear unbiased prediction (BLUP) and the nonparametric support vector machine (SVM) tended to outperform other parametric and nonparametric approaches in terms of prediction accuracy, so we use these to represent each class of GS methods. We define prediction accuracy as the correlation between the predicted and the true (simulated) phenotypes. Below, we introduce the concept of RSM, and present the vocabulary and notation used in RSM literature. Then, we describe the simulated data used to illustrate how RSM can be applied to evaluate GS methods. Finally, we show results of the RSM analysis on the diagnostic proposed for the discovery of unknown genetic architectures in experimental data.

3.3 Materials and Methods

3.3.1 Response Surfaces and Approximations

Response surface methodology (RSM) is used to approximate a functional relationship between a response variable $y$, and a set of design variables (Khuri and Mukhopadhyay 2010).
In particular, RSM can be used to find the combination of factor levels for which the response variables are optimized. RSM was first introduced by Box and Wilson (1951), and is used in many disciplines, including physical, biological, environmental and chemical sciences, engineering and economics. The main advantage of RSM is that it can reduce the number of experimental runs required to find the optimum conditions associated with responses.

In RSM, we have numerous design variables, \( z_1, z_2, \ldots, z_p \), which are associated with a response (Myers and Montgomery 1995). When only two design variables influence a single response variable, it is simple to visualize the response at each design variable combination with a contour plot, or level curves.

To illustrate the two-design variable case, we simulated hypothetical yield, temperature and drought data. The response in this example is grain yield, which is influenced by temperature and by the degree of drought. Average daily temperature is simulated to be between 64°F = 18°C and 80°F = 27°C, and drought is between −4 and 4 SPI (Standard Precipitation Index). Negative values for the drought index indicate dryer than normal conditions. The model we used to simulate data is yield = 110 + \( \cos(0.25 \text{drought})^2 + \sin(0.15 \text{temperature})^2 + 0.0024375 \text{drought} \times \text{temperature} \). Figure 3.1 shows the response surface of the simulated yield. It shows the relationship between yield, and the design variables; average daily temperature and drought.

![Figure 3.1 Response surface of yield in relation to temperature and drought.](image-url)
Figure 3.2 shows the level curves, or contour plot for yield. For these simulated data it is clear that the yield is maximized when temperature is between $73^\circ F = 22.78^\circ C$ and $74^\circ F = 23.33^\circ C$ and drought is around 2 SPI. In most situations, the true response surface is unknown, and is influenced by more than two design variables. In these more realistic scenarios, visualization of the data is nearly impossible, unless we project high dimensional observations onto lower dimensional surfaces. In RSM we develop a methodology where we model the relationship between the response and the design variables, and explore the surface to find the levels of the design variables that optimize the response.

![Contour plot](image)

**Figure 3.2** Contour plot (level curves) of the response surface of yield.

The model describing the relationship between the response $y$, and the $p$ design variables, $z_1, z_2, ..., z_p$ can be written in the form

$$y = f(z_1, z_2, ..., z_p) + \epsilon,$$

(3.1)

where $f$ is the true, unknown response function, and $\epsilon$ is the error term (Myers and Montgomery 1995). The error term is often assumed to have a normal distribution with mean 0 and variance
\( \sigma^2 \), although other distributions can be modeled. Herein we assume that \( \epsilon \) has a distribution that has mean 0, the expected value of the response can be written as

\[
E(y) = E[f(z_1, z_2, \ldots, z_p) + \epsilon] 
\]

\[
= E[f(z_1, z_2, \ldots, z_p)] + E[\epsilon] 
\]

\[
= f(z_1, z_2, \ldots, z_p). 
\]

Variables \( z_1, z_2, \ldots, z_p \) are called natural variables because they are measured in the natural unit of measurements (Myers and Montgomery 1995) such as °F and SPI. We can also transform these variables (or “code the variables”) to have mean 0 and the same standard deviation. The re-scaled variables are denoted as \( x_1, x_2, \ldots, x_p \).

Since the true response function is unknown, we have to approximate \( f \). Under standard smoothness assumptions, a low-order polynomial function is a good local approximation to the true \( f \). For example a first-order main effects model can be written as

\[
E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, 
\]

where \( x_1, x_2, \ldots, x_p \) are coded variables, \( \beta_0 \) is the unknown intercept, and \( \beta_1, \beta_2, \ldots, \beta_p \) are the unknown regression coefficients. Equation 5. is called the main effects linear model because it only contains the linear effects of the \( p \) factors on the response, but no interactions terms. When the model in equation 5. includes interactions, we call it the first-order model with interaction, and write it as

\[
E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \beta_{12} x_1 x_2, 
\]

where in our illustration only the first two design variables interact. If needed, a second-order model can be also used to model the unknown response function, \( f \). A second order polyno-
mial is a good local approximation to almost any surface because it can have different functional forms, and it is easy to estimate its parameters. In general, the second-order model can be written as

$$E(y) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{j=1}^{p} \beta_{jj} x_j^2 + \sum_{i<j\leq p} \beta_{ij} x_i x_j.$$  \hspace{1cm} (3.7)

Let \( \beta \) denote the vector of unknown regression coefficients with dimension depending on the model. With an interaction term or a second-order model we can introduce curvature into the estimated surface.

### 3.3.2 Experimental Design and Estimation

To estimate the regression coefficients, \( \beta \) in the response function, we must obtain data by conducting an experiment based on a treatment design (Myers and Montgomery 1995). In RSM, a frequently used design is the factorial design, which in the simplest case has two levels per factor. In the previous illustration temperature and the degree of drought were the two factors, but the number of levels per factor can be more than two. The factor levels can be qualitative, quantitative. To observe a response at each factor combination when there are two levels for the \( p \) factors requires \( 2^p \) unreplicated treatment combinations, and the design is called the \( 2^p \) factorial design. When \( p \) is large and the range of possible values of each factor is also large, finding the combination of the \( p \) factors that optimizes the response requires a large number of treatment combinations; as the number of factors increases, the number of treatment combinations increases dramatically. For example for 3 factors with 2 levels each, the number of unreplicated treatment combinations is \( 2^3 = 8 \), but if we increase the number of factors to 5, the number of factor combinations is \( 2^5 = 32 \). At least some of those treatment combinations would need to be replicated if we want an estimate of the experimental error variance, \( \sigma^2_\epsilon \). In a \( 2^p \) factorial design only \( p \) degrees of freedom of the total of \( 2^p - 1 \) are used to estimate the main effects. To avoid the cost of running multiple large experiments, we adopt a sequential approach, where we explore the surface using subset of factorial experiments located in different regions of the factorial space. To move from one region to the next, we use steepest ascent or
other methods for finding the direction along which the response increases (or decreases). If we assume that the second order and higher order terms in the model are not significant, we can consider a fractional factorial design where we only run a fraction of the full factorial design.

We illustrate the fractional factorial design using half of the full factorial design which is also called the $2^{p-1}$ design. For example, if we have three factors (eg. temperature, degree of drought, and level of fertilizer which we denote $A, B,$ and $C$), the half factorial design is a $2^{3-1}$ design, which requires $2^{3-1} = 2^2 = 4$ treatment combinations. Even though the fractional factorial experiment is less expensive than the complete factorial experiment, the fractional factorial design always involves aliasing of effects. In other words; some effects are confounded and cannot be estimated independently. For example in the case of the $2^{3-1}$ design the main effects are confounded with the two-factor interactions. Thus, there is a “cost” associated with implementing a fraction of the experiment.

Table 3.1 shows all of the treatment combinations $a, b, c, abc, ab, ac, bc,$ and $(1)$ in a $2^3$ design. The factorial effects are $I, A, B, C, AB, AC, BC, ABC,$ where $I$ is the identity column used to estimate the average response. The $-$ symbol stands for a low level (Level 1) of a factor, and $+$ stands for a high level (Level 2) of a factor.

