Machines will be capable, within twenty years, of doing any work that a man can do.

– Herbert Simon, 1965.

University of Alberta

BAGGING E-BAYES FOR ESTIMATED BREEDING VALUE PREDICTION

by

Jiaofen Xu

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Fall 2009
Edmonton, Alberta

Examining Committee

Guohui Lin, Computing Science

Paul Stothard, Agricultural, Food and Nutritional Science

Stephen Moore, Agricultural, Food and Nutritional Science

Randy Goebel, Computing Science

Abstract

This work focuses on the evaluation of a bagging EB method in terms of its ability to select a subset of QTL-related markers for accurate EBV prediction. Experiments were performed on several simulated and real datasets consisting of SNP genotypes and phenotypes. The simulated datasets modeled different dominance levels and different levels of background noises.

Our results show that the bagging EB method is able to detect most of the simulated QTL, even with large background noises. The average recall of QTL detection was 0.71. When using the markers detected by the bagging EB method to predict EBVs, the prediction accuracy improved dramatically on the simulation datasets compared to using the entire set of markers. However, the prediction accuracy did not improve much when doing the same experiments on the two real datasets. The best accuracy of EBV prediction we achieved for the dairy dataset is 0.57 and the best accuracy for the beef dataset is 0.73.

Acknowledgements

First of all, I would like to express my sincerest gratitude to my supervisors, Dr. Guohui Lin and Dr. Paul Stothard, who have supported me throughout my thesis with their encouragement and useful suggestions. Their moral support and continuous guidance enabled me to complete my master's degree successfully.

I am also grateful to Dr. Zhiquan Wang and Dr. Rong-Cai Yang for their valuable advice and insights throughout my thesis.

I would like to thank Honghao Li for her hard work in providing me with a high quality dataset. I would also like to thank all the members of Dr. Guohui Lin's group, for their support and advice whenever I needed it. Very special thanks for the continual help and cooperation from Dr. Zhipeng Cai.

I also wish to acknowledge all of my friends for their helpful criticisms and encouragement. I have been blessed with so many great friends who would always cheer me up when I was depressed by the exhausting and disappointing experiments.

Finally, I thank my family for backing me up throughout my graduate studies at the University of Alberta, a place far away from home. I would like to thank them for making my mind relax during the hard times of my thesis, and for giving me unlimited happiness and pleasure.

Table of Contents

1 Introduction **1**

2 Background **3**
2.1 QTL Mapping . . . 3
2.1.1 Goals of QTL Mapping . . . 3
2.1.2 The Distribution of QTL Effects . . . 3
2.1.3 Linkage Mapping . . . 4
2.1.4 SNP . . . 5
2.1.5 QTL Mapping Based on LD . . . 5
2.2 Breeding Value Estimation . . . 6
2.2.1 Quantitative Genetics Approaches . . . 7
2.2.2 MAS . . . 7
2.2.3 Machine Learning . . . 8
2.3 Genomic Selection . . . 8
2.3.1 What is Genomic Selection? . . . 8
2.3.2 Why use Genomic Selection? . . . 9
2.3.3 Challenges for Genomic Selection . . . 9

3 Related Work **10**
3.1 QTL Mapping . . . 10
3.1.1 Single-QTL Models . . . 10
3.1.2 Multiple-QTL Models . . . 12
3.2 EBV Prediction . . . 13
3.2.1 Genomic Selection . . . 13
3.2.2 Factors Affecting the Accuracy of EBV Prediction . . . 15

4 Datasets and Methods **16**
4.1 Datasets . . . 16
4.1.1 Dairy Dataset . . . 16
4.1.2 Beef Dataset . . . 17
4.1.3 Simulation Datasets . . . 17
4.1.4 Data Pre-processing . . . 18
4.2 Methods . . . 19
4.2.1 Feature Selection Methods . . . 19
4.2.2 Regression Methods . . . 22

5 Experiment Results and Discussion **28**
5.1 QTL Mapping . . . 28
5.1.1 Effects of Dominance Model . . . 28
5.1.2 Effects of Noise Level . . . 29
5.1.3 QTL Mapping using Real Datasets . . . 29
5.2 EBV Prediction . . . 46
5.2.1 Experiment Design . . . 46
5.2.2 Performance Measurements . . . 46
5.2.3 Methods Implementation . . . 47
5.2.4 Comparison of Algorithm Performance on Real Datasets . . . 47
5.2.5 Comparison of Algorithm Performance on Simulation Datasets . . . 50
5.2.6 Co-dominance Representation *vs.* Binary Representation . . . 58
5.2.7 Experimental Results of the Bagging EB Feature Selection Method . . . 59
5.3 Additional Results . . . 66
5.3.1 Comparison of Algorithm Performance on Real Datasets . . . 66

5.3.2 Comparison of Algorithm Performance on Simulation Datasets 68
5.3.3 Experimental Results of the Bagging EB Feature Selection Method 80

6 Conclusions **91**
6.1 Conclusions 91
6.2 Future Work 92
6.2.1 Using Haplotypes 92
6.2.2 Including Non-additive Effects in Simulation Models 93

Bibliography **94**

List of Tables

5.1 SNP Mapping precisions and recalls of the bagging EB method on the 5 types of dominance models. α describes the dominance model and β describes the noise level, respectively, in the simulation type. 28
5.2 SNP Mapping precisions and recalls of the bagging EB method on co-dominant model with 5 noise levels. α describes the dominance model and β describes the noise level, respectively, in the simulation type. 29
5.3 Significant SNPs with fat yield identified by bagging EB method. 30
5.4 Significant SNPs with fat yield identified by Kolbehdari *et al.* [33] 30
5.5 Significant SNPs with fat percentage identified by bagging EB method. 33
5.6 Significant SNPs with fat percentage identified by Kolbehdari *et al.* [33] 34
5.7 Significant SNPs with milk yield identified by the bagging EB method. 34
5.8 Significant SNPs with milk yield identified by Kolbehdari *et al.* [33] 35
5.9 Significant SNPs with protein yield identified by the bagging EB method. 36
5.10 Significant SNPs with protein yield identified by Kolbehdari *et al.* [33] 37
5.11 Significant SNPs with protein percentage identified by bagging EB method. 39
5.12 Significant SNP with protein percentage identified by Kolbehdari *et al.* [33] 40
5.13 Significant SNPs with ADG identified by the bagging EB method. 42
5.14 Significant SNPs with birth weight identified by the bagging EB method. 43
5.15 Significant SNPs with RFI identified by the bagging EB method. 45
5.16 R implementation for machine learning regression methods. 48
5.17 The average CCs, rCCs, and NRMSEs of 10 methods for EBV prediction on 5 traits from the dairy dataset. 49
5.18 The average CCs, rCCs, and NRMSEs of 10 methods for EBV prediction on 3 traits from the beef dataset. 50
5.19 The average CCs, rCCs, and NRMSEs of 10 methods for EBV prediction on 10 types of simulation datasets defined by (α, β), where α denotes the dominance model and β denotes the background noise ratio. 56
5.20 EBV prediction results on all 8 traits from the two real datasets by SVM with a linear kernel, using two SNP genotype encoding schemes. Bold text indicates the encoding scheme with better performance. 59
5.21 EBV prediction results on all 10 types of simulation datasets by BLUP, using two SNP genotype encoding schemes. Bold text indicates the encoding scheme with better performance. 59
5.22 The average CCs, rCCs, and NRMSEs of SVM and LR, with or without using a SNP selection method, for EBV prediction on 8 traits of the two real datasets. . . . 60
5.23 EBV prediction results on 10 types of simulation datasets by LR, with or without using the bagging EB method for SNP selection. 66

List of Figures

2.1 Distribution of QTL effects, which shows that there are many loci with small effects, and few loci with large effects [23]. 4

4.1 The Soft Margin Loss Function for linear SVM [15]. 23

5.1 The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the fat yield trait. 31
5.2 The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the fat percentage trait. 34
5.3 The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the milk yield trait. 35
5.4 The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the protein yield trait. 37
5.5 The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the protein percentage trait. 40
5.6 The distribution of SNPs located by the bagging EB method for the ADG trait. . . 42
5.7 The distribution of SNPs located by the bagging EB method for the birth weight trait. 44
5.8 The distribution of detected SNPs with RFI on genome located by the bagging EB method. 44
5.9 Experiment procedure used to evaluate the performance of EBV prediction methods 46
5.10 Performance of EBV prediction algorithms on dairy dataset by CC measurement, which shows that for fat yield trait, Ridge is the best method; for fat percentage trait, PLS is the best method; for milk yield trait, SVM (linear kernel) is the best method; for protein yield trait, PLS is the best method and for protein percentage trait, Ridge is the best method. 48
5.11 Performance of EBV prediction algorithms on beef dataset by CC measurement, which shows that for ADG trait, PLS is the best method; for birth weight trait, GP is the best method and for RFI trait, GP is the best method. 50
5.12 Performance of EBV prediction algorithms for the completely recessive model by CC measurement, SVM (rbfdot kernel), GP and BLUP are the top 3 methods for this model. 51
5.13 Performance of EBV prediction algorithms for the partially recessive model by CC measurement, LASSO, BLUP and SVM (rbfdot kernel) are the top 3 methods for this model. 51
5.14 Performance of EBV prediction algorithms for the co-dominant model by CC measurement, BLUP, LASSO and Ridge are the top 3 methods for this model. 52
5.15 Performance of EBV prediction algorithms for the partially dominant model by CC measurement, LASSO, BLUP and SVM (linear kernel) are the top 3 methods for this model. 52
5.16 Performance of EBV prediction algorithms for the completely dominant model by CC measurement, BLUP, SVM (rbfdot kernel) and LASSO are the top 3 methods for this model. 53
5.17 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.1 model by CC measurement, LASSO, BLUP and PLS are the top 3 methods for this model. 53
5.18 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.3 model by CC measurement, BLUP, LASSO and PLS are the top 3 methods for this model. 54
5.19 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.5 model by CC measurement, BLUP, PLS and SVM (rbfdot kernel) are the top 3 methods for this model. 54

5.20 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.7 model by CC measurement, BLUP, GP and PLS are the top 3 methods for this model. 55
5.21 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.9 model by CC measurement, BLUP, GP and PLS are the top 3 methods for this model. 55
5.22 The CCs of BLUP EBV prediction on 100 simulation datasets from different dominance models. The average CCs are $0.928, 0.891, 0.801, 0.802$, and 0.552 for co-dominant, partially dominant, completely dominant, partially recessive and completely recessive, respectively. 57
5.23 The CCs of BLUP EBV prediction on 100 simulation datasets with different levels of background noise. The average CCs are $0.928, 0.827, 0.701, 0.600, 0.535$, and 0.486 for $\beta = 0, 0.1, 0.3, 0.5, 0.7$, and 0.9, respectively. 58
5.24 Algorithm performance with and without bagging EB feature selection method for the completely recessive model by CC measurement. 61
5.25 Algorithm performance with and without bagging EB feature selection method for the partially recessive model by CC measurement. 61
5.26 Algorithm performance with and without bagging EB feature selection method for the co-dominant model by CC measurement. 62
5.27 Algorithm performance with and without bagging EB feature selection method for the partially dominant model by CC measurement. 62
5.28 Algorithm performance with and without bagging EB feature selection method for the completely dominant model by CC measurement. 63
5.29 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.1 by CC measurement. . . . 63
5.30 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.3 by CC measurement. . . . 64
5.31 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.5 by CC measurement. . . . 64
5.32 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.7 by CC measurement. . . . 65
5.33 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.9 by CC measurement. . . . 65
5.34 Performance of EBV prediction algorithms on dairy dataset by rCC measurement, which shows that for fat yield trait, Ridge is the best method; for fat percentage trait, Ridge is the best method; for milk yield trait, Ridge is the best method; for protein yield trait, PLS is the best method and for protein percentage trait, Ridge is the best method. 66
5.35 Performance of EBV prediction algorithms on dairy dataset by NRMSE measurement, which shows that for fat yield trait, PCA is the best method; for fat percentage trait, ElasticNet is the best method; for milk yield trait, SVM with rbfdot kernel is the best method; for protein yield trait, ElasticNet is the best method and for protein percentage trait, ElasticNet is the best method. 67
5.36 Performance of EBV prediction algorithms on beef dataset by rCC measurement, which shows for ADG trait, GP is the best method; for birth weight trait, Ridge is the best method and for RFI trait, GP is the best method. 67
5.37 Performance of EBV prediction algorithms on beef dataset by NRMSE measurement, which shows that if using this measurement, for ADG trait, PCA is the best method; for birth weight trait, SVM with rbfdot kernel is the best method and for RFI trait, ElasticNet is the best method. 68
5.38 Performance of EBV prediction algorithms for the completely recessive model by rCC measurement, SVM (rbfdot kernel), GP and BLUP are the top 3 methods for this model. 68
5.39 Performance of EBV prediction algorithms for the completely recessive model by NRMSE measurement, SVM (rbfdot kernel), GP and PCA are the top 3 methods for this model. 69
5.40 Performance of EBV prediction algorithms for the partially recessive model by rCC measurement, BLUP, LASSO and SVM (rbfdot kernel) are the top 3 methods for this model. 69
5.41 Performance of EBV prediction algorithms for the partially recessive model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and LASSO are the top 3 methods for this model. 70
5.42 Performance of EBV prediction algorithms for the co-dominant model by rCC measurement, BLUP, Ridge and LASSO are the top 3 methods for this model. 70

5.43 Performance of EBV prediction algorithms for the co-dominant model by NRMSE measurement, BLUP, LASSO and PLS are the top 3 methods for this model. 71
5.44 Performance of EBV prediction algorithms for the partially dominant model by rCC measurement, LASSO, BLUP and SVM (linear kernel) are the top 3 methods for this model. 71
5.45 Performance of EBV prediction algorithms for the partially dominant model by NRMSE measurement, BLUP, LASSO and SVM (linear kernel) are the top 3 methods for this model. 72
5.46 Performance of EBV prediction algorithms for the completely dominant model by rCC measurement, BLUP, LASSO and SVM (rbfdot kernel) are the top 3 methods for this model. 72
5.47 Performance of EBV prediction algorithms for the completely dominant model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and LASSO are the top 3 methods for this model. 73
5.48 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.1 model by rCC measurement, LASSO, BLUP and PLS are the top 3 methods for this model. 73
5.49 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.1 model by NRMSE measurement, BLUP, LASSO and PLS are the top 3 methods for this model. 74
5.50 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.3 model by rCC measurement, BLUP, LASSO and PLS are the top 3 methods for this model. 74
5.51 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.3 model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and PLS are the top 3 methods for this model. 75
5.52 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.5 model by rCC measurement, BLUP, PLS and SVM (rbfdot kernel) are the top 3 methods for this model. 75
5.53 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.5 model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and PLS are the top 3 methods for this model. 76
5.54 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.7 model by NRMSE measurement, BLUP, GP and PLS are the top 3 methods for this model. 76
5.55 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.7 model by NRMSE measurement, SVM (rbfdot kernel), BLUP and GP are the top 3 methods for this model. 77
5.56 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.9 model by NRMSE measurement, BLUP, GP and PLS are the top 3 methods for this model. 77
5.57 Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.9 model by NRMSE measurement, GP, SVM (rbfdot kernel) and BLUP are the top 3 methods for this model. 78
5.58 Performance of EBV prediction algorithms on 5 dominance models by rCC measurement. 78
5.59 Performance of EBV prediction algorithms on 5 dominance models by NRMSE measurement. 79
5.60 Performance of EBV prediction algorithms on co-dominant model with 5 noise levels by rCC measurement. 79
5.61 Performance of EBV prediction algorithms on co-dominant model with 5 noise levels by NRMSE measurement. 80
5.62 Algorithm performance with and without bagging EB feature selection method for the completely recessive model by rCC measurement. 80
5.63 Algorithm performance with and without bagging EB feature selection method for the completely recessive model by NRMSE measurement. 81
5.64 Algorithm performance with and without bagging EB feature selection method for the partially recessive model by rCC measurement. 81
5.65 Algorithm performance with and without bagging EB feature selection method for the partially recessive model by NRMSE measurement. 82
5.66 Algorithm performance with and without bagging EB feature selection method for the co-dominant model by rCC measurement. 82
5.67 Algorithm performance with and without bagging EB feature selection method for the co-dominant model by NRMSE measurement. 83

5.68 Algorithm performance with and without bagging EB feature selection method for the partially dominant model by rCC measurement. 83
5.69 Algorithm performance with and without bagging EB feature selection method for the partially dominant model by NRMSE measurement. 84
5.70 Algorithm performance with and without bagging EB feature selection method for the completely dominant model by rCC measurement. 84
5.71 Algorithm performance with and without bagging EB feature selection method for the completely dominant model by NRMSE measurement. 85
5.72 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.1 by rCC measurement. . . . 85
5.73 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.1 by NRMSE measurement. 86
5.74 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.3 by rCC measurement. . . . 86
5.75 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.3 by NRMSE measurement. 87
5.76 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.5 by rCC measurement. . . . 87
5.77 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.5 by NRMSE measurement. 88
5.78 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.7 by rCC measurement. . . . 88
5.79 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.7 by NRMSE measurement. 89
5.80 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.9 by rCC measurement. . . . 89
5.81 Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.9 by NRMSE measurement. 90

Nomenclature

EBV	estimated breeding value
QTL	quantitative trait loci
SNP	single nucleotide polymorphism
LD	linkage disequilibrium
EB	empirical Bayes
SVM	support vector machine
GP	Gaussian process
PCA	principal component analysis
PLS	partial least square
Ridge	ridge regression
LASSO	least absolute shrinkage and selection operator
ElasticNet	elastic-net regression
MAS	marker assisted selection
GWAS	genome wide association study
LR	least square regression
ML	maximum likelihood
EM	estimation maximization
LRS	likelihood ratio statistic
SIM	simple interval mapping
MCMC	Markov chain Monte Carlo
BLUP	best linear unbiased prediction
BW	birth weight
MY	milk yield
PY	protein yield
FY	fat yield
PP	protein percentage
FP	fat percentage
CC	correlation coefficient
rCC	rank correlation coefficient
NRMSE	normalized root mean square error
RMSE	root mean square error
IBD	identical by descent

Chapter 1

Introduction

Animal evaluation programs based on phenotypes or *estimated breeding values* (EBVs) have been used in many countries to aid breeders in their selection decisions. The phenotypes of quantitative traits typically vary along a continuous gradient depicted by a Gaussian distribution and are attributed to *quantitative trait loci* (QTL) and their interactions with the environment [4]. QTL refer to chromosomal loci containing mutations that have effects on the phenotypic trait being assayed or measured.