<table>
<thead>
<tr>
<th>Treatment Combination</th>
<th>Factorial Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
</tr>
<tr>
<td>a</td>
<td>+</td>
</tr>
<tr>
<td>b</td>
<td>+</td>
</tr>
<tr>
<td>c</td>
<td>+</td>
</tr>
<tr>
<td>abc</td>
<td>+</td>
</tr>
<tr>
<td>ab</td>
<td>+</td>
</tr>
<tr>
<td>ac</td>
<td>+</td>
</tr>
<tr>
<td>bc</td>
<td>-</td>
</tr>
<tr>
<td>(1)</td>
<td>+</td>
</tr>
</tbody>
</table>

The identity column, $I$ is always $+$, and we can write

$$I = ABC.$$  

(3.8)
Equation 3.8. is called the defining relation for the design, and represents the relationship between the identity and a particular factorial effect, which determines the aliasing pattern. Multiplying both sides of (8) by $C$ yields

$$C \times I = C \times ABC = ABC^2.$$  \hspace{1cm} (3.9)

However, the square of any column (factorial effect) is the identity $I$, so we get

$$C = AB.$$ \hspace{1cm} (3.10)

Using the defining relation, after defining the factorial effects for factors $A$ and $B$, we calculate the factorial effect of factor $C$ for every treatment combination. Table 3.1 shows the $+$ and $-$ signs for the factorial effects. To create a $2^{3-1}$ fractional factorial design we can consider the treatment combinations where the $ABC$ factorial effect has a $+$ sign. As we can see, these are the $a,b,c,$ and $abc$ treatment combinations which are listed in the top half of Table 3.1. We can also consider the 4 treatment combinations where the corresponding $ABC$ factorial effect has a $-$ sign. We call this the complementary fraction. These treatment combinations are the $ab, ac, bc,$ and (1). It does not matter which fraction we take because both of them belong to the same family.

### 3.3.3 Moving on the Estimated Response Surface

After designing the fractional factorial experiment, and running the 4 treatment combinations, we get at least four responses which provide information of the direction in which we need to move to obtain the second set of runs. One algorithm to do so is the method of steepest ascent (or steepest descent). Steepest ascent is a sequential approach for finding the maximum response, where we search for a region of the factor space where the response
is improved. Steepest descent also finds an optimum, but the term is used when the optimum is the minimum. The method of steepest ascent has three main steps (Myers and Montgomery 1995):

1. Experimental design
2. Model building
3. Sequential experimentation.

Since the method of steepest ascent is a sequential procedure, the three steps are typically repeated until we have found the region in factor space that is likely to contain the optimum response. That is, steepest ascent can involve several experiments, so careful planning of the steps and keeping the model and the experimental design simple is important.

In steepest ascent methodology, we conduct experiments along the path that leads to the maximum increase in the response. The coordinates of the steepest ascent path depend on the regression coefficients of the model. For illustrations consider a first-order regression model

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p. \]  

(3.11)

The movement of \( x_i(i = 1, 2, \ldots, p) \) relative to the other factors on the path of steepest ascent depends on the estimated regression coefficient \( b_i \). The magnitude of \( b_i \) determines how fast we are moving on the path relative to the \( x_i \) coordinate, and the sign of \( b_i \) tells us the direction of the movement. If \( b_i \) is positive, \( x_i \) will move in the positive direction on the path, and if \( b_i \) is negative, \( x_i \) will move in the negative direction on the path. If for example the magnitude of \( b_1 \) is twice as much as the magnitude of \( b_2 \), then \( x_1 \) will move twice as fast on the path as \( x_2 \). To illustrate the steepest ascent procedure, we use simulated data described in the following section.

Simple first-order models are typically used in the initial experiments. Unless the surface is complex or we have started far from the optimal region, we expect that only a few steps will be needed to move into the neighborhood of the optimum response. As we move closer to the optimum on the surface, we transition into second-order models that include interaction terms and curvature and that permit a better approximation of the underlying surface in the region of
interest. However, it is also possible that the additional steps of the steepest ascent procedure will not provide any new information for significant improvement compared to the first phase of the procedure where first-order model is used.

### 3.3.4 Scales and Other Considerations

In RSM, initial choices including the factors to consider and their range of values can have a large impact on the speed of approaching the optimal response. For example, for the factor temperature we can choose the range to be between 1 °F and 100 °F, or we can convert to the Celsius scale which will lead to a range between −17 °C and 38 °C. The estimated regression coefficients will be different depending on the temperature scale. Using different ranges of factors only influences the magnitude of the regression coefficients, but not the sign of the regression coefficients. This implies that using a different range for the factors would not modify the direction in which we are moving along the path, but it would change the speed of the movement relative to the scale used.

Another design aspect to consider is the choice of the metric for the response. Especially when the range of the response is large, it is useful to transform the response. One of the most commonly used transformations is the Box-Cox power transformation. The transformed response, \( w \) is defined as

\[
    w = \frac{y^\lambda - 1}{\lambda},
\]

where \( y \) is the untransformed response, and \( \lambda \) is the power parameter. For example, \( \lambda = -1 \) results in a reciprocal transformation, \( \lambda = 0.5 \) results in a square root transformation, and since \( \lim_{\lambda \to 0} \left[(y^\lambda - 1)/\lambda\right] = \ln y \), \( \lambda = 0 \) results in a logarithmic transformation. The power parameter \( \lambda \) can be estimated via maximum likelihood. This rank preserving transformation is also useful when we need to stabilize the variance of the response, or when the residual variance does not satisfy the normal assumptions. It is possible that the response surface has multiple peaks, and then more than one combination of the design variables satisfy the condition for
having the optimal response. It is also likely that the range of starting values do not include the global peaks, but in the procedure of steepest descent (or ascent) we would arrive in the required range.

### 3.3.5 Simulated Data

To illustrate the application of RSM to find the greatest impact of factors on GS, and show the steps in the implementation of the steepest ascent procedure, we simulated data with the R (R Development Core Team 2008) package QTL Bayesian Interval Mapping (“qtlbim”) (Yandell et al. 2012). R can be accessed from http://www.r-project.org, the qtlbim package can be obtained by library qtlbim (Yandell 2007) in R, and the reference manual of the package can be found at http://cran.r-project.org/web/packages/qtlbim/qtlbim.pdf. Other applications of qtlbim include Yi et al. (2007), Yi and Shriner (2008), and Piao et al. (2009).

In the simulations, we consider 5 factors with two levels each. The factors and the levels for the factors we consider are listed in Table 3.2.

Table 3.2 Specification of the two levels of the factors include \(n\), number of segregating progeny, \(m\), marker number, QTL number, proportion of genetic variance due to epistasis and narrow sense heritability. Epistasis 0 means that all of the genetic variance is additive variance, 0.5 epistasis means that half of the genetic variance is additive and the other half is epistatic.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Level 1</th>
<th>Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>200</td>
<td>1000</td>
</tr>
<tr>
<td>(m)</td>
<td>50</td>
<td>400</td>
</tr>
<tr>
<td>qtl</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>epi</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>(h)</td>
<td>0.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The initial experiment is a half-fractional factorial design, i.e., \(2^5 - 1 = 16\) experimental treatment combinations are examined. To evaluate the different treatment combinations, we used cross validation, where we divided the data into training sets and testing sets. The training sets were used to fit the models, and the testing sets were used to calculate the accuracy of prediction. Let \(\text{ph}_i\) denote the true phenotypic values at factor combination \(i\) and let \(\hat{\text{ph}}_i\)
denote the estimated phenotypic values at factor combination $i$. We define the accuracy of prediction $r(\hat{ph}, ph)$ as the correlation between the true phenotypic values and the predicted phenotypic values (Howard et al. 2014). We estimated the accuracy of prediction for each treatment combination using Best Linear Unbiased Prediction (BLUP) and Support Vector Machine (SVM) methods. BLUP is a parametric statistical procedure for prediction consisting a random effect term for the marker genotypes (Henderson et al. 1959, Henderson 1963, Bernardo 1994, Howard et al. 2014). SVM is a nonparametric machine learning technique that can model the relationship between the marker values and the phenotypes using a linear or a nonlinear mapping function (Vapnik 1995, Hastie et al. 2009, Howard et al. 2014). The response variable of interest is the difference between the accuracy of prediction of the SVM method and the BLUP method. For all of the treatment combinations we simulated marker and phenotypic data for a backcross population, and we created 20 replicates; within each replicate, we divided the marker and phenotypic data into 25 different training-testing data sets yielding a total of 500 replicates for each combination of factors. The training-testing data sets were created in a such way that a random 20% of the individuals belong to the testing set, and the remaining 80% belong to the training set.