Traditional phenotypic selection methods derive EBVs from phenotype and pedigree information. Though these methods have gained success in some economically important traits, they have many limitations [23]. The ideal situation for using the traditional selection methods is that the trait has high heritability, high quality data records and the phenotype can be observed in all individuals before reproductive age and with a relatively low cost. However, this ideal situation can hardly be achieved, which makes traditional selection methods costly or ineffective. Besides, the traditional selection methods lack detailed knowledge of the genetic architecture of the selected traits. If we can gain insight into the genetic architecture of the traits by using genetic markers, the selection progress may be greatly enhanced as we can know the actual chromosomal areas (*i.e.* QTL) affecting the traits. QTL mapping aims to identify QTL that are associated with the target phenotypic traits.

There are two main objectives in this research. One is QTL mapping to detect QTL for the target phenotypic trait. The other is EBV prediction, which is used for the selection of the best breeding animals. Due to the availability of a large volume of *single nucleotide polymorphism* (SNP) markers, we can exploit the *linkage disequilibrium* (LD) between SNP markers and QTL for QTL mapping and also use the genotype of SNP markers to predict individual EBV. As a result, the detection of QTL becomes the detection of the SNPs that are associated with the phenotypic trait. The focus of here is on the evaluation of a bagging *empirical Bayes* (EB) method, for its ability to select a subset of SNP markers that are significantly associated with the trait and then to use this subset of SNPs for accurate breeding value prediction. The evaluation was performed using several simulated and two real datasets consisting of genotypes and phenotypes.

The accuracy of EBV prediction is dependent on many factors, such as the number of samples and the proportion of the phenotypic variance attributable to genetic variation (*i.e.* heritability). The more phenotypic records available, the more observations there will be per haplotype, or per marker allele if single markers are used, and the higher the accuracy of QTL detection. However, for the SNP datasets we use, the number of SNPs is usually relatively large compared to the number of samples, which makes the detection of QTL-associated SNPs difficult. On the other hand, the higher the heritability, the more accurate the prediction of phenotype expected based on marker information, and the fewer records may be required while still achieving high prediction accuracy [23]. Nevertheless, high heritability cannot be guaranteed for most economically important traits.

The experiment design is as follows: For each dataset, the bagging EB method is used to identify potential trait-associated SNPs. These SNPs are then used in regression models to predict breeding values. Seven commonly employed machine learning regression models are used for the prediction: *support vector machine* (SVM), *Gaussian process* (GP), *principal component analysis* (PCA), *partial least square regression* (PLS), *ridge regression* (Ridge), *least absolute shrinkage and selection operator* (LASSO) and *elastic net* (ElasticNet). Ten types of simulation datasets are simulated, which model different dominance levels (completely recessive, partially recessive, co-dominant, partially dominant, completely dominant) and different levels of background noise. One of the real datasets is a dairy dataset, which contains 462 samples and 1341 SNPs. The other one is a beef dataset, which contains 433 samples and 47108 SNPs.

In summary, experimental results show that the bagging EB method is able to detect most of the simulated QTL even on datasets with large background noise. Also, when using the markers selected by the bagging EB method to predict breeding values, the prediction accuracy improved dramatically on the simulation datasets. However, the prediction accuracy did not improve much when using the markers selected by the EB method on the two real datasets. The best accuracy of EBV prediction achieve for the dairy dataset is 0.57 and for the beef dataset is 0.73. The top 3 regression methods for the dairy dataset are Ridge, SVM and PLS, and the top 3 regression methods for the dairy dataset are GP, Ridge and PLS.

The rest of this thesis is organized as follows: in the next chapter, some background information on QTL mapping and genomic selection are given. In the third chapter, we give a brief review of the related work in QTL mapping, EBV prediction. The fourth chapter presents the datasets and methods used in the experiment. The experimental design and results are given in the fifth chapter. Finally, chapter six concludes the thesis with some potential future research directions.

Chapter 2

Background

In livestock production systems, a large number of the economically important traits, such as milk yield and protein yield, are quantitative in nature. That is, they are measured to a numerical scale, as opposed to categorical variation traits (those that have two or several character values, *e.g.* disease, sex, eye color in humans). The phenotypes of quantitative traits typically vary along a continuous gradient depicted by a Gaussian distribution. The genetic variation of a quantitative trait is usually attributable to multiple genes and their interactions with the environment. For example, increases in milk production go together with increases in feed quality.

2.1 QTL Mapping

2.1.1 Goals of QTL Mapping

QTL refer to chromosomal loci each containing a gene that has an effect on the phenotypic trait being assayed or measured. QTL mapping is the technique that identifies QTL associated with that particular quantitative trait.

QTL mapping is a problem of great importance to biologists. First of all, QTL mapping can lead to the identification of the underlying genetic difference responsible for the phenotypic effect. Secondly, many QTL are associated with a particular trait and these QTL are often found on different chromosomes. Therefore, identification of QTL is critical for understanding the complexity of the genetic architecture of a trait. Moreover, knowledge of these loci can aid in selective breeding decision to improve the economically important trait [23].

2.1.2 The Distribution of QTL Effects

It has been verified that the amount of genetically inherited materials (*i.e.* genes) in the genome is finite, which means the number of loci underlying the variation in quantitative traits is also finite. In addition, the distribution of QTL effects may be described by a negative exponential distribution: there are only a few loci with moderate to large effects while the large number of remaining loci explain a relatively small portion of the phenotypic variability [45]. Although any locus with an

effect on the quantitative trait is called a QTL, we are only interested in the search for QTL that have moderate to large effects on the trait, and the use of this information to increase the accuracy of breeding selection. Figure 2.1 shows the number of QTL *vs.* QTL effects based on QTL mapping experiments in two datasets: pigs and dairy cattle [23].

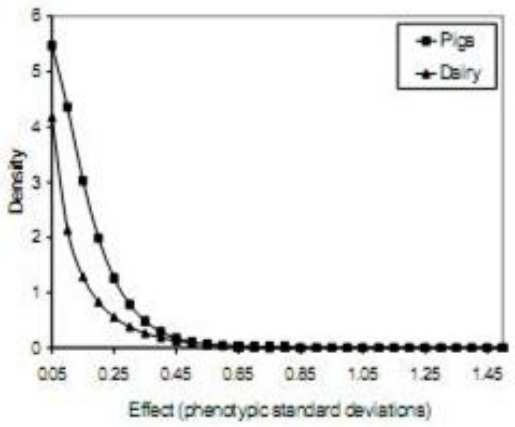

Figure 2.1: Distribution of QTL effects, which shows that there are many loci with small effects, and few loci with large effects [23].

2.1.3 Linkage Mapping

QTL detection using the linkage mapping approach has been performed for a large number of traits in livestock species for many years. To begin, a set of genetic markers must be developed, where a marker is an identifiable physical region of genome whose inheritance can be monitored. Markers are used because the actual genes that affect a quantitative trait are usually unknown and can only be identified using genetic markers that are linked to QTL. We can check the associations between allelic variation at the marker loci and variation in the quantitative trait to detect markers that are significantly more likely to co-occur with the trait than expected by chance. When there is evidence for co-occurrence, we say that the markers are linked to QTL which cause the variation of the quantitative trait.

Here is an example illustrating the procedures of QTL detection using the linkage mapping approach. We are assuming that a sire is heterozygous at a marker locus, and it has a large number of progeny. Suppose the alleles carried at this marker locus are ‘A’ and ‘B’ respectively. We can separate the progeny into two groups, those that receive allele ‘A’ and those that receive allele ‘B’. If there is a significant difference between these two groups for a particular quantitative trait, it indicates that this marker is linked to a QTL of that trait.

The disadvantage of the linkage mapping approach is that it requires thousands of progeny. Otherwise, the QTL are mapped to very large confidence intervals on the chromosome [1]. Such large confidence intervals will cause two problems. First, a large number of genes need to be investigated

in such large chromosome intervals if we want to identify the genes underlying the quantitative trait. Secondly, it will decrease the confidence of using QTL to assist selective breeding decisions, because the linkage between marker and QTL is not sufficiently ensured in such a large confidence interval.

Due to the availability of a large volume of SNP markers in livestock, we can exploit the LD between SNP markers and QTL to do QTL mapping, which overcomes the problem of linkage mapping as we can exploit LD in the whole population.

2.1.4 SNP

SNPs are an abundant source of genetic variation in the genome. It was shown that on average 99.9% of human's DNA sequence will be identical to that of another individual. However, over 80% of this 0.1% human genome variation comes in the form of SNPs [50]. Moreover, SNPs have a low mutation rate, which makes them useful in QTL analysis as markers in place of microsatellites.

As the name implies, a SNP marker is a DNA sequence difference occurring in a nucleotide between members of a species at a specific locus in the genome. For example, here are two sequenced DNA fragments from two people (DNA is comprised of four chemical entities, *i.e.* nucleotides A, G, C and T):

Individual 1: AGT**C**GCGC
Individual 2: AGT**T**GCGC

One can see that there is a difference in the fourth nucleotide, which is called a SNP if this variation is observed in the general population at a frequency greater than 1% [50]. With the advent of new molecular technology, a SNP marker can be genotyped in an individual for less than $1c$ USD.

2.1.5 QTL Mapping Based on LD

The classical definition of LD is the non-random association of alleles between two or more loci, and not necessarily on the same chromosome [26]. Consider two marker loci A and B each with two alleles. Suppose A has alleles A_1 and A_2, and B has alleles B_1 and B_2. If we look at the haplotypes for loci A and B, there are four possible haplotypes: A_1B_1, A_1B_2, A_2B_1 and A_2B_2. Assuming that the frequencies of alleles A_1, A_2, B_1 and B_2 in the population are all 0.5. If the two loci and the alleles are independent from each other, then the expected frequencies of each of the four haplotypes in the population would be 0.25. The LD level of the two loci is calculated by the deviation of the observed frequency of haplotypes from the expected value 0.25.

LD describes the situation that the combinations of genetic markers occur more or less frequently in a population than would be expected from random associations. This definition can be extended to the non-random association between markers and QTL. The underlying assumption of LD mapping of QTL is that a marker will affect the trait only if it is in LD with an unobserved QTL.

The difference between linkage mapping and LD mapping is the following. Linkage analysis describes the association between markers and QTL within families. The linkage can be broken down by recombination after only a few generations [23]. Therefore, the linkage between the markers and QTL is not sufficient to ensure that it will persist across the population. In contrast, LD mapping requires that a marker is in LD with a QTL across the whole population. To achieve this, the association must have persisted for a considerable number of generations. As a result, the marker should be closely linked to the QTL.

In conclusion, LD mapping of QTL exploits population LD levels between markers and QTL. The simplest method to perform LD mapping is to do a *genome wide association study* (GWAS) using single marker regressions.

2.2 Breeding Value Estimation

Genetic improvements through selective breeding have been used to improve the quality and profitability of livestock in many countries. Selective breeding is the process of selecting breeding plants and animals for particular phenotypic traits. Current selection programs in livestock are primarily based on selection of EBV derived from phenotype for traits of economic importance.

The EBV of an animal is an estimate of the animal's genetic breeding value for a particular trait. EBVs do not necessarily reflect animal's phenotypic performance, which is determined by both genetics and environment. In some cases, environmental factors can affect an animal's performance as much or more than the animal's genetic inheritance. The EBV is just the estimate of the genetic component of the phenotypic performance of an animal. As a result, it describes the heritability of the phenotypic trait, or that part of a trait that is expected to be inherited.

The task of genetic improvement of livestock is to select those animals that are genetically superior, not animals that have been raised under ideal management conditions. Therefore, we need to distinguish between genetic and environmental factors influencing phenotypic performance of a trait. As we know, performance that is the result of good management (*i.e.* environmental factors) will not be passed on to the next generation, whereas performance due to genetic superiority will be inherited by the progeny.

In order to explain the genetic variation of the phenotypic trait, two types of approaches have been proposed: the traditional quantitative genetics approach and the *marker assisted selection* (MAS) approach. There are basically three types of data available for breeding value estimation: pedigree, phenotypes and DNA markers. The quantitative genetics approach is based on the first two types of data, while MAS methods take advantage of DNA marker data in addition to the first two.

2.2.1 Quantitative Genetics Approaches

In the quantitative genetics approach, the genetic architecture of the trait of interest is treated as a "black box". EBV are derived from phenotype without knowing the specific role of any genes that affect the trait (*i.e.* without knowing where the genes that control the trait are located in the genome or what their individual effects are). The overall genetic variance of a quantitative trait is described by the infinitesimal model, which assumes that the trait is determined by an infinite number of unlinked and additive loci, and each with an infinitely small effect [18]. The genetic variances of individual loci are so small that they cannot be investigated separately, but the overall effects of all genes can be estimated via phenotypic resemblance between relatives. Therefore, the estimate is made by utilizing all the currently available information of the animal (*e.g.* the individual's performance, the parent's and sibling's performances and progeny performances). Naturally, the more information that is available, the greater the accuracy of the EBV, and the estimate will be less likely to change even when additional information becomes available.

Although the quantitative genetics approaches for selection have been effective for many traits, they have many limitations. The ideal situation for quantitative genetics selection is that the trait has high heritability, high quality of data recording and the phenotype can be observed in all individuals before reproductive age and with a relatively low cost. However, some traits, such as yield in plants, are of low heritability. Some traits, like milk production in dairy cattle, are only observable in one gender. And for some traits, such as carcass merit, slaughter quality, the selection decisions must be made before we can have the phenotype data.

Therefore, as we can see, many traits are not ideally suited to the quantitative genetics approach, which makes quantitative genetics selection difficult or costly for many traits. Moreover, if we can gain insights into the "black box" of genetics to elucidate the genetic architecture of the trait by using information of genetic markers, breeding progress could be greatly enhanced as we will be able to predict EBV more accurately and earlier in the life of animal (*e.g.* without examining its progeny).

2.2.2 MAS

MAS is the use of molecular genetic marker information in selection programs. MAS offers advantages over quantitative genetics approaches as marker information can be obtained for all animals at a young age and with low cost due to the advent of molecular genetic technology. In general, MAS is beneficial for traits with low heritability and traits with restrictions on phenotypic recordings, such as sex limited traits and traits recorded after selection [44].

When making selection decisions based on marker genotypes, a two-step procedure can be applied. Step 1 estimates the effects of markers in a reference population, while in step 2, the EBV of selection candidates are calculated using the information obtained in step 1.

Advances in genotyping technology and the discovery of thousands of SNPs in genome sequenc-

ing projects have provided new opportunities to find markers in linkage with QTL. Thus the use of molecular markers in genetics selection programs has become feasible. However, using the newly developed marker panels is especially challenging due to the high dimensionality of the datasets. More specifically, it is referred to as a large p (number of SNPs), small n (number of samples) problem, where the number of molecular markers is usually much larger than the number of samples.

2.2.3 Machine Learning

Machine learning involves the design and application of algorithms to automatically recognize complex patterns in large datasets. Machine learning has been applied in many fields, such as computer vision, natural language processing, medical diagnosis, bioinformatics, and speech and handwriting recognition.

When SNPs are used as molecular markers, the number of molecular markers is often much larger than the sample size (*i.e.* $n \ll p$). Traditional statistical approaches are severely challenged when dealing of this problem. Therefore, many machine learning techniques have been developed for large p small n datasets.

The prediction of EBV is essentially a regression problem, which is one of the most widely studied problems in machine learning. Given a set of known response variables , and a set of features (*i.e.* markers), the goal is to predict further responses for new values of the features. The problem becomes difficult when the sample size n is substantially smaller than the number of features p. Although many machine learning methods (*e.g.* SVM) can handle the high dimensionality, applying a variety of feature selection methods first to remove noise features and select informative features may still be useful for the prediction [16].

Another issue of potential importance is that the markers might act dependently. A framework for understanding interactions is necessary when analyzing genetic data; otherwise useful knowledge (*e.g.* gene-gene interactions) will go undetected. However, identifying interactions between marker loci is a challenging problem. Traditional statistical methods have difficulty modeling interactions because of combinatorial challenge. When the number of loci (*i.e.* the dimensionality of the dataset) increases, the number of combinations of interactions increases exponentially. Fortunately, machine learning offers many powerful models to identify interactions among features in high dimensionality datasets.

2.3 Genomic Selection

2.3.1 What is Genomic Selection?

Genomic selection is a form of MAS in which genetic markers, assumed to be in LD with QTL, covering the whole genome are used, so that potentially all the genetic variance is explained by the markers. Due to the large number of SNPs identified by genome sequencing in livestock species and

new technology for genotyping large numbers of SNPs for less than $1c$ USD per SNP, the genomic selection approach has become feasible for genetic improvement programs. The key feature of this method is that markers across the whole genome are used to divide the entire genome up into chromosomal segments (*e.g.* defined by adjacent markers) [45].

When performing genomic selection, the initial input is a SNP dataset assayed on a moderate number of animals with phenotypes. A regression model is then built from the input and used to predict the EBVs of testing samples, which have SNP genotypes but no phenotypic information.

2.3.2 Why use Genomic Selection?

Genomic selection has radically altered the structure of livestock breeding programs as the selection decision can be made based on genetic markers only. Optimal breeding programs designed with genomic selection can gain an increase in selection accuracy [23]. Furthermore, formal progeny testing will be unnecessary, which would potentially cut 92% of the cost for operating dairy breeding companies [55]. Moreover, genomic selection can predict EBVs for animals at a young age, which can reduce the generation interval by at least half [55].

2.3.3 Challenges for Genomic Selection

Implementation of Genomic selection proceeds in two steps:

1. Estimation of the effects of markers in a reference population for the quantitative trait.
2. Prediction of EBVs for selection candidates by summing across genome all the marker effects.

The above genomic selection approach can be used to map QTL as well as to predict EBVs. Note that, genomic selection can proceed using single markers, marker haplotypes, or using an IBD approach.

The factors governing the accuracy of estimates of QTL effects include the following:

- The number of samples in the reference population. The more phenotypic records available, the more observations there will be per haplotype, or per marker allele if single markers are used, and the higher the accuracy of genomic selection. This increased accuracy is because individual markers or haplotypes are likely to have small effects, so a large amount of data is needed to accurately estimate their effects.
- The proportion of the phenotypic variance explained by the DNA markers (*i.e.* heritability). The higher the heritability, the more accurate the prediction of phenotype based on marker information.

Chapter 3

Related Work

3.1 QTL Mapping

Understanding the complex genetic architecture of quantitative traits is a principal goal for biologists. If we can locate the genes affecting quantitative traits, it will lead to characterization and possible manipulation of these genes. Moreover, knowledge of these loci can aid in designing selection experiments to improve the traits.

Until recently, it was generally believed that quantitative traits are determined by a large number of QTL, each having a small effect on the phenotype. If this were true, then the molecular characterization of QTL would not be an attractive proposition, because the very large amount of effort required would probably not equal the value gained from characterizing each locus. However, analysis of QTL segregating from many experiments verified that only a small number of genetic loci contributed to a large proportion of the variance of each trait, whose effects can be obtained via segregation analysis [69].

Therefore, although the term QTL applies to any gene that has an effect on the quantitative trait, in practice, we seek only the major QTL, which have moderate to large effects, because only these have effects that are large enough to be detected and mapped on the genome.

The use of genetic markers to locate QTL is well established. There are a large number of different QTL mapping methods for identifying QTL by exploiting LD between markers and QTL [5]. Our concentration here is almost exclusively on detecting QTL, while the estimation of the QTL effects and precise locations are of secondary importance. There are basically two kinds of methods for QTL mapping, the ones that model a single QTL at a time and those that attempt to model several or all QTL at once.

3.1.1 Single-QTL Models

Considerable attention has been paid to the case of association between a single marker and a quantitative trait. The simplest one of all QTL mapping methods is the genome wide association test between the trait value and the genotypes of marker loci. In this method, each marker locus is

considered separately to test for association with the trait. If a marker is associated with the trait, it suggests that the trait is affected by a QTL close to the marker. To estimate the effects of marker loci, two general methods have been used: *least square regression* (LS) and *maximum-likelihood* (ML) estimation using the *estimation-maximization* (EM) method [22, 28, 9]. After obtaining the marker effects, for each marker locus, a number of statistical tests can be used to determine the significance of the association with the trait, *e.g.* student's t statistic, the LOD score, the *likelihood ratio statistic* (LRS) and a nonparametric statistic is also available [22, 37, 39, 35].

The second level of QTL mapping is *simple interval mapping* (SIM) and various modified versions of SIM. Lander and Botstein presented a likelihood-based framework for interval mapping [37]. The basic idea of these methods is to construct a marker genetic map first by dividing the entire genome into a finite number of intervals 1 or 2 cM apart, which contain putative QTL. The genotypes of the intervals are not observable but can be inferred from the genotypes of flanking markers. Two flanking markers define an interval that may contain several putative QTL. Interval mapping is a single-QTL model in the sense that it tests one interval at a time. Only effects of the putative QTL at the current position are included in the model and all other QTL effects are ignored. Interval mapping has several advantages over analysis at single marker locus. First of all and perhaps most important, interval mapping allows incomplete marker genotype data. If an individual is missing the marker genotype for a flanking marker, one can move to the next flanking marker whose genotype data is available. Secondly, interval mapping can improve estimates of QTL effects. If using single marker locus analysis, the effect at a marker locus might be attenuated as a result of recombination between the marker and the QTL [5].

The disadvantage of SIM is that it does not take account of all markers at once, which causes the problem called ghosting effects. The ghosting effects problem refers to the situation where if there is a QTL in one interval, adjacent intervals may also show peaks with "significant" likelihood ratios. "If a QTL is actually present in one interval, the hypothesis of a QTL in an adjacent interval will still fit the data better than the hypothesis of no QTL at all" [9].

Motivated by the dependency between markers, a variant of the interval mapping method, composite interval mapping, was proposed by Jansen [27]. Composite interval mapping treats QTL effects ignored by SIM as background effects, which are absorbed by the other selective markers (called the co-factors) outside the tested interval. Composite interval mapping can substantially improve the efficiency of QTL mapping as it reduces the confounding effects from nearby QTL if the background markers and the target interval are linked.

However, to maximize parameter estimation efficiency and statistical power and to estimate epistasis (*i.e.* interaction effects between QTL), multiple QTL must be mapped simultaneously. Moreover, because all QTL effects are estimated simultaneously, the problem of over-estimation of QTL effects due to single QTL analysis can be overcome [23].

3.1.2 Multiple-QTL Models

It is better to study all QTL jointly because they may be correlated due to physical linkage of their gene products to act interactively in determining the phenotype of the trait. For example, a disease might occur only if a particular combination of genotypes is present at different susceptibility loci, instead of just the result of a single disease gene alone. Therefore, multiple QTL mapping has become the state-of-the-art QTL mapping method. However, these methods require orders of magnitude more computation as the number of genetic markers is far larger than sample size, which is called the over-saturated model. Special techniques are required to deal with the over-saturated model if we want to study all of the QTL jointly. In general, there are two ways to handle such over-saturated model: variable selection and shrinkage estimation.

Variable selection is an important technique for dimension reduction. Several heuristic search approaches have been taken for selecting the optimal set of putative QTL. Kao *et al.* adopted a stepwise regression approach to add and delete QTL progressively until the model is stabilized [30]. QTL are added to the model if they significantly improve the fit of the existing model. It seems, however, quite arbitrary to set the effects of loci to zero that are below the significance threshold and include the full effects of those that are above this threshold. Furthermore, the selection of loci with the largest effects would probably result in the inclusion of over-predicted effects in the model.

Bayesian variable selection is an alternative approach. It assumes a prior for all loci and priori distributions for the unknowns in the model. Inference is then based on the conditional distribution of the unknowns given the observed data, the posterior distribution. A Bayesian method implemented via the *Markov chain Monte Carlo* (MCMC) algorithm has been developed for mapping multiple QTL [25]. The number of QTL is determined either by the Bayes factor or by reversible-jump MCMC. It has been noted that the reversible-jump MCMC for model selection usually has the problem of slow convergence.

Instead of deleting all non-significant variables from the model, in shrinkage analysis, all variables are included but their estimated effects are shrunk toward zero. Ridge regression is a typical example of shrinkage estimation. Recently, Xu showed that the usual ridge regression method can fail if the number of model effects is too large [68]. Ridge regression actually has a Bayesian analogy. The small positive number added to the diagonal elements of the coefficient matrix in ridge regression is equivalent to the ratio of the residual variance to the variance parameter of the QTL effects. Xu then modified the variance parameter of the prior distribution of the QTL effects and let the variance parameter vary across QTL [69]. As a result, the method can handle models with the number of effects many times larger than the number of samples.

3.2 EBV Prediction

3.2.1 Genomic Selection

A number of approaches have been proposed for breeding value estimation. The key difference between these approaches are the underlying genetic models and the assumptions they make about the variances of haplotype or single marker effects across chromosomal segments.

Least Square Regression (LR)

As described by Meuwissen *et al.*, LR genomic selection treats marker effects as fixed in the regression model and uses single marker regression analysis to test the statistical significance of each marker separately [45]. Next the multiple regression model is used to select the most significant markers, which pass the significance threshold, and to estimate their effects simultaneously, while all of the other marker effects are assumed to be zero. Due to the degrees of freedom shortage in LR, when the number of markers is larger than the number of phenotypic records, a regular LR approach would fail. Therefore, by imposing a significance threshold, the dimension of the model can be reduced significantly.

One of the problems that arises is the selection of the significance threshold. Another problem is the over-estimation of marker effects due to the single marker regression analysis for the selection of marker effects to be included in the final regression model.

Ridge Regression and Best Linear Unbiased Prediction (BLUP)

Whittaker *et al.* proposed a ridge regression MAS method in an attempt to avoid the problem of over-estimation [66]. In ridge regression, the marker effects are not treated as fixed effects, instead, the marker effects are estimated by shrinking toward the population mean. All marker effects can be estimated simultaneously, which overcomes the problem of significance testing.

In ridge regression, the effects of markers are estimated by:

$$\hat{g} = (X^T X + \lambda I)^{-1} X^T y, \tag{3.1}$$

where X is a matrix allocating all marker genotypes or haplotypes to phenotypes, and y is a vector of phenotypes. For BLUP as used by Meuwissen *et al.*, λ is equal to σ_e^2/σ_g^2 in the equation for ridge regression, where σ_e^2 is the error variance and σ_g^2 is the variance of the effects across all segments [45].

Both the ridge regression method and BLUP assume the variances of all marker effects to be the same. Actually, the ridge regression method uses the infinitesimal model by assuming that traits are determined by an infinite number of markers, each with an infinitesimally small effect scattered along the chromosomes. This model has been exceptionally valuable for animal breeding, and forms the basis for breeding value estimation theory. However, this assumption does not capture the "prior" knowledge that many markers have negligible effects while only a few have significant

effects. Therefore, the marker with the largest variance will tend to have its effect over-estimated, which will decrease the accuracy of EBV prediction. Better estimates of breeding value can be obtained by methods that allow varied variance for different markers.

Bayesian Methods

The Bayesian method assumes varied variance of effects for different markers, and the variances are estimated by using a prior distribution.

In probability theory, Bayes' theorem relates the conditional probabilities of two random events by a simple rule

$$P(x|y) = \frac{P(xy)}{P(y)} = \frac{P(y|x)P(x)}{P(y)}, \tag{3.2}$$

where

- $P(x)$ is the prior probability of x. It is "prior" in the sense that it does not take into account any information about y, and allows us to incorporate prior knowledge into the estimate of x.
- $P(x|y)$ is the conditional probability of x, given y. It is also called the posterior probability, because it is derived from or depends upon the specified value of y.
- $P(y|x)$ is the conditional probability of y given x.
- $P(y)$ is the prior probability of y, and always acts as a constant.

A number of approaches have been proposed to calculate the posterior distribution. To make things easy, it assumes that the data are normally distributed, and the prior and posterior distributions are also assumed to be normally distributed.

Meuwissen *et al.* described a Bayesian method to estimate both the chromosome segments effects and their variances simultaneously. The experimental results of Meuwissen *et al.* showed that the prior distribution of the variances of effects across chromosome segments was an inverted χ-square distribution. It is consistent with what would be expected from the theorem that many QTL have small effects while a few have large effects [45]. An advantage of using an inverted χ-square distribution as a prior for the variances is that with normally distributed data, the posterior is also inverted χ-square. Meuwissen *et al.*'s Bayesian method uses Gibbs sampling to estimate the variances of effects as we cannot estimate them directly because they are conditional on the unknown marker effects [45].

Xu also described a Bayesian method for single SNP markers with a similar prior for the variance of chromosome segment effects, *i.e.* an inverted χ-square distribution [68]. The implicit assumption of Xu is that the effect of marker i on the trait will absorb partly the effects of all QTL located between markers $i-1$ and $i+1$ [68]. The validity of this assumption will depend on the LD between the markers and the QTL. Actually, Xu proposed this method for QTL mapping instead of EBV prediction. As a QTL was restricted in adjacent marker brackets, the QTL was mapped

to a small interval. Therefore, this method gives a more precise estimate of QTL locations. Many Bayesian methods that have been developed for QTL mapping can also be used for EBV prediction, such as the Bayesian shrinkage method and empirical Bayes method.

3.2.2 Factors Affecting the Accuracy of EBV Prediction

The accuracy of EBV prediction depends on the number of markers, the number of phenotypic records per haplotype or marker allele if single markers are used, the heritability of the trait and the ratio of non-additive effects to additive effects for the trait.

The number of markers required is determined by the level of LD between adjacent markers. Meuwissen *et al.* stated that the level of LD should be > 0.2 for genomic selection to be successful [45]. Calus *et al.* used simulation to assess the effect of the level of LD between adjacent marker pairs on the accuracy of EBV prediction. They found that the accuracy increased dramatically as the level of LD between adjacent markers increased. In a dairy cattle population, in order to achieve an LD level of 0.2 between adjacent markers, the markers should spaced at most 100kb apart [8].

Meuwissen *et al.* compared the accuracy of EBV prediction with different numbers of phenotypic records. Their results suggest that 2000 phenotypic records are required to accurately estimate the haplotype or marker effects [45]. It is easy to see that the more phenotypic records are available, the more observations there will be per haplotype or marker, and the higher the accuracy of estimation of marker effects.

For heritability, higher heritability means that the phenotype is more dependent on the genotype. Therefore, the more accurate the prediction will be.

Chapter 4

Datasets and Methods

This thesis encompasses two main objectives:

- QTL detection, which can be regarded as an association study between QTL and the phenotypic trait. Our focus here is exclusively on detecting QTL, considering the estimation of the QTL effects and precise locations of secondary importance. As we are using a SNP dataset in our study, this objective can also be regarded as the detection of a subset of SNPs which are highly correlated with the trait.
- EBV prediction. The accuracy of the predicted EBV values is important for the selection of the best breeding or performing animals.

The focus of this work is mainly on the evaluation of an EB method, in terms of its ability to select a subset of SNP markers for accurate breeding value prediction. The evaluation was performed using several simulated and real datasets consisting of genotypes and phenotypes.

4.1 Datasets

4.1.1 Dairy Dataset

A total of 462 Canadian Holstein dairy bulls from Semex Canada (Guelph, ON) were used in the study. 319 of them originating from 10 core sire families, and the rest (143) from the general pedigree. A set of 1536 SNP markers was selected to strategically represent potential candidate genes across the bovine genome sequence assembly. After removing 139 poorly amplified and 56 monomorphic SNPs, 1341 SNPs were used for further analysis [34].

Five production traits: milk yield (MY), protein yield (PY), fat yield (FY), protein percentage (PP) and fat percentage (FP) are used in this study for prediction. The systematical environment effects of each phenotype trait have already been excluded by BLUP.

4.1.2 Beef Dataset

This dataset includes genotype, phenotype, and pedigree information for cattle located at the University of Alberta's Kinsella Ranch, Alberta. The dataset consists of a set of animals for which various traits were measured, including feed efficiency. The genotypes of these animals, and the genotypes of their sires were also determined.

A $58K$ genotyping panel developed by Illumina was used for genotyping [41]. Positions of SNPs on the bovine genome assembly version Bta V 4.0 were determined by BLAST and sequence alignment. There are 469 genotyped animals and a total of 51828 markers in the dataset. 47108 out of the 51828 total markers are polymorphic markers and 446 out of the 469 genotyped animals have both genotype and phenotype values.

We did some pre-processing for this dataset as the number of SNPs far exceeds the number of samples. Before applying EBV prediction algorithms to this dataset, we first removed some SNPs from the dataset, which we think are redundant for prediction purposes. We selected SNPs randomly from each chromosome but ensured that the proportion of SNPs on each chromosome is consistent with that in the original dataset. After the pre-processing, only 5000 SNPs remained in the dataset.

In summary, the dataset used in our study contains 446 samples and 5000 SNPs. Three production traits: ADG, birth weight (BW) and RFI are used in this study for prediction. The EBVs of these traits are derived from phenotype and pedigrees by BLUP.

4.1.3 Simulation Datasets

In the simulation, we assume that the QTL act additively because identifying interactions between loci is a much more difficult problem. Considering only the simple additive case will lead to greater clarity.

Suppose y_i is the phenotypic value of the i^{th} animal, $i = 1, 2, \ldots, n$. The simulation model for y_i is

$$y_i = \sum_{j=1}^{p} X_{ij} b_j + e_i \tag{4.1}$$

where

- p is the number of QTL included in the model.
- X_{ij} is the genotype information.
- b_j is the j^{th} QTL effect.
- e_i is the background error $\sim N(0, \beta \cdot std(QTL_i))$:
  - β is the level of background noise
  - $std(QTL_i)$ is the standard deviation of all the QTL effects of the i^{th} animal.

We also consider several dominance levels (α) to simulate. Suppose the alleles carried at a marker locus is A and B, and A is the QTL allele. Assuming that an animal is heterozygous at that marker locus, and then the effects caused by the heterozygous marker can be calculated in this way:

$$\text{Effect}_{\text{heterozygous}} = \alpha \cdot \text{Effects}_{\text{AA}} \tag{4.2}$$

- completely recessive: $\alpha = 0$.
- partially recessive: $\alpha = 0.25$.
- co-dominant: $\alpha = 0.5$.
- partially dominant: $\alpha = 0.75$.
- completely dominant: $\alpha = 1$.

We simulated 100 datasets for each of the following 10 types:

- $\alpha = 0, \beta = 0$.
- $\alpha = 0.25, \beta = 0$.
- $\alpha = 0.5, \beta = 0; \beta = 0.1; \beta = 0.2; \beta = 0.3; \beta = 0.4; \beta = 0.5$.
- $\alpha = 0.75, \beta = 0$.
- $\alpha = 1, \beta = 0$.

Therefore, a total of 1000 datasets were simulated, with each dataset containing 353 samples and 1341 SNPs on a chromosome. The reason that we only simulate different background noise for the co-dominant model is that it is believed that the real data usually follows the co-dominant model with background noise.

4.1.4 Data Pre-processing

As mentioned earlier, the big challenge of SNP datasets is that the number of features (*i.e.* SNPs) is far larger than the number of samples (*i.e.* animals). Therefore, our fist step in data pre-processing is to remove some features that are not informative (*i.e.* the same genotype is present in all or almost all animals). If a feature has the same values for almost all animals, we remove this feature from the dataset.

SNP encoding

The genotype of each SNP (feature) can only take three values (*e.g.* AA, AT, TT): 2 homozygous alleles (*i.e.* AA, TT) and 1 heterozygosis allele (*i.e.* AT). For each SNP, we define one of the homozygous allele as homozygous majority if its occurrence is more frequent and the remaining one is defined as homozygous minority.

- Co-dominance representation :
  - homozygous majority $\rightarrow 1$
  - heterozygosis $\rightarrow 0$
  - homozygous minority $\rightarrow -1$
- Binary representation :
  - homozygous majority $\rightarrow 001$
  - heterozygosis $\rightarrow 010$
  - homozygous minority $\rightarrow 100$

The binary representation seems to make more sense as it measures the presence and absence of the effect of each allele. SNP encoding is also one of the tasks that we want to evaluate in our experiments.