The simulated genome has 10 chromosomes, each having the same length. The markers were distributed throughout the genome in such a way that each chromosome had the same number of markers equally spaced along the length of each chromosome. We simulated no missing genotypic values and no missing phenotypic values. The phenotypic values are normally distributed.

Howard et al. (2014) discussed the role of epistasis in GS, but only evaluated GS methods for the two extreme cases of epistasis; they considered only additive gene action with 0% epistasis, and 100% epistasis without any additive gene action. Herein, the two levels we considered for epistasis were 0 and 0.5 indicating the proportion of total genetic variance that is explained by epistatic variance. Cooper et al. (2002) pointed out the importance of epistasis in terms of accounting for the genetic variation of a trait. They discussed the insignificant contribution of multi-way gene interactions to genetic gain. Their research suggested that response to selection - on an adaptive surface - is unlikely to happen if the average number of genetic interactions is
much greater than two, suggesting that response to selection is not likely for genetic architectures consisting of multi-way interactions. Thus we only considered two-way gene interactions among QTL when we included epistasis in the model. For each model the QTL had positive or negative additive effect, and for every model we considered the same QTL locations. For the epistatic part of the models we only considered two-way interactions between the QTL, and we did not simulate higher order interaction among QTL.

3.4 Results

3.4.1 Implementing Response Surface Methodology for Evaluating Factors Affecting Genomic Selection

We wish to find the combination of factor levels associated with the optimal response, $y$. In our example, $y$ is the difference between the accuracy of prediction using the parametric BLUP and the nonparametric SVM method. The response, $y$ depends on a set of design variables $x_{ind}$, $x_m$, $x_{QTL}$, $x_{epi}$, and $x_h$, where $x_{ind}$ is the number of individuals in the simulated backcross population, $x_m$ is the number of markers, $x_{QTL}$ is the number of QTL, $x_{epi}$ is the percentage of epistasis contributing to the genetic variability, and $x_h$ is the narrow sense heritability. The model can be written as

$$y = f(x_{ind}, x_m, x_{QTL}, x_{epi}, x_h) + \epsilon,$$

where $y$ is the response, $f$ is the unknown, possibly complex response function, which depends on the design variables $x_{ind}$, $x_m$, $x_{QTL}$, $x_{epi}$, and $x_h$, and $\epsilon \sim iid N(0, \sigma_\epsilon^2)$. The expected value of the response function can be written as

$$E(y) = E[f(x_{ind}, x_m, x_{QTL}, x_{epi}, x_h) + \epsilon]$$

$$= f(x_{ind}, x_m, x_{QTL}, x_{epi}, x_h).$$

We use a first-order polynomial to approximate the response function, $f$, so that

$$E(y) = \beta_0 + \beta_1 x_{ind} + \beta_2 x_m + \beta_3 x_{QTL} + \beta_4 x_{epi} + \beta_5 x_h,$$
where $\beta_0$ is the intercept, $\beta_1$ is the regression coefficient associated with the number of individuals, $\beta_2$ is the regression coefficient associated with the number of markers, $\beta_3$ is the regression coefficient associated with the number of QTL, $\beta_4$ is the regression coefficient associated with the proportion of genetic variation due to epistasis, and $\beta_5$ is the regression coefficient associated with the degree of heritability. Since we have 5 factors, the full factorial design would include $2^5 = 32$ experimental treatment combinations. To reduce the number of treatment combinations, we utilize a fractional factorial design consisting of half of the treatments in the factorial design.

With the RSM, we wish to find an optimum response, and to do so we explore the entire space of the factors where the space of the factors can be defined as $S_{\text{ind}} \in Z^+$, $S_m \in Z^+$, $S_{\text{QTL}} \in Z^+$, $0 \leq S_{\text{epi}} \leq 1$, and $0 \leq S_h \leq 1$, where $S$ stands for space, and $Z^+$ denotes the set of positive integers. We use steepest ascent for this purpose. After setting up the fractional factorial design, we get the response, $(r_{\text{SVM}} - r_{\text{BLUP}})$ for all of the 16 treatment combinations. The accuracy of the BLUP method, the accuracy of the SVM method, and the response for the 16 treatment combinations are listed in Table 3.3. For each of the treatment combinations, the value of the accuracies represents the average correlation over the 500 replicates.
Table 3.3  Mean accuracy of BLUP, mean accuracy of SVM, and the response (difference of mean accuracy of SVM and mean accuracy of BLUP) for 16 treatment combinations.

<table>
<thead>
<tr>
<th>Treatment Combination</th>
<th>BLUP Accuracy</th>
<th>SVM Accuracy</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>200 ind, 50 m, 10 QTL, 0 epi, 0.5 h</td>
<td>0.62</td>
<td>0.59</td>
<td>-0.02</td>
</tr>
<tr>
<td>1000 ind, 50 m, 10 QTL, 0 epi, 0.2 h</td>
<td>0.40</td>
<td>0.37</td>
<td>-0.03</td>
</tr>
<tr>
<td>200 ind, 400 m, 10 QTL, 0 epi, 0.2 h</td>
<td>0.23</td>
<td>0.22</td>
<td>-0.01</td>
</tr>
<tr>
<td>1000 ind, 400 m, 10 QTL, 0 epi, 0.5 h</td>
<td>0.63</td>
<td>0.61</td>
<td>-0.02</td>
</tr>
<tr>
<td>200 ind, 50 m, 50 QTL, 0 epi, 0.2 h</td>
<td>0.05</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>1000 ind, 50 m, 50 QTL, 0 epi, 0.5 h</td>
<td>0.29</td>
<td>0.27</td>
<td>-0.03</td>
</tr>
<tr>
<td>200 ind, 400 m, 50 QTL, 0 epi, 0.5 h</td>
<td>0.28</td>
<td>0.28</td>
<td>-0.01</td>
</tr>
<tr>
<td>1000 ind, 400 m, 50 QTL, 0 epi, 0.2 h</td>
<td>0.23</td>
<td>0.21</td>
<td>-0.02</td>
</tr>
<tr>
<td>200 ind, 50 m, 10 QTL, 0.5 epi, 0.2 h</td>
<td>0.14</td>
<td>0.13</td>
<td>-0.01</td>
</tr>
<tr>
<td>1000 ind, 50 m, 10 QTL, 0.5 epi, 0.5 h</td>
<td>0.48</td>
<td>0.51</td>
<td>0.03</td>
</tr>
<tr>
<td>200 ind, 400 m, 10 QTL, 0.5 epi, 0.5 h</td>
<td>0.27</td>
<td>0.28</td>
<td>0.01</td>
</tr>
<tr>
<td>1000 ind, 400 m, 10 QTL, 0.5 epi, 0.2 h</td>
<td>0.22</td>
<td>0.19</td>
<td>-0.03</td>
</tr>
<tr>
<td>200 ind, 50 m, 50 QTL, 0.5 epi, 0.5 h</td>
<td>0.06</td>
<td>0.05</td>
<td>-0.01</td>
</tr>
<tr>
<td>1000 ind, 50 m, 50 QTL, 0.5 epi, 0.2 h</td>
<td>0.08</td>
<td>0.07</td>
<td>-0.01</td>
</tr>
<tr>
<td>200 ind, 400 m, 50 QTL, 0.5 epi, 0.2 h</td>
<td>0.05</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>1000 ind, 400 m, 50 QTL, 0.5 epi, 0.5 h</td>
<td>0.25</td>
<td>0.23</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

Using the results for the response, we can fit a simple regression model, and get the estimate of the regression coefficients. The estimate of the response is

\[
\hat{y}_{BLUP-SVM} = -1.4310^{-2} - 9.2710^{-6} ind - 1.3810^{-6} m - 7.7710^{-5} QTL + 2.6410^{-2} epi + 1.7110^{-2} h.
\]  

(3.17)

The magnitude of the coefficients determines in what direction we move along the surface. Looking at the model fit and the estimated coefficients, we can see that increasing the proportion of total variance due to epistasis and the heritability can improve the response.