4.2 Methods

4.2.1 Feature Selection Methods

Because of the high dimensionality and small sample size of each SNP dataset, it is very necessary to select a subset of SNPs that are highly correlated with the phenotypic trait first. Feature selection is the commonly used technique in machine learning to select a subset of relevant features to build robust learning models.

Feature selection can help to improve the learning performance of many machine learning algorithms by removing irrelevant and redundant features from the high dimensional data.

The feature selection method also serves as a QTL detection method, *i.e.* we may consider the selected subset of SNPs as related to QTL. Therefore, we can locate the QTL for the phenotypic trait and gain a better understanding about the genetic structure of the trait. Moreover, if the number of markers required to apply genomic selection can be reduced, this could represent a large savings to the breeding program for marker genotyping.

Correlation-based Feature Selection

The correlation-based feature selection method evaluates features individually by measuring their correlation with the phenotypic trait. The aim is to select a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. In other words, subsets of features that are highly correlated with the phenotypic trait while having low inter-correlation are preferred.

The correlation-based feature selection method iteratively adds features with the highest correlation with the phenotypic trait as long as there is not already a feature in the selected subset that has a higher correlation with the feature currently in question.

M5

The basic idea behind the M5 feature selection is actually model selection using *Akaike Information Criterion* (AIC). The AIC is given by the following formula:

$$AIC = 2(p+1) + n[\ln(\frac{\ln(y_i - f(x_i))^2}{n}) + 1] \tag{4.3}$$

where k is the number of parameters in the statistical model, and $i = 1, ..., n$ is the observations, n is the number of observations.

Lower AIC values indicate a better model. The AIC methodology attempts to find the model that best explains the data with a minimum of free parameters. The M5 method adopts the greedy backward elimination approach for feature selection by stepping through the parameters and removing the one with the smallest standardized coefficient until no improvement is observed in the score given by the AIC.

Empirical Bayes (EB) [69]

The EB method is a selective shrinkage method for the "large p small n" model [69]. Selective shrinkage means different regression coefficients are assigned different prior variances. A smaller prior variance will cause the regression coefficient to shrink more, while a larger prior variance will lead to less shrinkage. As mentioned before, most quantitative traits are actually controlled by a large number of markers with small effects and a few markers with large effects. As a result, allowing the ridge factor to vary across different markers instead of allowing all model effects to be shrunk by the same factor would be better.

Let y_i for $i = 1, \ldots, n$ be the phenotypic value for the i^{th} animal in the population dataset. The linear model for y_i is

$$y_i = X\alpha + \sum_{j=1}^{p} z_{ij}\gamma_j + e_i \tag{4.4}$$

where X and α are defined as the environmental variants, which are not genetically interesting but should be included in the model to reduce the residual error. X is identity matrix. p is the

total number of markers (SNPs) in the entire genome, z_{ij} is the genotype of the j^{th} marker for individual i, γ_j is the QTL effect associated with marker j, and e_i is the residual error with a $N(0, \sigma^2)$ distribution.

This model is called an over-saturated regression model because $q + p \gg n$. We can always assume that $q \ll n$ for genomic selection datasets, so we can estimate α easily by ordinary least-squares method. However, p can be very large and most elements of γ will be zero or close to zero. When the number of markers exceeds the number of animals, the ordinary least-squares approach will have no unique solution. Therefore, Xu proposed a Bayesian regression method to handle the problem of over saturation [69].

In the Bayesian framework, every regression parameters of γ_j is considered as a random variable with a prior distribution. The purpose of Bayesian analysis is to infer the posterior distribution of the parameters given the observed data. The distribution of the observable is a likelihood function of the unknown parameters.

A normal prior distribution $N(0, \sigma_j^2)$ is assigned to γ with $\sigma_j^2 \sim Inv - \chi^2(\tau, \omega)$ as the prior for σ_j^2, where τ is the degree of freedom (prior belief) and $\omega > 0$ is the scale parameter. The actual form of the inverse χ-square distribution is

$$p(\sigma_j^2) \propto (\sigma_j^2)^{-(\tau+2)/2} \exp(-\frac{1}{2}\omega/\sigma_j^2). \tag{4.5}$$

The likelihood is proportional to

$$p(y|\theta) = \phi_n(y; X\alpha + Z\gamma, \sigma^2), \tag{4.6}$$

where $\theta = (\alpha, \sigma^2, \gamma)$, and $\phi_n(x; a, b)$ represents the n-dimensional multivariate normal density for vector x with mean a and covariance matrix b. Therefore, the prior density is proportional to

$$p(\theta|D) = \phi_p(\gamma; 0, D), \tag{4.7}$$

where $D = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$ and with a distribution of

$$p(D) = \varphi_p(D; \tau, \omega) \propto |D|^{-(\tau+2)/2} \exp -\frac{\omega}{2} tr(D^{-1}). \tag{4.8}$$

The posterior distribution of θ, $p(\theta|y)$, is proportional to $p(\theta, y, D)$

$$p(\theta, y, D) = p(y|\theta)p(\theta|D)p(D). \tag{4.9}$$

From above, we can then get the joint likelihood function for $\psi = (\alpha, \sigma^2, D)$ as proportional to

$$p(y|\psi) = \int \ldots \int p(\theta|y)d\gamma = \phi_n(y; X\alpha, V)\varphi_p(D; \tau, \omega), \tag{4.10}$$

where ψ is the vector of parameters in the model, and $V = \sum_{j=1}^{p} Z_j Z_j^T \sigma_j^2 + I\sigma^2$. Finally, the log likelihood function for parameter ψ is

$$L(\psi) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\alpha)^T V^{-1}(y - X\alpha) - \frac{1}{2}(\tau + 2)\log|D| - \frac{\omega}{2} tr(D^{-1}). \tag{4.11}$$

The maximum likelihood estimation of ψ can be estimated by using the following algorithm iteratively.

Start with some initialized values of the parameters $\psi^{(t)}$ for $t = 0$. We can obtain $\psi^{(t+1)}$ by maximizing the log likelihood function given $\psi^{(t)}$. Denote ψ_k as the k^{th} element of ψ, then the log likelihood function for ψ_k given $\psi^{(t)}$ is denoted by $L(\psi_k|\psi^{(t)})$. Then, the maximum likelihood estimation of ψ_k conditional on $\psi^{(t)}$ can be obtained by setting $\frac{\partial}{\partial\psi_k}L(\psi_k|\psi^{(t)}) = 0$ and solving for ψ_k. Here are the steps of the algorithm [69]:

- Step 1: Updating the environment effects using

$$\alpha^{(t+1)} = [X^T(V^{(t)})^{-1}X]^{-1}X^T(V^{(t)})^{-1}y; \tag{4.12}$$

- Step 2: Updating the residual variance using

$$\sigma^{2(t+1)} = \frac{\sigma^{2(t)}}{n}(y - X\alpha^{(t)})^T(V^{(t)})^{-1}(y - X\alpha^{(t)}); \tag{4.13}$$

- Step 3: Updating $\sigma_j^2(j = 1, \ldots, p)$ by maximizing

$$\begin{aligned} L(\sigma_j^2|\psi^{(t)}) = &-\frac{1}{2}\log(Z_j^T(V^{(t)})^{-1}Z_j(\sigma_j^2 - \sigma_j^{2(t)}) + 1) \\ &+ \frac{1}{2}\frac{(\sigma_j^2 - \sigma_j^{2(t)})[(y - X\alpha^{(t)})^T(V^{(t)})^{-1}Z_j]^2}{Z_j^T(V^{(t)})^{-1}Z_j)(\sigma_j^2 - \sigma_j^{2(t)}) + 1} \\ &- \frac{1}{2}(\tau + 2)\log\sigma_j^2 - \frac{\omega}{2\sigma_j^2}; \end{aligned} \tag{4.14}$$

- Step 4: Repeat Steps 1-3 until a certain criterion of convergence is satisfied.

The striking feature of the EB method is that most of the markers have an estimated effect close to zero. Therefore, the signal of QTL associated markers are so clear that we can detect these markers without effort.

Bagging EB

The EB method did capture the most significant features from the dataset, however, the problem is that it also ignores a lot of other informative features, which would definitely affect the accuracy of EBV prediction. Therefore, we proposed the bagging EB method to overcome this problem. Here is the procedure of the bagging strategy:

Given a dataset, we generated 999 new datasets having the identical genotype data but re-scaled trait values. The scale ratios are random numbers selected out of range $[0.001, 1000]$. On these 1000 datasets, the EB algorithm was run and the returned SNPs of significant effects were collected.

4.2.2 Regression Methods

We applied seven regression methods which are commonly used in the field of machine learning to deal with high dimensional data. These methods are SVM, GP, PCA, PLS, Ridge, LASSO and ElasticNet.

Support Vector Machine

The SVM was developed by Vapnik to solve the classification problem [62]. But gradually, SVMs have been successfully extended to regression problems. SVMs are gaining popularity due to many attractive features and promising empirical performance. They have been used for isolated handwritten digit recognition, object recognition, speaker identification, face detection in images, and text categorization [10, 2, 49, 29]. For the regression estimation case, SVMs have been compared on benchmark time series prediction tests, the Boston housing problem [49, 15, 48, 47]. In most of these cases, SVM generalization performance either matches or is significantly better than that of other competing methods.

The regression problem can be stated as: given a training data set $D = \{(x_i, y_i)|i = 1, 2, ..., n\}$, of input vectors x and associated targets y, the goal is to get a function $g(x)$ to infer the output y for a new input data vector x. Any practical regression algorithm has a loss function $L(y, g(x))$, which describes how the estimated values deviate from the true values. Many forms for the loss function can be found in the literature: *e.g.* linear, quadratic loss function, exponential. Here, Vapnik's loss function is used, which is known as the ε - insensitive loss function (also called the soft margin loss function) and defined as [62]:

$$L(y, g(x)) = \begin{cases} 0 & \text{if} |\text{y} - \text{g(x)}| \leq \varepsilon \\ |y - g(x)| - \varepsilon & \text{otherwise} \end{cases} \tag{4.15}$$

where $\varepsilon > 0$ is a predefined constant which controls the noise tolerance. With the ε - insensitive loss function, the goal is to find $g(x)$ that has at most ε deviation from the actually obtained targets y_i for all training data, and at the same time is as flat as possible. In other words, the regression algorithm does not care about errors as long as they are less than ε, but will not accept any deviation larger than this.

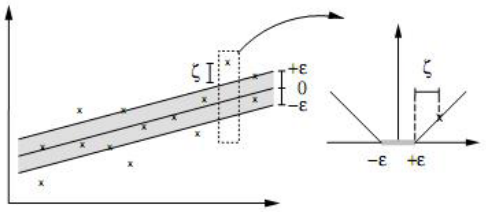

Figure 4.1: The Soft Margin Loss Function for linear SVM [15].

For the simplest case, function g can take the following linear form: $g(x) = w.x + b$. Thus, the regression problem can be written as a convex optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 \\ \text{subject to} & |y_i - (w \cdot y + b)| \leq \varepsilon. \end{array} \tag{4.16}$$

As shown in Figure 4.1, only the points outside the shaded region contribute to the cost, as the deviations are penalized in a linear fashion. For complicated problems, non-linear kernel functions

have been used to replace the dot product, particularly to reduce the effect of outliers on the classifier. Some common used kernel functions are listed below:

- Polynomial (homogeneous): $k(x_i, x_j) = (x, x')^p, p \in N$
- Polynomial (inhomogeneous): $k(x_i, x_j) = (x_i \cdot x_j + c)^p, p \in N, c \geq 0$
- Gaussian Radial Basis Function: $k(x_i, x_j) = \exp^{-\frac{\|x_i - x_j|^2}{2\sigma^2}}$
- Radial Basis Function: $k(x_i, x_j) = \exp^{-\gamma\|x_i - x_j\|^2}$, for $\gamma > 0$

GP

In the mathematical theory of probability, a GP is a stochastic process for which any finite combination of random variables have joint Gaussian distributions (or more generally, any function applied to the random variables will give a normally distributed result) [67]. From the above definition, we can say that GP specifies a distribution over functions. Inference of continuous values with a Gaussian process prior is known as Gaussian process regression.

GP is completely specified by a mean function $\mu(x)$ and positive definite covariance function $k(x, x')$. In most cases, we can assume the random variables have mean zero. For any finite set of variables, it will have a joint multivariate Gaussian distribution. Therefore, given a set of noise free training data $x_i|i = 1, ..., n$, we may draw samples $f(x_1), \ldots, f(x_n)$ from the GP prior with $f(x_1), \ldots, f(x_n) \sim N(0, k)$, where k is the covariance function $k(x, x')$. We want to maker predictions f^* for testing data x^*. The joint distribution of f and f^* is as follows:

$$\begin{bmatrix} f \\ f^* \end{bmatrix} \sim N(\begin{bmatrix} \mu \\ \mu^* \end{bmatrix}, \begin{bmatrix} k(x,x)k(x,x^*) \\ k(x^*,x)k(x^*,x^*) \end{bmatrix}), \tag{4.17}$$

where μ is the training mean and μ^* is the testing mean, $k(x, x)$ is the training set covariances, $k(x, x^*)$ is the training-testing set covariances and $k(x^*, x^*)$ is the testing set covariances. Therefore, the conditional posterior distribution f^* given $D = \{X, f\}$ is $P(f^*|x^*x, f)$ with Gaussian distribution $N(\mu, \sum)$, where

$$\begin{aligned} &\mu = k(x, x^*)k(x, x)^{-1}f, \text{and} \\ &\textstyle\sum = k(x^*, x^*) - k(x, x^*)k(x, x)^{-1}k(x^*, x). \end{aligned} \tag{4.18}$$

From the posterior distribution, we can sample function values for the testing data. Then we can either use the mean values as our predictions, or express our uncertainty of the predictions by confidence intervals. There are many choices of reasonable prior mean and covariance functions. Therefore, one important issue is to choose the best prior mean and covariance functions for the specified dataset.

PCA

PCA is a mathematical procedure that reduces the dimensionality of the data while retaining most of the variation in the dataset. It accomplishes this reduction by transforming a number of possibly

correlated variables into a smaller number of uncorrelated variables called principal components, along which the variation in the data is maximal [56]. More generally, PCA projects the data along the directions where the data varies the most. By using a few components, each sample can be represented by relatively few features – the principal components, which are orthogonal linear transforms of the original variables.

The first principal component accounts for the largest percentage of the variation and the second principal component explains the second largest percentage and so on. Typically the first few principal components account for most of the variation while the remaining principal components make a negligible contribution. PCA is theoretically the optimal orthogonal linear transformation for given data in least square terms at the price of greater computational requirement compared with other dimensionality reduction techniques.

Using PCA to do regression is to predict continuous responses based on estimated regression coefficients of constructed principal components instead of original variables. Typically, only a subset of the principal components are used in regression. Although it is usual to select the principal components with the highest variances, the low variance principal components may also be important. In practice, we often use cross-validation to select a subset of principal components.

PLS

PLS is a statistical method that is similar to PCA as both PLS and PCA produce factor score variables used in the predictive regression model as linear combinations of the original predictor variables. PLS and PCA differ in the methods used in extracting the factor scores. In short, PCA reflects the covariance structure between the predictor variables, while PLS examines the covariance structure between the responses and corresponding variables [19]. PLS regression finds components from variables that are also relevant for the responses.

PLS is an extension of the multiple linear regression model. Note that the emphasis is on predicting the responses but not necessarily on trying to understand the underlying relationship between the variables. PLS is a method for constructing predictive models when the factors are many and highly co-linear.

LR

Consider the usual linear regression model: Given p predictors $x_1, \ldots, x_p$, predict the response variable Y by the linear model

$$y = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p. \tag{4.19}$$

Given a dataset, a model fitting procedure gives the vector of coefficients $\beta = (\beta_0, \ldots, \beta_p) = (X'X)^{-1}X'y$ by minimizing the residual sum square error. However, ordinary least square regression may result in large variable estimates of the regression coefficients when the number of

parameters to be estimated (p) is large compared to the number of observations (n). From a mathematical standpoint, in this situation, the $X'X$ matrix is ill-conditioned: the value of its determinant is nearly 0, and attempts to calculate the inverse of the matrix run into numerical snags with uncertain final values, and so $var(\beta)$ will have very large elements. For a statistical model with too many degrees of freedom, the learner may adjust to very specific random features of the training data, which have no causal relation to the target function. Therefore, although the performance on the training examples still increases, the performance on testing data becomes worse. This is the problem called overfitting.

Several regularized regression methods have been developed in the last few decades to overcome the overfitting problem of ordinary least squares regression, such as Ridge, LASSO, and more recently LARS and ElasticNet [60, 17, 75].

Ridge

Ridge is the most commonly used method of regularization and works by adding a quadratic penalty term for large parameter estimates to the residual sum square error. Ridge reduces this variability by shrinking the coefficients, resulting in better prediction accuracy at the cost of usually only a small increase of bias. In Ridge, the sum of squares of the coefficients is constrained as follows:

$$L^{ridge}(\beta_1, \ldots, \beta_p) = ||y - \sum_{j=1}^{p} \beta_j x_j||^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2, \tag{4.20}$$

where p is the number of predictor variables, $\beta_j, j = 1, .., p$ are the regression coefficients, $|| \cdot ||$ denotes the squared Euclidean norm, and λ_1 is the ridge parameter, which determines how much ridge regression departs from least square regression. The optimal value for the ridge parameter is estimated by a series of trial and errors, and involves cross validation.

LASSO

LASSO was developed by Tibshirani to improve both prediction accuracy and model interpretability [60]. The LASSO method reduces the variability of the estimates by shrinking the coefficients and at the same time produces interpretable models by shrinking some coefficients to exactly zero. In [60], it was demonstrated that in terms of prediction accuracy and interpretability, the LASSO method outperforms ridge regression for data with a small to moderate number of moderate-sized effects, but ridge regression performs better with a large number of small effects.

LASSO minimizes residual sum squared error subject to a bound on the L_1 norm of the coefficients. LASSO constrains the sum of the absolute values of the coefficients using the following formula:

$$L^{lasso}(\beta_1, \ldots, \beta_p) = ||y - \sum_{j=1}^{p} \beta_j x_j||^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|, \tag{4.21}$$

where λ_2 is the LASSO parameter again estimated by a series of trial and errors and cross validation.

The limitation of LASSO is that if $p > n$, the LASSO method selects at most n variables, *i.e.* the number of selected coefficients is bounded by the number of observations.