To determine the direction in which we should move and at what rate in the response surface to improve the response, we define the base and calculate the increment of each factor. Table 3.4 shows the low and high levels (level 1 and level 2) of the factors, the average of the two levels of the factors, and the distance between the average levels and either of the levels. These values will be helpful for the later steps.

Since in our example, the fitted model has the largest estimated coefficient for epistasis, the basis is chosen to be 1% of epistasis. It corresponds to \[ \frac{1}{0.01} = 4 \] design units, and it will influence
Table 3.4  The levels of factors, the mean of the levels of the factors and half of the difference between the levels of factors.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Level 1</th>
<th>Level 2</th>
<th>( \frac{\text{Level1} + \text{Level2}}{2} )</th>
<th>( \frac{\text{Level1} + \text{Level2}}{2} - \text{Level1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ind</td>
<td>200</td>
<td>1000</td>
<td>600</td>
<td>400</td>
</tr>
<tr>
<td>m</td>
<td>50</td>
<td>400</td>
<td>225</td>
<td>175</td>
</tr>
<tr>
<td>qtl</td>
<td>10</td>
<td>50</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>epi</td>
<td>0</td>
<td>0.5</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>h</td>
<td>0.2</td>
<td>0.5</td>
<td>0.35</td>
<td>0.15</td>
</tr>
</tbody>
</table>

the increment (\( \Delta \)) of the other factors as well. For the other four factors the movement will be the following:

\[
\left( \frac{-9.271 \times 10^{-6}}{2.641 \times 10^{-2}} \right) 4 = -0.0014 \quad (3.18)
\]

for individuals,

\[
\left( \frac{-1.381 \times 10^{-6}}{2.641 \times 10^{-2}} \right) 4 = -0.0002 \quad (3.19)
\]

for markers,

\[
\left( \frac{-7.771 \times 10^{-5}}{2.641 \times 10^{-2}} \right) 4 = -0.0118 \quad (3.20)
\]

for QTL,

\[
\left( \frac{1.711 \times 10^{-2}}{2.641 \times 10^{-2}} \right) 4 = 2.5943 \quad (3.21)
\]

for heritability. Using the information above, we can calculate the values of the factors for which we will calculate response in terms of examining the response surface. They determine the movement along the path of each factors on the response surface. Table 3.5 shows the base and the increment for each of the factors that are used to calculate the positions of the factors that need to be next evaluated, and see whether the response can be improved by moving on the response surface in the specified direction.

The base is the mean of the low and the high levels of the factors. The increment is calculated as the product of \( \frac{\text{Level1} + \text{Level2}}{2} \) and the corresponding movement. For example for the number of individuals the increment is calculated as \( 400(-0.0014) \). Base+\( \Delta \), Base+2\( \Delta \), and Base+3\( \Delta \) indicate in what direction and what speed we have to move on the response surface for each factor. As Table 3.5 shows, we don’t move away from the base significantly for the number of individuals, and number of QTL, and the number of markers
Table 3.5  Base, increment and the coordinates of the steepest ascent for the number of individuals, number of markers, number of QTL, proportion of epistasis and the degree of heritability for the initial 16 experimental runs.

<table>
<thead>
<tr>
<th></th>
<th>Individuals</th>
<th>Markers</th>
<th>QTL</th>
<th>Epistasis</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>600</td>
<td>225</td>
<td>30</td>
<td>0.25</td>
<td>0.35</td>
</tr>
<tr>
<td>Increment</td>
<td>400(-0.0014)</td>
<td>175(-0.0002)</td>
<td>20(-0.0118)</td>
<td>0.25</td>
<td>0.15(2.5943)</td>
</tr>
<tr>
<td>(\Delta)</td>
<td>-0.56</td>
<td>-0.04</td>
<td>-0.24</td>
<td>0.25</td>
<td>0.39</td>
</tr>
<tr>
<td>Base+(\Delta)</td>
<td>599</td>
<td>225</td>
<td>30</td>
<td>0.5</td>
<td>0.74</td>
</tr>
<tr>
<td>Base+2(\Delta)</td>
<td>599</td>
<td>225</td>
<td>30</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>Base+3(\Delta)</td>
<td>598</td>
<td>225</td>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Base+4(\Delta)</td>
<td>598</td>
<td>225</td>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Base+5(\Delta)</td>
<td>597</td>
<td>225</td>
<td>29</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

stays the same. The percentage of epistasis and heritability are the only factors for which we need to move significantly on the response surface. For both epistasis and heritability, we have to move into the positive direction. At Base+3\(\Delta\) we are at the upper boundary of the surface for the percentage of epistasis, and for heritability. It indicates that after running the treatment combinations with factor levels for Base+\(\Delta\), Base+2\(\Delta\), Base+3\(\Delta\), not many more experimental runs will be necessary, because we reached the limits for epistasis and heritability, and there is not further improvement in \(y\) with respect to \(n\), the number of individuals, \(m\), the number of markers, and the number of QTL. The number of experimental runs depends on how far away we start the experimentation from the optimum and from the boundaries of the response surface.

Table 3.6 shows the description of the additional treatment combinations, the accuracy of BLUP, the accuracy of SVM, and the response for the additional runs.

As we can see in Table 3.6, the response values for the additional runs are larger, in most cases, than the response values were for the 16 initial treatment combinations. When the number of individuals are around 597, the number of markers is 225, the number of QTL is 29, the proportion of epistasis in the total genetic variance is 100\%, and the narrow sense heritability is 1, the response value is maximized. Since the increment of the number of individuals, the number of markers and the number of QTL is small, and the response values are close to each other in this area of the response surface, we cannot neglect the possibility of having the optimum
Table 3.6  Mean accuracy of BLUP, mean accuracy of SVM, and the response (difference of
mean accuracy of SVM and mean accuracy of BLUP) for the additional treatment
combinations for the additional runs.

<table>
<thead>
<tr>
<th>Addit. Runs</th>
<th>Treatment Combination</th>
<th>BLUP Ac.</th>
<th>SVM Ac.</th>
<th>Resp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base+Δ</td>
<td>599 ind, 225 m, 30 QTL, 0.5 epi, 0.74 h</td>
<td>0.36</td>
<td>0.37</td>
<td>0.01</td>
</tr>
<tr>
<td>Base+2Δ</td>
<td>599 ind, 225 m, 30 QTL, 0.75 epi, 1 h</td>
<td>0.29</td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>Base+3Δ</td>
<td>598 ind, 225 m, 29 QTL, 1 epi, 1 h</td>
<td>0.01</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>Base+4Δ</td>
<td>598 ind, 225 m, 29 QTL, 1 epi, 1 h</td>
<td>0.01</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>Base+5Δ</td>
<td>597 ind, 225 m, 29 QTL, 1 epi, 1 h</td>
<td>-0.01</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Base+6Δ</td>
<td>597 ind, 225 m, 29 QTL, 1 epi, 1 h</td>
<td>-0.01</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>Base+7Δ</td>
<td>596 ind, 225 m, 28 QTL, 1 epi, 1 h</td>
<td>0.00</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Base+8Δ</td>
<td>596 ind, 225 m, 28 QTL, 1 epi, 1 h</td>
<td>0.00</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Base+9Δ</td>
<td>595 ind, 225 m, 28 QTL, 1 epi, 1 h</td>
<td>0.00</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Base+10Δ</td>
<td>594 ind, 225 m, 28 QTL, 1 epi, 1 h</td>
<td>-0.01</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

not exactly at the case when the number of individuals is around 597, the number of markers
is 225, the number of QTL is 29 but close to it. However, it is certain that the optimum can be
found when 100% of the total genetic variance is explained by epistasis, and the heritability is 1.