ElasticNet

ElasticNet is a least squares method with both an L_1 penalty and a quadratic penalty. The L_1 penalty is a LASSO type threshold that performs variable selection thus inducing a sparse model. The quadratic penalty, related to Ridge, encourages a grouping effect and places no limitation on the number of variables that may be selected for the model. Therefore, the ElasticNet combines the Ridge and the LASSO constraints:

$$L^{elasticnet}(\beta_1, \ldots, \beta_p) = ||y - \sum_{j=1}^{p} \beta_j x_j||^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|, \tag{4.22}$$

where λ_1 is the Ridge penalty parameter, penalizing the sum of the squared regression coefficients and λ_2 is the LASSO penalty, penalizing the sum of the absolute values of the regression coefficients.

BLUP

BLUP assumes the infinitesimal model that traits are determined by a large number of genes each with small effects scattered along the chromosomes. All marker effects can be estimated simultaneously through fitting the marker effects as random effects which does not require degrees of freedom without the problem of significant testing and overestimation. Random effects require however an estimate of the variance of the marker effects. BLUP assumes that the marker effects are drawn from normal distribution with constant variance across chromosome segments [45, 32]. The marker effects can be calculated through:

$$y = 1_n \mu + \sum_i X_i g_i + e \tag{4.23}$$

where X is the design matrix allocating all marker genotypes to phenotypes, y is a vector of phenotypes; g_i is the genetic effects of the marker at the i^th segment.

The estimates of g_i are obtained from the Henderson Mixed model equations:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \mu \\ g \end{bmatrix} = \begin{bmatrix} X'Y \\ Z'Y \end{bmatrix} \tag{4.24}$$

where $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$, σ_g^2 is the variance of marker effects at a chromosome segment. The problem with BLUP is the same variance assumption made for every marker. The majority of the markers which have very little effect on the trait and being estimated to be small but non-zero and their cumulative effect adds noise to the estimates and dominate the variance of the marker effects.

Chapter 5

Experiment Results and Discussion

5.1 QTL Mapping

There are 20 pre-defined SNPs in each simulation dataset that are related to QTL. The purpose of this set of experiments is to detect those SNPs using the bagging EB method. We use two measurements to guage the performance on each dataset:

- precision = number of correctly detected SNPs / number of all detected SNPs
- recall = number of correctly detected SNPs / number of pre-defined QTL related SNPs=20

We care about recall more than precision, considering precision as a measurement of false positives. A small number of false positive SNPs should not adversely affect EBV prediction. We use the average precision and recall over 100 datasets for each simulation type as the precision and recall for that simulation type.

5.1.1 Effects of Dominance Model

α	β	Precision	Recall
0	0	0.1259	0.4000
0.25	0	0.1887	0.8170
0.5	0	0.5472	**1.0000**
0.75	0	0.1957	0.8790
1	0	0.1806	0.7754

Table 5.1: SNP Mapping precisions and recalls of the bagging EB method on the 5 types of dominance models. α describes the dominance model and β describes the noise level, respectively, in the simulation type.

From Table 5.1 , it is apparent that all of the pre-defined QTL-related SNPs can be detected with the co-dominant model (*i.e.* $\alpha = 0.5$). Even for the completely recessive model (*i.e.* $\alpha = 0$), 40% of the pre-defined QTL-related SNPs can still be detected. The average recall of QTL detection for these five types of dominance models is 0.77428, which is quite a good result for QTL detection according to our knowledge.

5.1.2 Effects of Noise Level

α	β	Precision	Recall
0.5	0	0.5472	**1.0000**
0.5	0.1	0.2557	0.8589
0.5	0.3	0.2504	0.7538
0.5	0.5	0.2240	0.6352
0.5	0.7	0.2135	0.5400
0.5	0.9	0.1963	0.4480

Table 5.2: SNP Mapping precisions and recalls of the bagging EB method on co-dominant model with 5 noise levels. α describes the dominance model and β describes the noise level, respectively, in the simulation type.

From Table 5.2, it is apparent that, even when the background noise level is as large as 0.9, more than 40% of the pre-defined QTL-related SNPs can still be detected. The average recall of QTL detection for these six types of background noise levels is 0.70598.

5.1.3 QTL Mapping using Real Datasets

Dairy Dataset

On the dairy dataset, we compare the SNPs identified by the bagging EB method with the previously reported SNPs that are in LD with the target traits, identified by Kolbehdari *et al.* using more conventional approaches [33].

- Fat Yield

  The bagging EB method identified 59 SNPs for the fat yield trait (see Table 5.3 for more details). In Kolbehdari *et al.*, 20 SNPs were identified as being linked to fat yield QTL (see Table 5.4 for more details). 14 of these were located by the bagging EB method (high-lighted in Table 5.4) [33]. The chromosomal distribution of the 59 SNPs detected by the bagging EB method and the 20 SNPs detected by Kolbehdari *et al.* is illustrated in Figure 5.1. Every circle (asterisk) corresponds to identified SNP by the bagging EB method (by Kolbehdari *et al.*, respectively) and is placed according to the physical position of the SNP on the chromosome. Every line for each chromosome starts from 0 to the length of that chromosome.

SNP ID	SNP NCBI ID	Chr	Position (bp)	SNP ID	SNP NCBI ID	Chr	Position (bp)
BTA-25798	rs41631818	1	57471639	BTA-67252	rs43625320, rs43710950	10	47378178
BTA-51387	rs41637121	1	9825544	BTA-62316	rs43619490, rs41645645	10	32078348
BTA-51334	rs41582208 rs43266357	1	129330659	BTA-85521	rs41660431	11	27756098
BTA-37517	rs41634488 rs43245382	1	82400004	BTA-85502	rs41570374	11	14380343
BTA-56358	rs41643471	1	147169741	BTA-35689	rs41580517	14	8445108
BTA-58283	rs41584322	1	155161056	BTA-34428	rs41579049	14	600616
BTA-32668	rs41629125	1	70388299	BTA-36296	rs41568366	15	27439579
BTA-48448	rs41641845	2	101743148	BTA-37330	rs41606411	15	65711950
BTA-49758	rs41637017	2	133582794	BTA-05419	rs29019578	17	69608829
BTA-67720	rs43712273	3	48985550	BTA-41575	rs41642341	17	66728166
BTA-67008	rs43709929	3	27529775	BTA-41937	rs41634832	17	73637371
BTA-69049	rs41601701	3	107600745	BTA-43272	rs41581637	18	43245535
BTA-03119	rs29010885	3	120330624	BTA-43383	rs41606586	18	45878855
BTA-69344	rs43714172	3	113526879	BTA-44132	rs41583655 rs41897273	18	63546879
BTA-70719	rs41588659	4	12041038	BTA-42463	rs41578926 rs41852614	18	1829874
BTA-72027	rs41591535	4	21357942	BTA-45168	rs41571919	19	36145441
BTA-72646	rs41653936	4	121014987	BTA-44625	rs41640962	19	19223211
BTA-71929	rs41590706	4	104509458	BTA-44769	rs41584901	19	22105731
BTA-70031	rs41648823	4	42247660	BTA-54527	rs42007974 rs41603503	22	43868960
BTA-74391	rs41591894	5	89352039	BTA-55670	rs41640789	23	17701316
BTA-73124	rs41604534	5	30267651	BTA-19607	rs41626402	23	31741848
BTA-74666	rs41592942	5	102184668	BTA-56065	rs41642095	23	27420716
BTA-74378	rs41591891	5	89675691	BTA-56444	rs41588598	23	35537384
BTA-74374	rs41591890	5	89820118	BTA-57584	rs41645253	24	22884730
BTA-76476	rs41653363	6	60256066	BTA-57948	rs41584244	24	35759732
BTA-38255	rs41799542 rs41578761	7	39069058	BTA-58138	rs41601307	24	42308240
BTA-79577	rs41655323 rs43522598	7	62174885	BTA-60610	rs41588786 rs29017003	25	3625175
BTA-48814	rs41585631	7	91545853	BTA-63460	rs41650226	27	16214506
BTA-80348	rs43536843 rs41657346	7	101670999	BTA-63997	rs41653491	28	30354771
BTA-82018	rs41658330	8	86000952				

Table 5.3: Significant SNPs with fat yield identified by bagging EB method.

SNP NCBI ID	Chr	Position (bp)	SNP NCBI ID	Chr	Position (bp)
rs29020642	1	8944582	rs41653025	10	55554590
rs41637121	1	9825544	rs43703342	11	70223472
rs41631818	1	57471639	**rs41579049**	14	600616
rs41629125	1	70388299	**rs41580517**	14	8445108
rs41634488	1	82400004	rs41644615	21	59512214
rs41588659	4	12041038	**rs41640789**	23	17701316
rs41652648	5	89734677	**rs41645253**	24	22884730
rs41591894	5	89352039	rs41653440	28	27379015
rs41592942	5	102184668	rs41569649	28	29126150
rs41578761	7	39069058	**rs41653491**	28	30354771

Table 5.4: Significant SNPs with fat yield identified by Kolbehdari *et al.* [33]

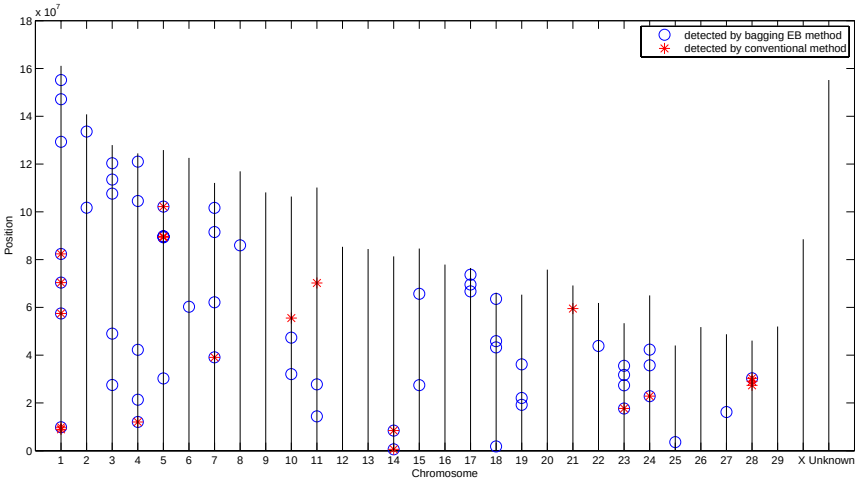

Figure 5.1: The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the fat yield trait.

- Fat Percentage

  The bagging EB method identified 137 SNPs for the fat percentage trait (see Table 5.5 for more details). In Kolbehdari *et al.*, 18 SNPs were identified as being linked to fat percentage QTL (see Table 5.6 for more details). 14 of these were located by the bagging EB method (high-lighted in Table 5.6) [33]. The chromosomal distribution of the 137 SNPs detected by the bagging EB method and the 18 SNPs detected by Kolbehdari *et al.* is illustrated in Figure 5.2.

SNP ID	SNP NCBI ID	Chr	Position (bp)	SNP ID	SNP NCBI ID	Chr	Position (bp)
BTA-50282	rs43266506 rs41639301	1	124766305	BTA-35689	rs41580517	14	8445108
BTA-49418	rs41578215 rs43264225	1	124619364	BTA-63809	rs41587081	14	36931412
BTA-90879	rs41593713 rs43652289	1	64972044	BTA-35529	rs41633631	14	7751463
BTA-51387	rs41637121	1	9825544	BTA-34504	rs41732038 rs41579063	14	31239726
BTA-49758	rs41637017	2	133582794	BTA-63808	rs41587080	14	36931826
BTA-47550	NoName	2	50485349	BTA-34365	rs41627981	14	21749163
BTA-47554	rs41636197	2	50531534	BTA-36071	rs41746511	15	22061623
BTA-49211	rs41578169	2	124588209	BTA-37321	rs41606408	15	64762463
BTA-48448	rs41641845	2	101743148	BTA-14362	rs17870648	15	77489176
BTA-47345	rs41601048	2	37581122	BTA-36903	NoName	15	42919183
BTA-66873	rs41587426	3	21330178	BTA-36417	rs41629524	15	30748428
BTA-69473	rs43714209	3	9265271	BTA-06131	rs29020495	15	4707548
BTA-66828	rs41587408	3	23563371	BTA-93018	rs41663273	15	12055763
BTA-67710	rs43712268	3	46520676	BTA-37898	rs41632676 rs41783454	15	82813803
BTA-68437	rs41591426	3	85497115	BTA-13080	rs29018045	15	33209840
BTA-66884	rs43709850	3	21729249	UCP2-119F1-SNP2	rs41255549	15	52967147
BTA-67008	rs43709929	3	27529775	BTA-40104	rs41634808	16	70486364
BTA-67032	rs43710740	3	28510731	BTA-39644	rs41640597	16	61438252
BTA-68440	rs43709367	3	85504135	BTA-40115	rs41634811	16	70316526
BTA-67006	rs43709927	3	27529597	BTA-41122	rs41570561	17	53839339
BTA-69971	rs43381342 rs41648794	4	32921996	BTA-05419	rs29019578	17	69608829
BTA-70630	rs41651931	4	53780816	BTA-41575	rs41642341	17	66728166
BTA-71929	rs41590706	4	104509458	BTA-06993	rs29021347	18	23275330
BTA-71995	rs41590720 rs43372455	4	21506496	BTA-43272	rs41581637	18	43245535
BTA-71709	rs41603667 rs43412863	4	99108543	BTA-43456	rs41639020	18	52971579
BTA-70719	rs41588659	4	12041038	BTA-42474	rs41862956 rs41635530	18	35101896
BTA-70305	rs41586929 rs43385300	4	51290584	BTA-43586	rs41667443	18	46227259
BTA-72172	rs41591551	4	106221452	BTA-45612	rs41577553	19	49300176
BTA-70031	rs41648823	4	42247660	BTA-45318	rs41576373	19	38107017
BTA-03561	rs29011323	4	63659276	BTA-45274	rs41576348 rs41910839	19	37412442
BTA-74704	rs41654440	5	106919474	BTA-45999	rs41578668 rs41926374	19	57489886
BTA-74374	rs41591890	5	89820118	BTA-46516	rs41641481	19	14133569
BTA-74708	rs41592948	5	107086994	BTA-45737	rs41577583	19	51307694
BTA-14452	rs29023214	5	72940885	BTA-46471	rs41579796	19	13241023
BTA-73401	rs43435053 rs41656716	5	43735760	BTA-50505	rs41640212	20	39860784
BTA-74041	rs41590820	5	80725826	BTA-68718	rs43711332 rs43350564	20	4745147

Continued From Previous Page							
SNP ID	SNP NCBI ID	Chr	Position (bp)	SNP ID	SNP NCBI ID	Chr	Position (bp)
BTA-76476	rs41653363	6	60256066	BTA-52607	rs41608371	21	57134316
BTA-77588	rs41654079	6	17969409	BTA-52481	rs41986690 rs41643783	21	50199271
BTA-77190	rs41655369	6	14138172	BTA-52116	rs41641678 rs41965754	21	6373428
BTA-48814	rs41585631	7	91545853	BTA-53018	rs41565637	21	65108809
BTA-78503	rs41591999	7	19087354	BTA-54823	rs41637661	22	53094097
BTA-00518	rs29013673 rs41588250	7	10644280	BTA-54815	rs41603523	22	53524088
BTA-78885	rs41593188	7	5690741	BTA-53782	rs41640891	22	22353504
BTA-01023	rs29011990	8	31322025	BTA-54516	rs41643865 rs42007232	22	44041777
BTA-63744	rs41587051	8	88219117	BTA-54885	rs42017576 rs41585965	22	54680513
BTA-82711	rs41591739 rs43580165	8	110811676	BTA-54520	NoName	22	43903694
BTA-85058	rs41657163	9	15356131	BTA-54684	rs41644468	22	48236460
BTA-84818	rs41656367	9	98959285	BTA-55670	rs41640789	23	17701316
BTA-67252	rs43625320 rs43710950	10	47378178	BTA-56497	rs41644254	23	40127883
BTA-75267	rs41593881	10	75989024	BTA-56444	rs41588598	23	35537384
BTA-72346	rs41591576	10	61401897	BTA-56321	rs41643446	23	33379805
BTA-15159	rs29026524	10	25201003	BTA-58227	rs41567447	24	42941049
BTA-60288	rs41644756	10	21174069	BTA-57584	rs41645253	24	22884730
BTA-85502	rs41570374	11	14380343	BTA-60378	rs41587828	25	41030299
BTA-88633	NoName	11	30892953	BTA-59785	rs41649668	25	29478794
BTA-96128	rs41665730	11	51401510	BTA-59747	rs42063277 rs41649644	25	26760577
POMC-J00021-254	NoName	11	76263770	BTA-60552	rs41605903	25	4269013
BTA-110430	rs41606063	11	90467185	BTA-61400	rs41648148	26	34505207
BTA-109326	rs41569023	11	89254142	BTA-62026	rs41650578 rs42088425	26	11162266
BTA-85521	rs41660431	11	27756098	BTA-62917	rs41647955	27	39075830
BTA-116618	rs41568304 rs29025976	11	101649690	BTA-63971	rs41587125	28	29071239
BTA-20730	rs41623430	12	36358485	BTA-63997	rs41653491	28	30354771
BTA-02763	rs29010310	12	35004860	BTA-64756	rs41648852	28	11042931
BTA-32313	rs41628258 rs41683033	13	32892499	BTA-63978	rs41587129	28	29116905
BTA-33038	rs41698732 rs41630688	13	58697423	BTA-03525	rs29011289	29	36374464
BTA-32394	rs41566165 rs41686749	13	37823571	BTA-65265	rs41650027	29	31192423
BTA-34428	rs41579049	14	600616	BTA-65568	rs41586159	29	36378367
BTA-35480	rs41567322	14	7873197	BTA-66351	rs43707575	?	137090
BTA-35305	rs41632222	14	62549835				

Table 5.5: Significant SNPs with fat percentage identified by bagging EB method.