3.5 Discussion

Originally, RSM (Box and Wilson, 1951) were developed to help design experiments to find
the combination of conditions that could be controlled to maximize the output of some process
(Naylor, 1969). Myers et al. (1989) summarized the extensive applications of RSM’s to engi-
neering systems, Bezerra et al (2008) summarized applications in analytical chemistry and more
recently RSM’s have been applied to systems of interest to biologists including pharmaceutical
production (Koyamada et al., 2004), water purification (citation), food processing (citation),
fermentation systems (Zhang et al., 2010). Despite clear efficiencies of response surface meth-
ods, they have not been utilized in agricultural or genetic improvement systems. This is due
in part to the misperception that RSM’s can only be applied to systems where the factors and
factor levels can be explicitly planned and controlled as fixed effect treatments (Menke, 1973).
Indeed, biological researchers in disciplines from Agronomy to Biotechnology often sequentially
evaluate levels of a single treatment in a series of experiments or attempt to design a full fac-
torial treatment experiment only to learn that the costs are prohibitive (consulting experiences of authors Beavis and Carriquiry). RSM’s were introduced to address these situations in all scientific disciplines. Indeed RSM’s have been shown to be effective in determining optimal conditions from sampling studies of complex systems where many of the factors are unknown and even unrecorded, e.g., ecosystem studies (Menke, 1973). We hypothesize that application of RSM's to 'big data' samples from molecular and cellular systems to cultivar x environment field systems will be effective and efficient at finding the optimal conditions of the system.

Herein, we have demonstrated that RSMs can be used to find the conditions under which applications of two different statistical methods for genomic selection can be optimized. The critical first step is to establish objective criteria that match goals of the study. It is possible that we might want to find the combination of factors that will maximize the accuracy of the predictions for each method. However, such criteria would not address the goals of the study. In this study we wanted to explore the conditions under which the difference between a non-parametric GS method and a parametric GS method was maximized. That is, we wanted to determine the conditions under which the two GS methods could be a diagnostic of the true underlying genetic architecture. Studies like Clowers et al. (2010), where they examine genetic variation associated with natural variation in cold tolerance in Drosophila, can gain advantage from diagnostics where the underlying genetic architecture can be revealed with help of comparing prediction accuracies for parametric and nonparametric statistical methods.

Our findings show that for this particular example the proportion of epistasis and heritability had the greatest impact on the response which supports the results of Howard et al. (2014). The difference between the parametric and the nonparametric GS methods is maximized in terms of accuracy of prediction at the limits of the parameter space for these two parameters. Thus, differences between parametric and nonparametric GS methods are minimal except when heritability and proportion of epistasis reach their maximum. We have to point out that for purposes of GS, some of the parameters (eg. heritability and genetic architecture) cannot be controlled by experimentalists, and they can only be modified as part of a simulation. Also, the prediction accuracy results are not exactly the same as results previously reported (Howard et al., 2014) because we modified the data simulation. Previously (Howard et al., 2014), we sim-
ulated the chromosomes with different length, and in this article we wanted to have uniform chromosomes in terms of length. Consequently, this changed the positions of the simulated QTL. This can lead to future work of examining how chromosome structure can influence prediction accuracies.

Even though we demonstrated RSM using a GS example, the methodology can be applied to any novel data analysis technique (eg. for genome-wide association study). We believe that this article gives the first description of RSM for GS, and how a response can be maximized using limited number of experimental runs. However, we understand that this area of research requires more extension and validation.
Bibliography


Henderson CR. Selection index and expected genetic advance. NAS-NRC Publ. 1963; 982.


CHAPTER 4. PREDICTING PROGENY’S PHENOTYPE USING PARENTAL PHENOTYPIC AND GENOTYPIC INFORMATION

4.1 Abstract

Genomic prediction techniques that take advantage of inexpensive marker data are widely used in plant and animal breeding. These techniques utilize whole genome DNA marker data and phenotypic information to model the association between the genotype and the phenotype, and then use the model to predict the phenotypes of individuals for which only marker information is available. Most genomic prediction research focuses on predicting phenotype within the same generation. However, it is important to recognize that genomic prediction could be more valuable in identification of the of the most likely parents to produce superior progeny. Herein, we use parametric and nonparametric genomic prediction techniques to identify crosses that are most likely to produce superior progeny. We investigate how precisely we can predict which crosses we should make to produce doubled haploid lines with the best phenotypic values through use of simulation models. We define a metric to evaluate genomic prediction techniques, where we compare simulated progeny having the highest phenotypic values with predicted progeny having the highest phenotypic values based on their parental phenotypic and genotypic values. To aid in the decision process we developed a simple metric based on the assumption that breeders are interested in the tails of distributions. We compared the methods for genetic architectures consisting of only additive effects and only epistatic effects. Results indicate that parents of progeny in the best 5% of the distribution of predicted individuals will be the same parents as those simulated to have the best progeny in 20 – 50% of the families depending on the prediction method, the heritability and genetic architecture. Generally, the Best Linear Unbiased Predictor and the Support Vector Machines methods performed simi-
larly for additive genetic architecture and high heritability, but when epistasis was simulated, the Support Vector Machines had higher percentage of identical parents than the Best Linear Unbiased Prediction method.

4.2 Introduction

Genomic Prediction (GP) is widely used in plant breeding because it can increase genetic gain, and lower the cost of delivering an improved variety. GP is a technique that utilizes DNA marker information throughout the whole genome and phenotypic values to predict the phenotype of cultivars for which phenotypic information is not available, thus GP enables breeders to select without extensive field testing. GP was first introduced by Meuwissen et al. (2001), but since then there is large variety of parametric and nonparametric statistical methods developed for GP (de los Campos et al. 2013; Gianola et al. 2010), and there is a large amount of effort devoted to comparing the methods (VanRaden et al. 2009; Daetwyler et al. 2010; Jannink et al. 2010; de los Campos et al. 2010; Clark et al. 2011; Heslot et al. 2012; Riedelsheimer et al. 2012; Lorenz 2013; Howard et al. 2014; Thavamanikumar et al. 2015). In GP, a training set consisting of phenotypic and genotypic marker information is used to model the relationship between the genotype and the phenotype of cultivars, and then the model is used to predict the phenotype of cultivars in the testing set. To date the training and testing sets are in the same generation. However, GP may have it’s greatest impact when used to predict distributions of progeny using the parents’ genotypic and phenotypic information.

Simulation techniques are useful to evaluate methods, because they mimic real-life situations. With the use of simulation techniques we can evaluate processes, methods, experiments, situations, etc. that are otherwise not feasible due to constraints (e.g. financial, time or place limitations). Geneticists first used computers to implement simulation models to evaluate limits to artificial selection (Hill and Robertson, 1966) in closed breeding populations. By 1988, Oscar Kempthorne, one of R.A. Fisher’s disciples, pointed out that the classical experimental and algebraic approaches were limited to unrealistic assumptions in breeding and evolutionary systems. Since 1988, plant and animal geneticists and breeders have used simulation models
to evaluate the limits to emerging statistical methods (Beavis, 1994) and to choose among selection methods because experimental evaluation of breeding methods is time and resource limited (Podlich et al., 1999). To date, there have been over 15,000 publications in which the terms simulation and breeding occur in the title. Currently there are numerous simulation software packages that have been developed and implemented for public and private research enterprises. Some are quite simple, while others are very flexible and complex.

In Howard et al. (2014) method comparisons were based on simulated phenotypic and genotypic data developed using progeny from a single cross. In our study we simulate genotypic and phenotypic data for a family structure consisting of doubled haploid (DH) parents and their progeny. The motivation for the particular approach of simulation is to mimic a real life situation of starting a breeding population with existing cultivars adapted to maturity zone 3. Thus, we are sampling from cultivars that represent the population structure of publicly available cultivars adapted to maturity zone 3. The advantage of simulation in our situation is that we can have information about all of the single crosses, their genotype and their phenotype. We can also produce as many progeny for each of the crosses as needed. When the number of possible crosses and the number of generated progeny are high, it is not feasible to run all of the experiments so simulation enables us to evaluate crosses without field testing. With simulation, each of the variables are controlled, and we can avoid confounding effects.

In previous simulations we assumed that all of the markers are polymorphic, thus the alleles at all of the loci are distinctive. Current high throughput technologies have generated a set of marker loci where some proportion of markers are polymorphic and some are monomorphic, thus the alleles can not be distinguished. For our simulations we used real marker information as a starting point to obtain a much more realistic simulation technique than only using polymorphic markers. In Howard et al. (2014), the authors evaluated parametric and nonparametric statistical methods for GP. They simulated data with different genetic architectures, and different levels of heritability. They compared GP methods in terms of the correlation between the simulated phenotypes and the predicted phenotypes which they called the accuracy, and the mean squared error (MSE) of the prediction. They concluded that the parametric methods performed similarly, and most of the nonparametric methods performed similarly in terms
of prediction accuracy and MSE. Since Howard et al. (2014) concluded that there is not a superior parametric GP method nor a superior nonparametric GP method, we only chose one parametric method and one nonparametric method to evaluate. To perform GP we used the Best Linear Unbiased Prediction (BLUP) and the Support Vector Machine (SVM) methods. The BLUP is a parametric technique (Henderson, 1949) used in GP where the genomic information is treated as a random effect in the mixed effect model. SVM is a nonparametric technique originally used for classification (Cortes and Vapnik, 1995). In both techniques, the prediction model is built based on the parents’ phenotypic and marker data. Then, the model is used to predict the progeny’s phenotype using the progeny’s marker data.

Instead of comparing prediction methods in terms of prediction accuracy and MSE, plant breeders are interested in predicting which crosses are most likely to produce the best progeny in terms of phenotypic value i.e. the phenotypic tails of the distribution. We compare the individuals with the highest phenotypic values in the simulated data set with the individuals with the highest phenotypic values in the predicted data, and calculate the proportion of common parents. We compare predictions performed with BLUP and SVM, and we simulate two different genetic architectures. Either only additive effects or only epistatic effects are present in the simulation data.

### 4.3 Materials and Methods

For the purpose of simulating genotypic and phenotypic data for doubled haploid parents and progeny, we used our own software (available at http://gfspopgen.agron.iastate.edu/resources.html). To start the simulation we used real SNP data consisting 4600 markers of 41 parents of the Soybean Nested Association Mapping (SoyNAM) project. The marker values were coded as \( \{0, 1, 2\} \) where the values 0, and 2 coded for the homozygotes. Stupar et al. (2013) give a description of the SoyNAM parents, and what crosses were performed for the SoyNAM project. We performed k-means cluster analysis (Hartigan and Wong, 1978) to designate 10 genetically diverse parents based on the 4600 markers, and we used them to simulate every possible
pairwise cross where we did not allow selfing. Since we generated doubled haploid genotypes, we changed the heterozygote genotypes (coded as 1) randomly in the SoyNAM parent genotype data to one of the two homozygote genotypes (coded as 0 or 2).

Using the marker information of the SoyNAM parents enabled us to have a more pragmatic simulation technique since only 31% of the markers are polymorphic on average. Figure 4.1 shows the proportion of polymorphic markers in a set of 45 crosses obtained by mating 10 SoyNAM parents. As Figure 4.1 shows at most 40% of the markers are polymorphic, which means that at least 60% of the markers are monomorphic (same in both parents at a locus). This simulation method enables us to model recombination more realistically since not 100% of the markers are polymorphic.

For each of the \(\frac{10 \times 9}{2} = 45\) crosses we simulated 100 DH progeny. The progeny had a simulated phenotype with mean 55, and 4600 biallelic markers. The phenotypes were generated based on the \(P = G + E\) model which can be interpreted that the phenotypic expression is determined by a genetic and a non-genetic factor, and where \(G\) was calculated depending on the genetic architecture. The simulated trait was influenced by 20 quantitative trait loci (QTL), and they represented the genetic variability either in an additive or an epistatic fashion. When we simulated additive effects, we used calculated the genetic term of the \(P = G + E\) model as the product of the QTL matrix and the corresponding additive effects that were calculated as shown in equation (1).

\[
a = \sqrt[\frac{\sigma^2_g}{n_{QTL}}], \tag{4.1}
\]

where \(a\) is the additive effect, \(\sigma^2_g\) is the genetic variance, and \(n_{QTL}\) is the number of QTL which is 20 in our case. The QTL were selected such that they are equally spaced throughout the genome.

Equation (1) can be derived as the following. Let’s recall that

\[
\sigma^2_g = \sigma^2_a + \sigma^2_d + \sigma^2_{epi}, \tag{4.2}
\]

where \(\sigma^2_a\) is the additive genetic variance, \(\sigma^2_d\) is the dominance genetic variance, \(\sigma^2_{epi}\) is the epistatic genetic variance, and \(\sigma^2_a = 2pq[a + d(q - p)]^2\) where \(p\) and \(q\) are the frequency of the two alleles in the genotype, and \(d\) is the dominance effect. Since we decided to not simulate any
dominance effect, \( d = 0 \), thus \( \sigma_a^2 = 2pq \sigma_a^2 \). For half of the simulated data sets we also assumed no epistasis, thus \( \sigma_g^2 = \sigma_a^2 = 2pq \sigma_a^2 \). From here, we can see that

\[
a = \sqrt{\frac{\sigma_g^2}{2pq}}.
\]  
(4.3)

Since in our case the expected value of \( p \) and \( q \) are 0.5, we can conclude that

\[
a = \sqrt{2\sigma_g^2},
\]  
(4.4)

and

\[
a = \sqrt{\frac{2\sigma_g^2}{n_{QTL}}}
\]  
(4.5)

for each of the simulated QTL.

When we simulate epistasis, the general formula (equation (3)) can also be applied for calculating the epistatic effect. However, the allele frequencies differ because of the interaction of the alleles. We simulated gene interaction among the 20 QTL in such a way that only the neighboring QTL interact, resulting 10 epistatic pairs. Since the alleles are coded as 0 or 1, we generated the epistatic pairs by multiplying the neighboring vectors of alleles in the QTL matrix. Since in our original QTL matrix the expected value of \( p \) and \( q \) were 0.5, in the epistatic matrix the expected frequencies were 0.25, and 0.75 because it is expected that 1/4 of the products of the alleles had a value of 1, and the rest had value of 0.

After altering the allele frequencies, the epistatic effect can be written as

\[
a = \sqrt{\frac{8\sigma_g^2}{3n_{QTL}}},
\]  
(4.6)

The genetic variance, \( \sigma_g^2 \) is calculated as

\[
\sigma_g^2 = \frac{r \sigma_e^2}{1 - r},
\]  
(4.7)

where \( r \) is the repeatability, and \( \sigma_e^2 \) is the error variance. In our software the repeatability and the error variance can be specified by the user, and we used \( r = 0.7 \) as the high repeatability, \( r = 0.3 \) as the low repeatability, and \( \sigma_e^2 = 64 \). To simulate genotypic and phenotypic information for a population we have to incorporate the recombination frequency in the simulation model. Instead of using a constant as the recombination frequency, we decided to use the
recombination frequencies in the SoyNAM parent genotypic data as the expected value of the recombination frequencies for simulation purposes. We compared these recombination values with random $U(0,1)$ values, which determined whether an actual recombination occurred at the given site of the genome.

The 4500 individuals that were simulated using 10 of the genetically most diverse SoyNAM parents represent the parental generation. We used all of the 4500 simulated individuals, and we phenotypically selected the best 50 to be parents of the next generation, resulting in $\frac{50 \times 49}{2} = 1225$ pairwise crosses excluding self-fertilization. Again, we generated 100 progeny for all of the crosses resulting a total of 122500 DH progeny in generation 2. Then, we performed prediction using the parametric BLUP and the nonparametric SVM methods. We examined the the performance with two different training sets. The training set for the prediction was either the phenotype and genotype of the 4500 individuals in the parental generation, or the phenotype and genotype of the top 50 individuals in terms of phenotypic performance chosen from the 4500 parental generation. We performed the predictions using one family at a time, and we repeated the process 1225 times.

In each of the 1225 families we calculated the 90th percentile of the simulated and predicted phenotypes. And we compared whether the best simulated and predicted phenotypes were from the same crosses.