SNP NCBI ID	Chr	Position (bp)	SNP NCBI ID	Chr	Position (bp)
rs41587408	3	23563371	**rs41580517**	14	8445108
rs43709929	3	27529775	**rs41579063**	14	31239726
rs29018853	6	80568731	**rs41587081**	14	36931412
rs41657163	9	15356131	rs41639879	17	4328233
rs41592660	9	35968706	**rs41570561**	17	53839339
rs43710950	10	47378178	**rs41641678**	21	6373428
rs41579049	14	600616	**rs41643783**	21	50199271
rs41633631	14	7751463	**rs41640789**	23	17701316
rs41567322	14	7873197	rs41648176	26	37373547

Table 5.6: Significant SNPs with fat percentage identified by Kolbehdari *et al.* [33]

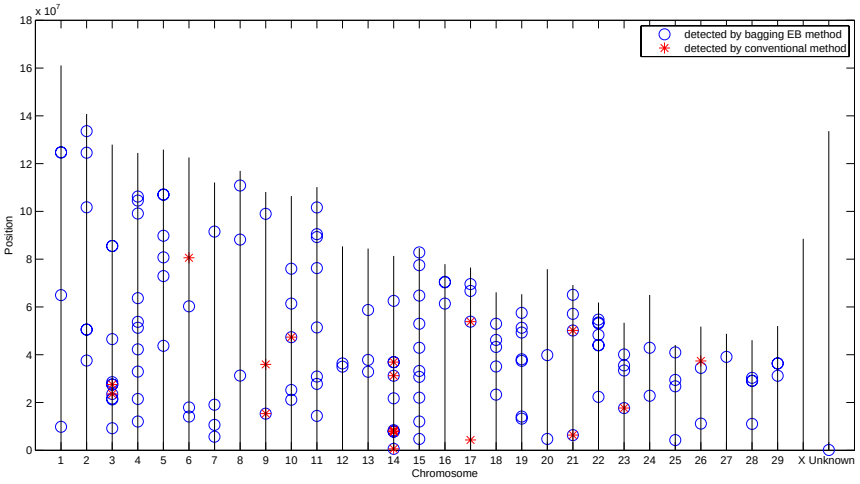

Figure 5.2: The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the fat percentage trait.

- Milk Yield

  The bagging EB method identified 8 SNPs for the milk yield trait (see Table 5.7 for more details). In Kolbehdari *et al.*, 21 SNPs were identified as being linked to milk yield QTL (see Table 5.8 for more details). 4 of these were located by the bagging EB method (high-lighted in Table 5.8) [33]. The chromosomal distribution of the 8 SNPs detected by the bagging EB method and the 21 SNPs detected by Kolbehdari *et al.* is illustrated in Figure 5.3.

SNP ID	SNP NCBI ID	Chr	Position (bp)	SNP ID	SNP NCBI ID	Chr	Position (bp)
BTA-25798	rs41631818	1	57471639	BTA-82018	rs41658330	8	86000952
BTA-74391	rs41591894	5	89352039	BTA-37379	rs41634441	15	65118611
BTA-01562	rs29012523	6	96681568	BTA-52804	rs41585246	21	59696422
BTA-79411	rs41568570	7	61523523	BTA-57215	rs41643632	23	9148248

Table 5.7: Significant SNPs with milk yield identified by the bagging EB method.

SNP NCBI ID	Chr	Position (bp)	SNP NCBI ID	Chr	Position (bp)
rs41631818	1	57471639	rs41633631	14	7751463
rs41629125	1	70388299	rs41628862	14	31340161
rs41633664	1	72671082	rs41587081	14	36931412
rs41643471	1	147169741	rs41632222	14	62549835
rs43709850	3	21729249	rs41581694	18	53174491
rs41655901	5	31294522	rs41577598	19	53085652
rs41656714	5	34116566	rs41608371	21	57134316
rs41592943	5	102164054	rs41644615	21	59512214
rs41658330	8	86000952	**rs41585246**	21	59696422
rs41662488	9	72868606	**rs41643632**	23	9148248
rs41569023	11	89254142			

Table 5.8: Significant SNPs with milk yield identified by Kolbehdari *et al.* [33]

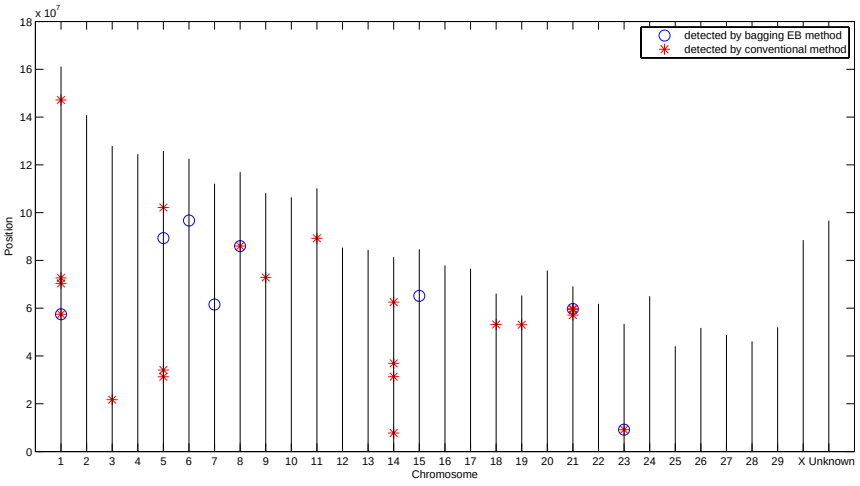

Figure 5.3: The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the milk yield trait.

- Protein Yield

  The bagging EB method identified 73 SNPs for the protein yield trait (see Table 5.9 for more details). In Kolbehdari *et al.*, 15 SNPs were identified as being linked to protein yield QTL (see Table 5.10 for more details). 14 of these were located by the bagging EB method (highlighted in Table 5.10) [33]. The chromosomal distribution of the 73 SNPs detected by the bagging EB method and the 15 SNPs detected by Kolbehdari *et al.* is illustrated in Figure 5.4.

SNP ID	SNP NCBI ID	Chr	Position (bp)	SNP ID	SNP NCBI ID	Chr	Position (bp)
BTA-25798	rs41631818	1	57471639	BTA-109326	rs41569023	11	89254142
BTA-40545	rs41570536	1	138164597	BTA-87550	rs43673339	11	23267364
BTA-37517	rs41634488 rs43245382	1	82400004	BTA-100126	rs41255194 rs41610129	11	5734704
BTA-56358	rs41643471	1	147169741	BTA-33166	rs41576572	13	63707683
BTA-32668	rs41629125	1	70388299	BTA-35305	rs41632222	14	62549835
BTA-47302	rs43303789 rs41645054	2	35552829	BTA-35100	rs41580090	14	53159463
BTA-47097	rs41644050	2	28058808	BTA-36218	rs41628605	15	5117022
BTA-49693	rs41636979	2	130329371	BTA-36533	rs41606337	15	33741493
BTA-49690	rs41579185	2	131520473	UCP2-119F1-SNP1	NoName	15	52967013
BTA-03117	rs29010883	3	120343156	BTA-37712	rs41789652 rs41635079	15	76184495
BTA-67720	rs43712273	3	48985550	BTA-39553	rs41581442	16	2632304
BTA-67710	rs43712268	3	46520676	BTA-41575	rs41642341	17	66728166
BTA-13888	rs29018845	3	127404885	BTA-43450	rs41581694 rs41881775	18	53174491
BTA-08621	rs29022178	3	1748195	BTA-43860	rs41893922 rs41636749	18	56665828
BTA-70719	rs41588659	4	12041038	BTA-43473	rs41639029	18	52354543
BTA-72014	rs41590729	4	21363040	BTA-43272	rs41581637	18	43245535
BTA-70439	rs41650728 rs43389706	4	46392237	BTA-44769	rs41584901	19	22105731
BTA-70974	rs41589533 rs43407614	4	67929148	BTA-46467	rs41641448	19	13446096
BTA-72511	rs41652790	4	117074929	BTA-52702	rs41644615	21	59512214
BTA-03561	rs29011323	4	63659276	BTA-52804	rs41585246	21	59696422
BTA-74376	rs41652648	5	89734677	BTA-51859	rs41973813 rs41640667	21	20350613
BTA-73161	rs41655901	5	31294522	BTA-14847	rs29023606	22	32356601
BTA-73124	rs41604534	5	30267651	BTA-57215	rs41643632	23	9148248
BTA-74666	rs41592942	5	102184668	BTA-56601	rs41588624	23	51470673
BTA-73401	rs43435053 rs41656716	5	43735760	BTA-56127	rs41642130	23	28771706
BTA-01562	rs29012523	6	96681568	BTA-56133	rs41642736	23	28774680
BTA-76424	rs41597173	6	59468702	BTA-57584	rs41645253	24	22884730
BTA-05111	rs29019275	6	31895195	BTA-60610	rs41588786 rs29017003	25	3625175
BTA-78885	rs41593188	7	5690741	BTA-62026	rs41650578 rs42088425	26	11162266
BTA-38255	rs41799542 rs41578761	7	39069058	BTA-61900	rs42109870	26	51720076
BTA-78503	rs41591999	7	19087354	BTA-08995	rs29022551	26	34481741
BTA-79543	rs41595443	7	60350545	BTA-63997	rs41653491	28	30354771
BTA-01023	rs29011990	8	31322025	BTA-63780	rs41606880 rs42142454	28	23479471
BTA-80822	rs41607547	8	18591285	BTA-64465	rs41647684 rs42150400	28	41464821
BTA-82018	rs41658330	8	86000952	BTA-64112	rs41646367	28	34713358
BTA-85294	rs41610029	9	14494768	BTA-65667	rs41652241 rs42183180	29	37723374
BTA-62324	rs41584755	10	6357459				

Table 5.9: Significant SNPs with protein yield identified by the bagging EB method.

SNP NCBI ID	Chr	Position (bp)	SNP NCBI ID	Chr	Position (bp)
rs41631818	1	57471639	**rs41581694**	18	53174491
rs41629125	1	70388299	**rs41636749**	18	56665828
rs41591535	4	21357942	**rs41644615**	21	59512214
rs41578761	7	39069058	**rs41585246**	21	59696422
rs29011990	8	31322025	**rs41643632**	23	9148248
rs41662488	9	72868606	rs41648723	26	45430061
rs41569023	11	89254142	**rs41606880**	28	23479471
rs41628862	14	31340161			

Table 5.10: Significant SNPs with protein yield identified by Kolbehdari *et al.* [33]

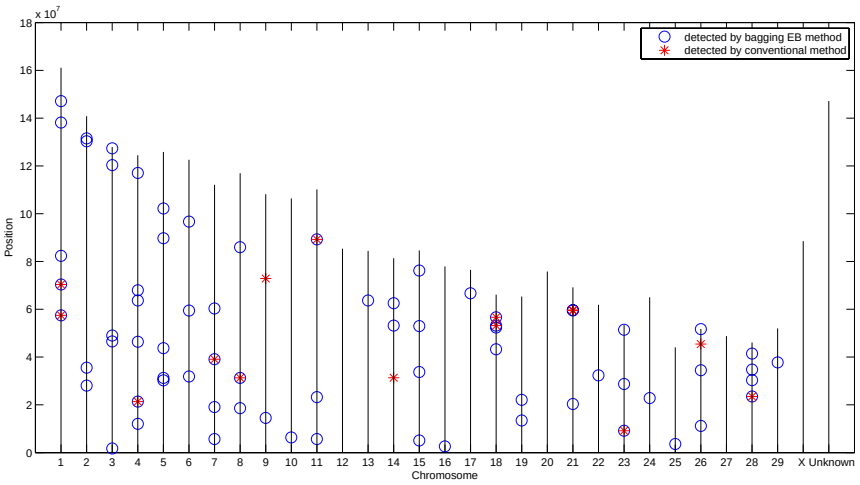

Figure 5.4: The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the protein yield trait.

- Protein Percentage

  The bagging EB method identified 144 SNPs for the protein percentage trait (see Table 5.11 for more details). In Kolbehdari *et al.*, 11 SNPs were identified as being linked to protein percentage QTL (see Table 5.12 for more details). 7 of these were located by the bagging EB method (high-lighted in Table 5.12) [33]. The chromosomal distribution of the 144 SNPs detected by the bagging EB method and the 11 SNPs detected by Kolbehdari *et al.* is illustrated in Figure 5.5.

SNP ID	SNP NCBI ID	Chr	Position (bp)	SNP ID	SNP NCBI ID	Chr	Position (bp)
BTA-50282	rs43266506 rs41639301	1	124766305	BTA-35305	rs41632222	14	62549835
BTA-49418	rs41578215 rs43264225	1	124619364	BTA-63809	rs41587081	14	36931412
BTA-43363	rs41581670	1	99940814	BTA-34428	rs41579049	14	600616
BTA-47237	rs41573015	1	114530617	BTA-35120	rs41580092	14	53423375
BTA-25798	rs41631818	1	57471639	CRH-AF340152-240	NoName	14	30476920
BTA-56358	rs41643471	1	147169741	BTA-37379	rs41634441	15	65118611
BTA-58283	rs41584322	1	155161056	BTA-36296	rs41568366	15	27439579
BTA-66873	rs41587426	3	21330178	BTA-13080	rs29018045	15	33209840
BTA-68437	rs41591426	3	85497115	BTA-10839	rs29016390	15	29531720
BTA-66828	rs41587408	3	23563371	BTA-36533	rs41606337	15	33741493
BTA-69473	rs43714209	3	9265271	BTA-37362	rs41634436	15	65444437
BTA-03117	rs29010883	3	120343156	BTA-36821	rs41606367	15	41084436
BTA-67008	rs43709929	3	27529775	UCP3-119F1-SNP1	NoName	15	53109621
BTA-67150	rs43710804	3	18511564	BTA-37330	rs41606411	15	65711950
BTA-66884	rs43709850	3	21729249	BTA-93018	rs41663273	15	12055763
BTA-67006	rs43709927	3	27529597	BTA-36403	rs41629522	15	30711830
BTA-67726	rs43712279	3	48936455	BTA-37921	rs41783616 rs41632689	15	84403742
BTA-69502	rs43715231	3	117182169	UCP2-119F1-SNP2	rs41255549	15	52967147
BTA-68440	rs43709367	3	85504135	BTA-38218	rs41634657	16	24139267
BTA-70719	rs41588659	4	12041038	BTA-38214	NoName	16	24118162
BTA-02213	rs29009770	4	52897683	BTA-41122	rs41570561	17	53839339
BTA-05538	rs29019697	4	120976229	BTA-41575	rs41642341	17	66728166
BTA-70628	rs43390740 rs41651229	4	53778230	BTA-41557	rs41605775	17	67182104
BTA-71427	rs41653715	4	78988138	BTA-41937	rs41634832	17	73637371
BTA-72640	rs41653934 rs43421610	4	121112757	BTA-41178	rs41639856	17	56142931
BTA-71995	rs41590720 rs43372455	4	21506496	BTA-06993	rs29021347	18	23275330
BTA-70627	rs41602694	4	53769553	BTA-43586	rs41667443	18	46227259
BTA-72646	rs41653936	4	121014987	BTA-43456	rs41639020	18	52971579
BTA-70640	rs41651938	4	53871899	BTA-43817	rs41635828 rs41890212	18	55262674
BTA-14452	rs29023214	5	72940885	BTA-45318	rs41576373	19	38107017
BTA-73397	rs41656714	5	34116566	BTA-45122	rs41915188 rs41585787	19	35996053
BTA-01251	rs29012216	5	43717398	BTA-45675	rs41636128	19	47491519
BTA-74704	rs41654440	5	106919474	BTA-45737	rs41577583	19	51307694
BTA-74077	rs41590827	5	81313130	BTA-45999	rs41578668 rs41926374	19	57489886
BTA-74399	rs41652659	5	89297080	BTA-45274	rs41576348 rs41910839	19	37412442
BTA-77147	rs41655339	6	94872476	BTA-03003	rs29010770	19	23025802
BTA-76945	rs41654417 rs43703009	6	86123130	BTA-44625	rs41640962	19	19223211
BTA-01562	rs29012523	6	96681568	BTA-48448	rs41641845	2	101743148

Continued From Previous Page							
SNP ID	SNP NCBI ID	Chr	Position (bp)	SNP ID	SNP NCBI ID	Chr	Position (bp)
BTA-75696	rs41595102	6	31963668	BTA-68718	rs43711332 rs43350564	20	4745147
BTA-78536	rs41592010	7	19658645	BTA-50442	rs41942245 rs41640182	20	38231035
BTA-78624	rs41658022	7	38195149	BTA-50648	rs41641089	20	3918842
BTA-79604	rs41595457	7	62825607	BTA-51721	rs41639611	21	30670019
BTA-80071	rs41656596	7	88265052	BTA-52629	rs41644583	21	55368160
BTA-80658	rs41658112	7	8962192	BTA-52485	rs41643786	21	50151532
BTA-78828	rs41659935	7	37422303	BTA-52804	rs41585246	21	59696422
BTA-78885	rs41593188	7	5690741	BTA-51888	rs41583332	21	24520032
BTA-82798	rs41591758	8	114064165	BTA-54695	rs41584607	22	48502182
BTA-32914	rs41693478 rs41630632	8	60094652	BTA-54895	rs41565711	22	54320773
BTA-09463	rs29025624	9	41600425	BTA-56321	rs41643446	23	33379805
BTA-84964	rs41588204	9	100456943	BTA-10604	rs29016156	23	16037615
BTA-85058	rs41657163	9	15356131	BTA-56497	rs41644254	23	40127883
BTA-75267	rs41593881	10	75989024	BTA-56444	rs41588598	23	35537384
BTA-67252	rs43625320 rs43710950	10	47378178	BTA-55670	rs41640789	23	17701316
BTA-73370	rs41656695	10	10350925	BTA-58227	rs41567447	24	42941049
BTA-60288	rs41644756	10	21174069	BTA-01508	rs29012470 rs41588810	25	1413389
BTA-62316	rs43619490 rs41645645	10	32078348	BTA-60133	rs41651853 rs43725940	25	38292098
BTA-87573	rs41661853	11	2672910	BTA-60552	rs41605903	25	4269013
BTA-110430	rs41606063	11	90467185	BTA-60378	rs41587828	25	41030299
BTA-96122	rs43663179 rs41665725	11	3083261	BTA-60610	rs41588786 rs29017003	25	3625175
BTA-87550	rs43673339	11	23267364	BTA-59785	rs41649668	25	29478794
BTA-88633	NoName	11	30892953	BTA-62084	rs41606816	26	17965776
BTA-114551	rs41616840	11	96375898	BTA-61400	rs41648148	26	34505207
BTA-20730	rs41623430	12	36358485	BTA-62109	rs41650612 rs42084993	26	18119148
BTA-21642	rs41626908	12	36683471	BTA-63448	rs42116333 rs41650218	27	15950856
BTA-11237	rs29017000	13	18150380	BTA-66883	rs43709849 rs43333444	27	25391156
BTA-32708	rs41687245	13	42753587	BTA-64756	rs41648852	28	11042931
BTA-32313	rs41628258 rs41683033	13	32892499	BTA-63971	rs41587125	28	29071239
BTA-33849	rs41633517	13	74140464	BTA-66155	rs43706971	29	49418532
BTA-33334	rs41566230 rs41709315	13	67626097	BTA-03525	rs29011289	29	36374464
BTA-33840	rs41633516	13	74160790	BTA-65263	rs41584970	29	29174128
BTA-33108	rs41702018 rs41601522	13	60331867	BTA-65386	rs42176863 rs41650995	29	30611970
BTA-63808	rs41587080	14	36931826	BTA-66351	rs43707575	?	137090

Table 5.11: Significant SNPs with protein percentage identified by bagging EB method.