We simulated DH family structure consisting 4500 parents and 122500 progeny using the genotypes of the SoyNAM parents as starting point. We used the SoyNAM parents’ recombination frequencies to establish recombination in the simulated data sets. We simulated high and low heritabilities (repeatabilities), and genetic architectures only consisting additive genetic variance or only epistatic genetic variance. We also predicted the phenotypes of the progeny using the parents’ genotypic anad phenotypic information. We either used all of the 4500 parents as the training set, or only the phenotypically best 50 individuals. For GP we utilized the parametric BLUP and nonparametric SVM methods, and we compared the phenotypically best simulated individuals with the phenotypically best predicted individuals. The comparison was based on the proportion of parents that were identical in the simulated and in the predicted sets.
4.4 Results and Discussion

The predictions were carried out in the 1225 families separately. In each of the 1225 predicted and simulated data sets, we calculated the 90\textsuperscript{th} percentile of the phenotypes, and we ranked the 1225 values. Next, we compared the parents of the highest 5\%, 10\%, 15\%, ..., 95\%, and 100\% of individuals, and we counted the number of parents that were the same in the simulated and the predicted data sets.

Figures 4.2 – 4.17 show the proportion of parents that are in the simulated and the predicted DH data sets for different genetic architectures, different values of heritability, size of training sample and different GP methods. Figures 4.2 – 4.17 indicate that among the top 5\% of the simulated and predicted progeny, about 20 – 30\% of their parents are identical. Whereas, among the top 50\% of the simulated and predicted progeny, about 60 – 70\% of their parents are identical. Generally the proportion of identical parents in the simulated and in the predicted data sets are higher when the 4500 parental genotype and phenotype was used in the training set instead of only the top 50 parents. When we look at the high heritability results we can conclude that in general the proportion of identical parents are higher than in the low heritability cases. There are a few replicates that do not follow this trend but we believe that it might be the consequence of sampling or non-genetic sources of variability. When additive genetic architecture is present, in most of the cases we notice more variability among the replicates when the top 50 parents are in the training set instead of the 4500 parents.

When we compare the data with additive genetic architecture with the data with epistatic genetic architecture for high heritability, the results agree with earlier findings (Howard et al. 2014) that when epistasis is present the nonparametric SVM method outperform the parametric BLUP method in terms of prediction when a single family is simulated. Howard et al. (2014) compared prediction accuracy, and found that when the genetic architecture consisted only additive genetic effects, the BLUP and SVM methods had similar prediction accuracy. However, when epistasis was present the SVM method produced higher prediction accuracy than the BLUP method. We could distinguish between BLUP and SVM more when heritability was high, whereas when heritability was low the difference was not as substantial. Here also, we
observe a clear difference between the parametric BLUP method and the nonparametric SVM method in terms of the proportion of identical parents when heritability is high and epistasis is present. When heritability is low and the genetic architecture is epistatic we can not conclude that the SVM method outperforms the BLUP method, but it is likely that it is due to sampling or non-genetic sources of variability. We have to point out that Howard et al. (2014) evaluated GP for single $F_2$ and $BC$ populations, and in this study we evaluate GP for structured DH families. We mimic starting a breeding program consisting of existing cultivars that were developed for maturity zone 3, and use GP techniques to help guide selection for different levels of heritability, different genetic architectures, and for different number of individuals being in the training set.

The results confirmed that when heritability is high and only additive genetic architecture is present in the simulated data the parametric and nonparametric methods perform similarly, and when only epistatic genetic architecture is present the SVM method outperforms the BLUP method. In our opinion the metric for comparing the low and high heritability, different genetic architectures, the size of the training set, and the GP methods is more sensible for practical use because the plant breeders are more interested in knowing which parents they have to cross to get improved progeny than the accuracy of prediction (based on personal communication with Tracy Doubler).

Also, our simulation technique where we use real genotypic information of the structured SoyNAM parents represents an improved method to simulate family structure compared to simulation techniques where no real data is used for the simulation, and at the starting point all of the markers are polymorphic.

We realize that future work is needed to evaluate more scenarios where we study the impact of different family structures, the distribution of the recombination frequency across the genome, the number of QTL affecting the trait, the magnitude of the environmental error, etc.
Figure 4.1  The proportion of polymorphic markers in the 45 crosses constructed by 10 genetically diverse SoyNAM parents.
Figure 4.2  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications.
Additive DH Population with Heritability=0.7, BLUP(50)

Figure 4.3 The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications.
Figure 4.4  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications.
Figure 4.5  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.7. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications.
Figure 4.6  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications.
Additive DH Population with Heritability=0.3, BLUP(50)

Figure 4.7  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications.
Figure 4.8 The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications.
Figure 4.9  The proportion of identical parents in the simulated and in the predicted DH data with additive genetic architecture and heritability of 0.3. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications.
Figure 4.10 The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications.
Figure 4.11  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications.
Figure 4.12 The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications.
Figure 4.13  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.7. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications.
Figure 4.14  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction BLUP was used with 4500 parents in the training set. The plot shows 5 replications.
Figure 4.15  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction BLUP was used with 50 parents in the training set. The plot shows 5 replications.
Figure 4.16  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction SVM was used with 4500 parents in the training set. The plot shows 5 replications.
Figure 4.17  The proportion of identical parents in the simulated and in the predicted DH data with epistatic genetic architecture and heritability of 0.3. For the prediction SVM was used with 50 parents in the training set. The plot shows 5 replications.
Bibliography


Riedelsheimer C., Technow F., Melchinger A. E., 2012 Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13 : 452.


The main objective of the dissertation was to evaluate parametric and nonparametric statistical methods for the purpose of predicting phenotypes in plant breeding populations using simulated data. While there are many existing methods in Genomic Prediction (GP), and much research has been conducted in this area, this dissertation offered new contributions by reviewing GP methods in detail, examining the difference among prediction methods when epistasis is present in the data, introducing Response Surface Methodology (RSM) in GP, simulating progeny with family structure instead of progeny from a single cross, and providing a metric that focuses on predicting the parents that need to be crossed to produce superior progeny.

In Chapter 2 we reviewed 10 parametric and 4 nonparametric statistical methods used in GP. We compared the methods based on accuracy of prediction and mean squared error (MSE) using simulated $F_2$ progeny. In half of the data the heritability was 70%, and in the other half it was 30%. We simulated progeny with only additive genetic architecture, and with only epistatic genetic architecture. The results show that there is no superior method among the parametric methods and there is no superior method among the nonparametric methods. Also, there is a clear advantage of nonparametric methods when epistasis is present in the simulated data. Heritability affected the estimated accuracy and MSE because when the heritability was high, the estimate of accuracy and MSE were higher than in the low heritability cases.

In Chapter 3 we introduced Response Surface Methodology as a systematic approach to evaluating GP methods. We were particularly interested to find the combination of attribute levels that results in an accurate phenotypic prediction, and to maximize the difference between prediction accuracy using the parametric Best Linear Unbiased Prediction (BLUP) method and the nonparametric Support Vector Machine (SVM) method. The results confirmed our find-
ings in the first research project. When epistasis is present and heritability is the highest, the advantage of using a nonparametric GP method versus using a parametric GP method is the greatest.

Chapter 4 is devoted to comparing parametric and nonparametric GP methods using simulated data from structured families. For this project we used real soybean genotypic data to mimic a real breeding program where 4500 parents and their 122500 progeny were simulated. Instead of evaluating the methods based on accuracy of prediction or MSE, we compared the individuals with the highest phenotypic values in the simulated data set with the individuals with the highest phenotypic values in the predicted data set, and calculated the proportion of common parents. This metric represents one of the plant breeders’ main objective better which is predicting the parents that need to be crossed to produce superior progeny. Using simulated family structure and this new metric for the evaluation of parametric and nonparametric GP resulted similar conclusion than in the first two research projects. There was an advantage of using nonparametric GP method when epistasis was present, but there was no clear distinction between the parametric and nonparametric methods when only additive genetic effects were simulated.