SNP NCBI ID	Chr	Position (bp)	SNP NCBI ID	Chr	Position (bp)
rs41587408	3	23563371	rs41566192	13	38542259
rs41650658	4	51278419	**rs29021058**	13	70531797
rs29014633	5	80625519	**rs41570561**	17	53839339
rs41590827	5	81313130	rs41637636	22	51046178
rs41578761	7	39069058	**rs29016156**	23	16037615
rs41593881	10	75989024			

Table 5.12: Significant SNP with protein percentage identified by Kolbehdari *et al.* [33]

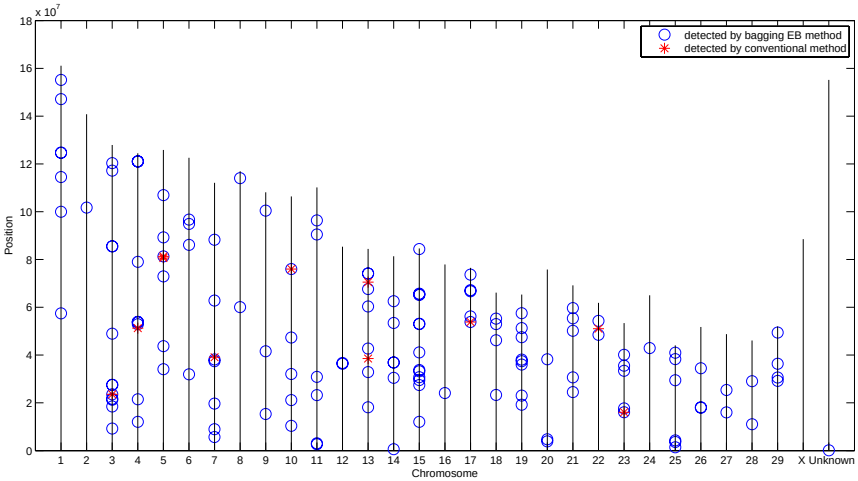

Figure 5.5: The distribution of SNPs located by the bagging EB method (circles) and Kolbehdari *et al.* (asterisks) for the protein percentage trait.

Beef Dataset

- ADG

  The bagging EB method identified 118 SNPs for the ADG trait (see Table 5.13 for more details). The chromosomal distribution of the 118 SNPs (circles) is illustrated in Figure 5.6.

SNP ID	Chr	Position (bp)	SNP ID	Chr	Position (bp)
ARS-BFGL-NGS-49636	1	72115871	ARS-BFGL-NGS-104129	13	14703664
ARS-BFGL-BAC-3570	1	68876072	ARS-BFGL-NGS-25461	13	55420918
ARS-BFGL-NGS-23990	1	94409183	BFGL-NGS-114069	13	80191338
ARS-BFGL-NGS-104338	1	143049917	BTA-26152-no-rs	13	26932538
BTA-89820-no-rs	1	47065260	ARS-BFGL-NGS-11692	13	57473029
BTA-123503-no-rs	1	5222500	UA-IFASA-2669	13	8918717
BFGL-NGS-116361	1	64054878	ARS-BFGL-BAC-1991	14	79193975
BTB-01996267	1	4284068	ARS-BFGL-NGS-107714	14	61455546
ARS-BFGL-BAC-34933	2	117246367	ARS-BFGL-BAC-25195	14	46871848
BTA-29589-no-rs	2	69069976	ARS-BFGL-NGS-84700	14	36497204
ARS-BFGL-BAC-33346	2	14563745	ARS-BFGL-NGS-94777	14	19847065
ARS-BFGL-NGS-104615	2	51431329	BTB-01582460	14	47833670
Hapmap44041-BTA-23382	2	11060396	Hapmap27935-BTC-065354	14	25031801
Hapmap53196-rs29011523	2	122576368	UA-IFASA-7214	14	29767185
ARS-BFGL-NGS-91059	3	17276446	BFGL-NGS-115168	15	77151019
Hapmap35643-SCAFFOLD318144-4314	3	118929393	Hapmap42192-BTA-37799	15	78182163
ARS-BFGL-NGS-104320	3	12634099	ARS-BFGL-NGS-23826	16	18523072
BFGL-NGS-112252	3	1480364	ARS-BFGL-NGS-77702	16	76450290
Hapmap50807-BTA-86717	3	50351708	BFGL-NGS-113782	16	73039934
Hapmap57193-rs29021598	3	57084078	ARS-BFGL-NGS-24777	16	62219285
BTB-00194057	4	68649907	ARS-BFGL-NGS-80663	17	73245015
Hapmap27013-BTA-158242	4	24624405	ARS-BFGL-NGS-107698	17	65658383
BFGL-NGS-110933	5	105049077	ARS-BFGL-NGS-41524	18	65248992
BTA-26132-no-rs	5	45035030	ARS-BFGL-BAC-33724	19	31556338
BTA-15444-no-rs	5	109182394	ARS-BFGL-NGS-28151	19	11121958
Hapmap26308-BTC-057761	6	37963147	ARS-BFGL-NGS-38061	20	69341094
BFGL-NGS-119662	6	121559687	ARS-BFGL-NGS-21312	21	22277790
BTB-00259424	6	63426671	ARS-BFGL-NGS-5033	21	30421616
BTB-01322000	6	10526583	BTB-01275085	21	44103471
Hapmap39094-BTA-75706	6	33328169	Hapmap48168-BTA-106480	22	42520656
Hapmap48464-BTA-77646	6	18119103	Hapmap36428-SCAFFOLD91128-4487	22	55536152
ARS-BFGL-NGS-107035	7	91836262	ARS-BFGL-NGS-28703	22	50897530
BTB-01269021	7	75830004	ARS-BFGL-NGS-62095	22	4942114
BTB-00956439	7	105905909	BFGL-NGS-119695	23	2310949
ARS-BFGL-NGS-44014	7	1343222	BFGL-NGS-116395	23	11896040
Hapmap42417-BTA-108192	7	1304084	ARS-BFGL-BAC-43516	23	1481289
Hapmap24348-BTA-159633	8	70345145	ARS-BFGL-BAC-30064	23	13056038
ARS-BFGL-NGS-67446	8	82931653	ARS-BFGL-NGS-102124	24	9983648
ARS-BFGL-NGS-44247	8	95681398	Hapmap42819-BTA-114761	24	62604210

Continued From Previous Page					
SNP ID	Chr	Position (bp)	SNP ID	Chr	Position (bp)
ARS-BFGL-NGS-34990	8	11558500	ARS-BFGL-BAC-30721	24	3829304
BFGL-NGS-110811	8	72054743	BTB-01448403	24	1670594
Hapmap41091-BTA-81593	8	65835035	Hapmap60714-rs29019480	24	31555267
Hapmap44492-BTA-88469	9	3523789	Hapmap38513-BTA-58574	24	51201344
Hapmap45612-BTA-28859	9	80520854	Hapmap41212-BTA-26034	25	21490449
ARS-BFGL-NGS-55693	9	22678278	ARS-BFGL-NGS-1638	26	44493220
ARS-BFGL-NGS-10386	9	79108555	ARS-BFGL-NGS-32517	26	48311186
BTA-117313-no-rs	9	42477655	ARS-BFGL-NGS-39006	27	38059196
ARS-BFGL-NGS-247	10	69088790	ARS-BFGL-NGS-18023	28	6656202
Hapmap48018-BTA-60331	10	2258968	BTA-100905-no-rs	28	2832256
ARS-BFGL-NGS-55327	10	87407821	Hapmap59655-rs29010842	28	5993743
BTB-00445816	10	95733847	ARS-BFGL-NGS-86495	29	41659431
ARS-BFGL-NGS-105394	10	11991392	BTA-66199-no-rs	29	49924569
ARS-BFGL-NGS-93718	10	15252637	BTA-27090-no-rs	X	82480759
BTB-00445989	10	95935849	ARS-BFGL-NGS-91969	X	28044
ARS-BFGL-NGS-107825	11	88002621	ARS-BFGL-NGS-30224	?	204045
ARS-BFGL-NGS-12433	11	51613089	BFGL-NGS-111321	?	48338
ARS-BFGL-NGS-105689	11	106007656	ARS-BFGL-NGS-70477	?	184737
ARS-BFGL-NGS-28721	12	36308360	BTB-01349004	?	442886
BFGL-NGS-116551	12	11153570	ARS-BFGL-NGS-102954	?	564808

Table 5.13: Significant SNPs with ADG identified by the bagging EB method.

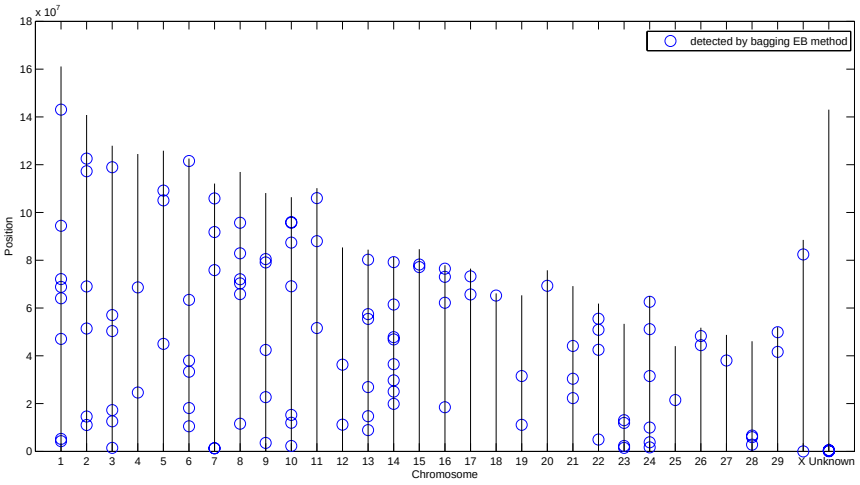

Figure 5.6: The distribution of SNPs located by the bagging EB method for the ADG trait.

- Birth Weight

  The bagging EB method identified 74 SNPs for the birth weight trait (see Table 5.14 for more details). The chromosomal distribution of the 74 SNPs is illustrated in Figure 5.7.

SNP ID	Chr	Position (bp)	SNP ID	Chr	Position (bp)
Hapmap47302-BTA-45176	1	102365562	ARS-BFGL-BAC-23075	14	66302948
ARS-BFGL-NGS-34687	1	46663776	ARS-BFGL-BAC-19427	15	57106189
Hapmap41804-BTA-24071	1	93283602	BTB-01786587	15	10317507
ARS-BFGL-NGS-36803	2	63623562	BTB-00587441	15	23622087
Hapmap44597-BTA-47277	2	33028755	BFGL-NGS-114653	15	54681301
BTB-01071390	2	63146406	BTB-01296218	15	15505958
ARS-BFGL-NGS-17147	2	6502336	Hapmap47274-BTA-37477	15	70477467
ARS-BFGL-NGS-103120	3	14284629	ARS-BFGL-BAC-34362	17	10728397
BTA-68842-no-rs	3	101025935	ARS-BFGL-NGS-31362	17	13089206
ARS-BFGL-NGS-36551	3	109411154	ARS-BFGL-NGS-101525	17	9367689
BTB-01493530	4	119764535	BFGL-NGS-118873	17	73529631
BFGL-NGS-110046	4	87394228	BTB-00745347	19	28700644
Hapmap23022-BTA-161235	5	34648814	ARS-BFGL-NGS-37850	19	32807055
ARS-BFGL-NGS-72188	6	41541414	BTB-00745293	19	28671067
ARS-BFGL-NGS-29722	6	97364829	ARS-BFGL-BAC-27914	20	30128561
BTB-01900621	6	81244923	ARS-BFGL-BAC-31759	20	71553129
Hapmap23186-BTC-046762	6	41208356	ARS-BFGL-NGS-30365	21	22006620
BTA-26162-no-rs	6	57416736	ARS-BFGL-NGS-89251	21	33370896
Hapmap33170-BTC-071249	6	38756335	ARS-BFGL-NGS-4067	22	30952220
BTA-100891-no-rs	6	38076963	BFGL-NGS-118012	23	38509097
Hapmap27537-BTC-060891	6	38638962	ARS-BFGL-BAC-29235	24	49701953
ARS-BFGL-NGS-57673	8	87229588	Hapmap50403-BTA-58218	24	44777710
ARS-BFGL-NGS-10386	9	79108555	ARS-BFGL-NGS-37419	24	1011550
BTB-00385217	9	25430999	ARS-BFGL-NGS-36333	25	39561967
BTA-84883-no-rs	9	97381353	Hapmap54366-rs29015052	25	17517068
Hapmap41972-BTA-79298	10	87622309	ARS-BFGL-NGS-86242	27	4521654
ARS-BFGL-NGS-32292	10	25221613	BTA-63422-no-rs	27	15259667
ARS-BFGL-NGS-29976	11	102812100	Hapmap49694-BTA-64179	28	36295371
BFGL-NGS-114345	11	23286507	Hapmap38041-BTA-64824	28	12122846
ARS-BFGL-BAC-15562	11	70364978	Hapmap51476-BTA-63589	28	14635797
ARS-BFGL-NGS-26350	11	76122798	ARS-BFGL-NGS-101448	29	45972848
Hapmap26194-BTA-158111	11	20901523	Hapmap30567-BTA-151983	X	43569929
Hapmap52802-ss46526242	11	106092699	ARS-BFGL-NGS-36429	X	64480988
ARS-BFGL-BAC-14357	12	64406055	ARS-BFGL-NGS-26893	X	65813619
ARS-BFGL-NGS-16128	13	15267627	Hapmap60033-rs29012094	X	39452099
BTA-03159-no-rs	13	50414986	ARS-BFGL-NGS-102959	X	66444708
Hapmap39263-BTA-33871	13	77245638	Hapmap39109-BTA-30591	?	364178

Table 5.14: Significant SNPs with birth weight identified by the bagging EB method.

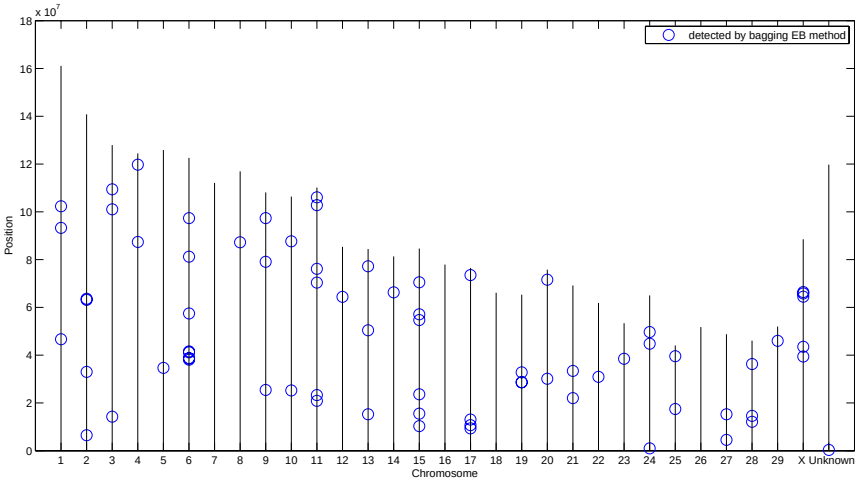

Figure 5.7: The distribution of SNPs located by the bagging EB method for the birth weight trait.

- RFI

  The bagging EB method identified 108 SNPs for the RFI trait (see Table 5.15 for more details).

  The chromosomal distribution of the 108 SNPs is illustrated in Figure 5.8.

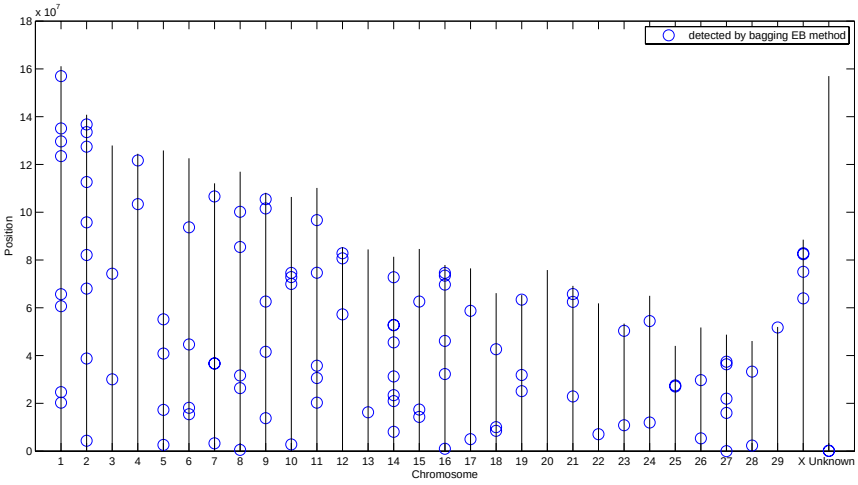

Figure 5.8: The distribution of detected SNPs with RFI on genome located by the bagging EB method.