The results of the three research projects lead to many more questions. Future research projects include the study of recombination, and how recombination effect prediction accuracy. Introducing the new metric implies the need of comparing metrics (MSE, correlation, usefulness, selection of the tails of the phenotypic distributions) used for GP, and whether the selection decisions made based on the different metrics would be different.
APPENDIX  PARAMETRIC AND NONPARAMETRIC STATISTICAL METHODS FOR SIMULATED BC POPULATION

The parametric and nonparametric methods were also evaluated on a simulated backcross (BC) populations. The results for the simulated BC populations were similar to the results for the $F_2$ populations, so we are only providing the results for the BC populations in this section. The specifications of the BC populations are given in Table A.1. Figure A.1 shows the histograms of the simulated phenotypic values for the 4 population - genetic architecture - heritability combinations. Table A.2 shows the bandwidth values that minimized each of the 4 data combinations for the Nadaraya-Watson prediction.

Table A.1  Specification of the simulated BC population. The table contains information about the genetic architecture and the heritability.

<table>
<thead>
<tr>
<th>Genetic architecture</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>additive</td>
<td>0.70</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.70</td>
</tr>
<tr>
<td>additive</td>
<td>0.30</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Table A.2  Bandwidth values used for each of the four combinations of genetic architectures, heritabilities and population types for the Nadaraya-Watson prediction.

<table>
<thead>
<tr>
<th>Genetic architecture</th>
<th>Heritability</th>
<th>Bandwidth value</th>
</tr>
</thead>
<tbody>
<tr>
<td>additive</td>
<td>0.70</td>
<td>100</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.70</td>
<td>100</td>
</tr>
<tr>
<td>additive</td>
<td>0.30</td>
<td>105</td>
</tr>
<tr>
<td>epistatic</td>
<td>0.30</td>
<td>105</td>
</tr>
</tbody>
</table>

There is little difference in estimated accuracies between the $F_2$ and BC progeny that is not accounted for by differences in genetic architecture and heritability. While there is little impact on accuracy of GS in these two types of progeny from a single cross of inbred lines,
Figure A.1  The histogram of the simulated phenotypic values. The histograms represent the distribution of the phenotypic values for the $BC$ population.

Further research is needed to determine the affects of more complex plant breeding population structures.
Table A.3 The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the BC population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 183 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 494 replicates, and for the rest, the calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>BC, $h^2 = 0.70$, Accuracy</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>0.53</td>
<td>0.08</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>0.78</td>
<td>0.00</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>0.78</td>
<td>0.00</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>BLUP</td>
<td>0.78</td>
<td>0.01</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.80</td>
<td>0.00</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayes Lasso</td>
<td>0.79</td>
<td>0.00</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes A</td>
<td>0.78</td>
<td>0.00</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayes B</td>
<td>0.79</td>
<td>0.00</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayes C</td>
<td>0.79</td>
<td>0.00</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayes Cπ</td>
<td>0.81</td>
<td>0.00</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>0.66</td>
<td>0.23</td>
<td>0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>RKHS</td>
<td>0.77</td>
<td>0.18</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.76</td>
<td>0.22</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.75</td>
<td>0.00</td>
<td>0.03</td>
<td>0.07</td>
</tr>
</tbody>
</table>
The mean and standard error of the prediction accuracy values for the parametric and the nonparametric methods for the BC population with heritability $h^2 = 0.30$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations for the epistatic mean and epistatic s.e. for the LASSO method are based on 201 replicates, for the epistatic mean and epistatic s.e. for the Neural Network method are based on 494 replicates, and for the rest, the calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>BC, $h^2 = 0.30$, Accuracy</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>0.32</td>
<td>0.08</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>0.47</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>0.47</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>BLUP</td>
<td>0.47</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.47</td>
<td>-0.01</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes Lasso</td>
<td>0.47</td>
<td>0.00</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Bayes A</td>
<td>0.46</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayes B</td>
<td>0.47</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayes C</td>
<td>0.47</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>Bayes C$\pi$</td>
<td>0.46</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>0.38</td>
<td>0.09</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>RKHS</td>
<td>0.46</td>
<td>0.03</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.45</td>
<td>0.07</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.46</td>
<td>0.00</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Table A.5  The mean and standard error of the mean squared error values for the parametric and the nonparametric methods for the BC population with heritability $h^2 = 0.70$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>Method</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>3.07</td>
<td>4.52</td>
<td>0.38</td>
<td>0.49</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>1.27</td>
<td>2.74</td>
<td>0.13</td>
<td>0.26</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>1.24</td>
<td>2.63</td>
<td>0.12</td>
<td>0.22</td>
</tr>
<tr>
<td>BLUP</td>
<td>1.24</td>
<td>2.58</td>
<td>0.12</td>
<td>0.21</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.15</td>
<td>2.58</td>
<td>0.12</td>
<td>0.24</td>
</tr>
<tr>
<td>Bayes Lasso</td>
<td>1.22</td>
<td>2.59</td>
<td>0.13</td>
<td>0.23</td>
</tr>
<tr>
<td>Bayes A</td>
<td>1.27</td>
<td>2.88</td>
<td>0.13</td>
<td>0.25</td>
</tr>
<tr>
<td>Bayes B</td>
<td>1.23</td>
<td>2.83</td>
<td>0.13</td>
<td>0.25</td>
</tr>
<tr>
<td>Bayes C</td>
<td>1.22</td>
<td>2.65</td>
<td>0.13</td>
<td>0.22</td>
</tr>
<tr>
<td>Bayes Cπ</td>
<td>1.10</td>
<td>2.58</td>
<td>0.11</td>
<td>0.22</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>2.32</td>
<td>2.47</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td>RKHS</td>
<td>1.32</td>
<td>2.46</td>
<td>0.13</td>
<td>0.22</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>1.35</td>
<td>2.45</td>
<td>0.13</td>
<td>0.24</td>
</tr>
<tr>
<td>Neural network</td>
<td>1.43</td>
<td>2.58</td>
<td>0.16</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table A.6  The mean and standard error of the mean squared error values for the parametric and the nonparametric methods for the BC population with heritability $h^2 = 0.30$. The table contains the mean and standard error of the prediction accuracy values for both the additive and the epistatic cases. The first 10 methods are parametric, and the last 4 are nonparametric. The calculations are based on 500 replicates.

<table>
<thead>
<tr>
<th>Method</th>
<th>Additive mean</th>
<th>Epistatic mean</th>
<th>Additive s.e.</th>
<th>Epistatic s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least squares regression</td>
<td>1.96</td>
<td>2.33</td>
<td>0.21</td>
<td>0.27</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>1.09</td>
<td>1.41</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Bayesian Ridge regression</td>
<td>1.10</td>
<td>1.38</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>BLUP</td>
<td>1.09</td>
<td>1.33</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Lasso</td>
<td>1.09</td>
<td>1.35</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Bayes Lasso</td>
<td>1.10</td>
<td>1.33</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Bayes A</td>
<td>1.10</td>
<td>1.44</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>Bayes B</td>
<td>1.10</td>
<td>1.43</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>Bayes C</td>
<td>1.08</td>
<td>1.36</td>
<td>0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>Bayes Cπ</td>
<td>1.10</td>
<td>1.34</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Nadaraya-Watson estimator</td>
<td>1.26</td>
<td>1.32</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>RKHS</td>
<td>1.11</td>
<td>1.33</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>1.13</td>
<td>1.37</td>
<td>0.11</td>
<td>0.13</td>
</tr>
<tr>
<td>Neural network</td>
<td>1.12</td>
<td>1.34</td>
<td>0.11</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Figure A.2  The boxplots of accuracy of prediction for the BC population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure A.3  The boxplots of accuracy of prediction for the BC population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.
Figure A.4  The boxplots of accuracy of prediction for the BC population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure A.5  The boxplots of accuracy of prediction for the BC population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.
Figure A.6  The boxplots of mean squared error for the BC population with additive genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure A.7  The boxplots of mean squared error for the BC population with epistatic genetic architecture and heritability of 0.70. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.
Figure A.8  The boxplots of mean squared error for the BC population with additive genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.

Figure A.9  The boxplots of mean squared error for the BC population with epistatic genetic architecture and heritability of 0.30. The first 10 boxplots correspond to the parametric methods, and the last 4 (gray) boxplots correspond to the nonparametric methods.