SNP ID	Chr	Position (bp)	SNP ID	Chr	Position (bp)
ARS-BFGL-NGS-34197	1	129626479	ARS-BFGL-NGS-103773	12	80698939
BTA-49236-no-rs	1	24657856	ARS-BFGL-BAC-15708	13	16221252
BTB-00027149	1	65712986	ARS-BFGL-NGS-65910	14	72750829
BTA-49381-no-rs	1	123444391	BTB-00563522	14	31305251
ARS-BFGL-BAC-11057	1	135081108	Hapmap49133-BTA-35076	14	52852261
ARS-BFGL-NGS-31767	1	156966213	ARS-BFGL-NGS-101613	14	45512732
BFGL-NGS-110620	1	60663283	BTB-02056709	14	23394459
BFGL-NGS-116430	1	20294179	ARS-BFGL-NGS-66553	14	20929826
ARS-BFGL-BAC-33163	2	4344036	ARS-BFGL-NGS-71623	14	8065775
ARS-BFGL-NGS-88498	2	133522697	BTB-01790975	14	52715978
UA-IFASA-3968	2	68006157	ARS-BFGL-NGS-13356	15	17437755
ARS-BFGL-NGS-36162	2	112633454	Hapmap44375-BTA-37785	15	14339378
ARS-BFGL-NGS-15468	2	127414821	ARS-BFGL-NGS-71816	15	62626005
BTB-01860738	2	82120284	ARS-BFGL-NGS-106467	16	46125996
ARS-BFGL-NGS-39303	2	136766494	ARS-BFGL-NGS-102714	16	73465805
Hapmap35082-BES9-Contig524-1040	2	95762272	Hapmap60572-rs29010980	16	74557208
Hapmap52382-rs29019574	2	38762006	ARS-BFGL-NGS-36880	16	69727165
ARS-BFGL-NGS-26956	3	30065978	Hapmap38110-BTA-38625	16	32278543
BTB-00134661	3	74265239	Hapmap42977-BTA-55653	16	965541
BTA-71817-no-rs	4	103395988	BTB-00667048	17	5044986
BTB-01148141	4	121700061	Hapmap48751-BTA-41232	17	58707513
Hapmap55237-rs29010308	5	17272101	Hapmap34615-BES10-Contig586-1594	18	8552751
BTB-00214293	5	2595077	ARS-BFGL-NGS-21227	18	42661333
Hapmap58596-rs29013530	5	40852253	Hapmap30624-BTA-42324	18	10031352
BTA-73549-no-rs	5	55176318	ARS-BFGL-NGS-88748	19	63380264
Hapmap30691-BTC-038216	6	44643940	ARS-BFGL-NGS-57209	19	31896076
BTB-01428731	6	93704721	BFGL-NGS-110331	19	25110983
Hapmap48464-BTA-77646	6	18119103	ARS-BFGL-BAC-2524	21	65804493
Hapmap27487-BTA-90787	6	15457973	BTA-24891-no-rs	21	62502471
Hapmap40325-BTA-80477	7	106599922	ARS-BFGL-NGS-71975	21	22886913
ARS-BFGL-NGS-109201	7	36645610	ARS-BFGL-BAC-2600	22	7116084
BFGL-NGS-115333	7	3274931	Hapmap49546-BTA-25249	23	50364385
Hapmap47734-BTA-78811	7	36784990	Hapmap54795-rs29014478	23	10843024
BFGL-NGS-111964	8	100157121	BTB-01970944	24	11943588
ARS-BFGL-NGS-33628	8	26365756	Hapmap55314-rs29026474	24	54401219
ARS-BFGL-NGS-5096	8	567527	ARS-BFGL-NGS-21527	25	27022645
BTB-01092452	8	85420151	BTA-110448-no-rs	25	27513378
BTB-00285653	8	31663727	ARS-BFGL-NGS-13248	26	5318047
ARS-BFGL-NGS-10254	9	41579851	Hapmap59240-rs29019735	26	29771636
BTB-01715634	9	13821086	ARS-BFGL-NGS-108861	27	37445592
ARS-BFGL-NGS-85461	9	105471834	ARS-BFGL-NGS-105349	27	36341427
ARS-BFGL-NGS-104698	9	101590454	ARS-BFGL-NGS-42178	27	22286
Hapmap43073-BTA-83925	9	62572059	BTB-00959704	27	21942706
ARS-BFGL-BAC-7166	10	74550952	ARS-BFGL-NGS-39294	27	15979955
ARS-BFGL-NGS-106325	10	2841923	BFGL-NGS-111706	28	33312204
BTB-00434073	10	72886486	ARS-BFGL-NGS-14076	28	2334737
Hapmap42386-BTA-98732	10	69999527	BFGL-NGS-116005	29	51766334
ARS-BFGL-NGS-1846	11	96707761	Hapmap41822-BTA-30608	X	75065677
ARS-BFGL-NGS-14449	11	30628461	BTA-21001-no-rs	X	82814214
ARS-BFGL-BAC-16175	11	20279452	ARS-BFGL-NGS-18028	X	63961868
BFGL-NGS-115114	11	74699947	BTA-27090-no-rs	X	82480759
BTB-01293391	11	35716271	ARS-BFGL-NGS-16263	?	181736
ARS-BFGL-NGS-18439	12	82833172	BTB-00118945	?	74933
BTB-02035517	12	57219962	BFGL-NGS-117409	?	51744

Table 5.15: Significant SNPs with RFI identified by the bagging EB method.

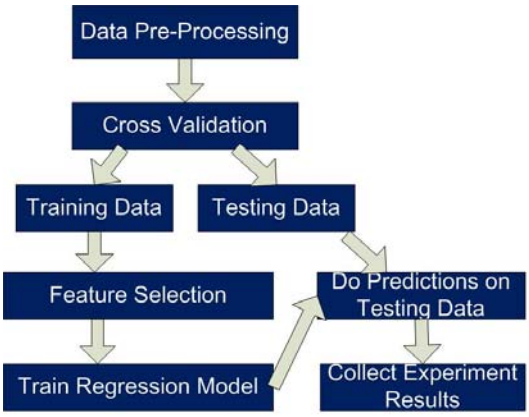

Figure 5.9: Experiment procedure used to evaluate the performance of EBV prediction methods

5.2 EBV Prediction

5.2.1 Experiment Design

Our experiment design for EBV prediction is illustrated in Figure 5.9. After data pre-processing as mentioned in the previous section, we use 10-fold cross validation by partitioning the data into 10 folds. One round of cross-validation involves validating the model on one fold (called the testing set) while the model is learned by performing feature selection and model selection on the other folds (called the training set).

Cross-validation is a way to estimate the average accuracy of a model. Suppose we have a model with one or more unknown parameters, and a dataset to which the model can be fit (the training dataset). The fitting process optimizes the model parameters to make the model fit the training data as well as possible. As a result, the model fits perfectly on the training data, but it might not fit well on the independent validation data extracted from the same population as the training data. This problem is particularly likely to happen when the size of the training data set is small, or when the number of parameters in the model is large.

5.2.2 Performance Measurements

Three performance measurements are used in our experiment, which are the *correlation coefficient* (CC), the *rank correlation coefficient* (rCC) and the *normalized root mean square error* (NRMSE).

CC

CC indicates the strength and direction of a linear relationship between two variables. The range of CC is $[-1, 1]$. A CC value of 1 indicates an increasing linear relationship, a CC value of -1 indicates a decreasing linear relationship and CC values between -1 and 1 indicate the degree of linear dependence between the two variables. We usually take the absolute value of CC. The closer the coefficient is to 1, the stronger the correlation between the variables. A CC value of 0 means the

variables have no correlation. The equation for the calculation of CC is as follows:

$$CC_{y'y} = \frac{n\sum y_i' y_i - \sum y_i' \sum y_i}{\sqrt{n\sum y_i'^2 - (\sum y_i')^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{5.1}$$

where n is the number of observations, y is the vector of target values and $y^{'}$ is the vector of the predicted values.

rCC

rCC is the CC between target ranking and predicted ranking. We rank both the target trait values and predicted trait values, and then calculate the CC between these two rankings using the same formula as CC. The reason for using this measurement is, in livestock industry, breeding animals are usually selected according to the EBV ranking, *e.g.* the top 20% animals are selected by ranking EBVs of the economically important trait.

NRMSE

The *root mean square error* (RMSE) is a frequently used measurement of the differences between values predicted by a model and the actual values. NRMSE is a further normalization of the RMSE, calculated by dividing the RMSE by the standard deviation of the actual values that we are trying to predict. NRMSE is quite useful because we can compare the accuracy of the model across different datasets. NRMSE will reach the value of 1 if the method of prediction is no more accurate than forecasting the unconditional mean of the prediction set while a value of 0 corresponds to a perfect fit. The equation for the calculation of NRMSE is as follows:

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n}(y_i' - y_i)^2/n}}{std(y)} \tag{5.2}$$

where n is the number of observations, y is the vector of target values, $y^{'}$ is the vector of the predicted values and $std(y)$ is the standard deviation of the target values.

5.2.3 Methods Implementation

As discussed previously, seven machine learning regression methods were used for EBV prediction. All of these methods were implemented in R by using existing R packages. The specific packages and methods used are given in Table 5.16:

5.2.4 Comparison of Algorithm Performance on Real Datasets

Dairy Dataset

We used 10 methods, SVM with a linear kernel, SVM with an rbfdot kernel, GP, PCA, PLS, LASSO, ElasticNet, Ridge, LR, and BLUP for EBV prediction and measured their performance using CC, rCC, and NRMSE. Figure 5.10 plots the CCs of these 10 methods for the 5 traits from the dairy

	R Package	Version	R Method
SVM (linear kernel)	kernlab	0.9-5	ksvm
SVM (rbfdot kernel)	kernlab	0.9-5	ksvm
GP	kernlab	0.9-5	gausspr
PCA	pls	2.1-0	pcr
PLS	pls	2.1-0	plsr
LASSO	lars	0.9-7	lars
ElasticNet	elasticnet	1.0-3	cv.enet
Ridge	MASS	7.2-40	lm.ridge

Table 5.16: R implementation for machine learning regression methods.

dataset. Table 5.17 lists all the CCs, rCCs, and NRMSEs of the 10 methods for the 5 traits from the dairy dataset. For each trait, the top 3 performing methods are highlighted. As we can see, the CC and rCC measurements are consistent with each other in most cases, while NRMSE does not always match up with the above two measurements in terms of ranking. We focus only on the CC measurement of performance during our experiments. Please refer to Section 5.3 for comparisons of rCCs and NRMSEs.

From the table, one can see that the machine learning methods outperform the traditional genomic selection method BLUP on all of the five traits. No single machine learning method performed the best on all traits. The best method in this experiment is Ridge with an average CC of 0.49724 on the 5 traits, followed by SVM (linear kernel) with an average CC of 0.4514 and then PLS with an average CC of 0.4496.

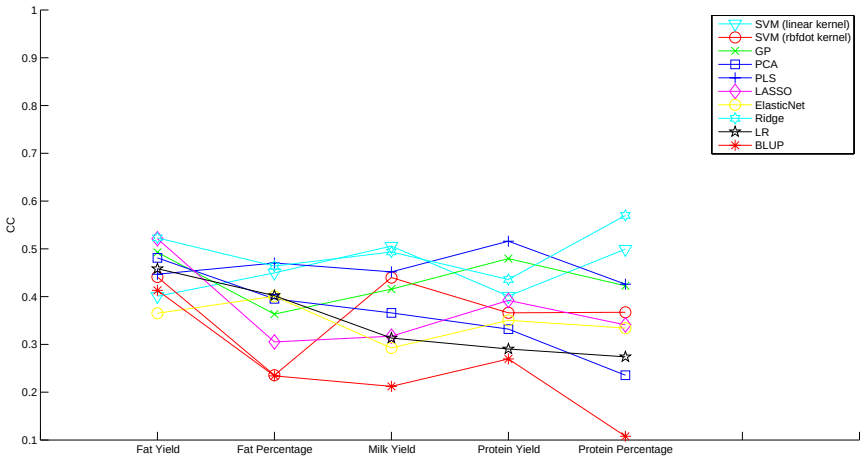

Figure 5.10: Performance of EBV prediction algorithms on dairy dataset by CC measurement, which shows that for fat yield trait, Ridge is the best method; for fat percentage trait, PLS is the best method; for milk yield trait, SVM (linear kernel) is the best method; for protein yield trait, PLS is the best method and for protein percentage trait, Ridge is the best method.

Trait	Measure	SVM (linear)	SVM (rbfdot)	GP	PCA	PLS	LASSO	Elastic Net	Ridge	LR	BLUP
FY	CC	0.401	0.441	**0.493**	0.481	0.447	**0.521**	0.365	**0.523**	0.459	0.413
	rCC	0.393	0.448	0.487	**0.493**	0.461	**0.510**	0.376	**0.526**	0.425	0.472
	NRMSE	0.991	**0.674**	0.858	**0.658**	0.901	0.854	0.740	0.903	0.908	**0.732**
FP	CC	**0.450**	0.236	0.363	0.395	**0.470**	0.305	0.401	**0.464**	0.402	0.234
	rCC	**0.462**	0.227	0.345	0.406	**0.434**	0.331	0.423	**0.473**	0.392	0.252
	NRMSE	0.955	0.816	0.934	**0.701**	0.899	0.977	**0.662**	0.926	0.948	**0.785**
MY	CC	**0.506**	0.440	0.416	0.366	**0.452**	0.317	0.293	**0.493**	0.313	0.212
	rCC	0.420	**0.432**	0.391	0.313	**0.441**	0.305	0.298	**0.470**	0.328	0.237
	NRMSE	0.872	**0.660**	0.901	**0.708**	0.895	0.965	**0.768**	0.933	0.975	0.799
PY	CC	0.401	0.366	**0.480**	0.332	**0.516**	0.392	0.351	**0.436**	0.290	0.270
	rCC	0.322	0.380	**0.438**	0.350	**0.461**	0.385	0.345	**0.408**	0.289	0.289
	NRMSE	1.031	**0.737**	0.875	**0.709**	0.889	0.932	**0.696**	0.895	0.973	0.767
PP	CC	**0.500**	0.367	0.423	0.236	**0.426**	0.341	0.334	**0.570**	0.274	0.108
	rCC	**0.504**	0.362	0.357	0.215	**0.401**	0.350	0.326	**0.523**	0.291	0.276
	NRMSE	0.890	**0.765**	0.950	**0.794**	0.930	0.960	**0.714**	0.852	0.955	0.828

Table 5.17: The average CCs, rCCs, and NRMSEs of 10 methods for EBV prediction on 5 traits from the dairy dataset.

Beef Dataset

Similarly for beef dataset, Table 5.18 lists the CCs, rCCs, and NRMSEs of the 10 methods on the 3 traits. For each trait, the top 3 performing methods are again highlighted. From the table, one can see that BLUP again performed poorly. The best accuracy achieved by the machine learning methods is over 20% higher than that of BLUP on all of the three traits. Similarly, no single machine learning method won out all the time. Specifically, PLS is the best method for predicting ADG trait, while GP is the best method for predicting both birth weight and RFI traits. Compared to the dairy dataset, the CCs achieved on the three traits here are much higher, for example, 0.733 for the RFI trait. Note that beef dataset contains many more SNP markers genotyped using a better chip, which likely make it a better dataset for quantitative association studies. By averaging the results of each method on all of the 3 traits, GP is the best method in this experiment with an average CC measurement of 0.6421, followed by Ridge with an average CC of 0.6128 and then PLS with an average CC of 0.6062.

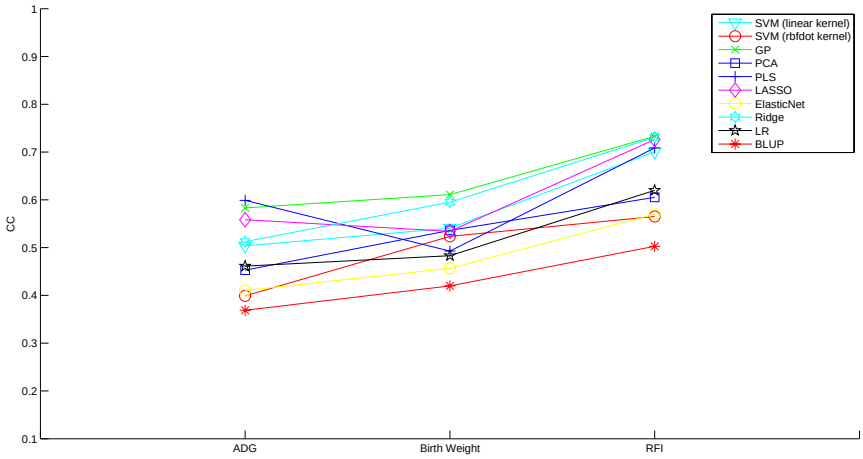

Figure 5.11: Performance of EBV prediction algorithms on beef dataset by CC measurement, which shows that for ADG trait, PLS is the best method; for birth weight trait, GP is the best method and for RFI trait, GP is the best method.

Trait	Measure	SVM (linear)	SVM (rbfdot)	GP	PCA	PLS	LASSO	Elastic Net	Ridge	LR	BLUP
ADG	CC	0.504	0.399	**0.583**	0.453	**0.599**	**0.558**	0.410	0.513	0.461	0.369
	rCC	0.470	0.369	**0.547**	0.414	**0.530**	**0.527**	0.353	0.462	0.471	0.421
	NRMSE	0.879	**0.681**	0.819	**0.635**	0.816	0.827	**0.687**	0.875	0.943	0.762
BW	CC	**0.541**	0.523	**0.611**	0.536	0.493	0.534	0.457	**0.595**	0.483	0.420
	rCC	0.512	0.479	**0.560**	**0.516**	0.467	0.480	0.431	**0.584**	0.380	0.392
	NRMSE	0.852	**0.578**	0.807	**0.598**	0.866	0.837	**0.644**	0.810	0.888	0.748
RFI	CC	0.701	0.565	**0.733**	0.605	0.709	**0.726**	0.570	**0.731**	0.620	0.503
	rCC	**0.646**	0.469	**0.653**	0.543	0.584	0.602	0.520	**0.634**	0.505	0.466
	NRMSE	0.713	**0.546**	0.672	**0.537**	0.713	0.687	**0.511**	0.685	0.854	0.705

Table 5.18: The average CCs, rCCs, and NRMSEs of 10 methods for EBV prediction on 3 traits from the beef dataset.

5.2.5 Comparison of Algorithm Performance on Simulation Datasets

On the simulation datasets, we followed the same procedure as used with the two real datasets for EBV prediction. Again we used 10 methods, SVM with a linear kernel, SVM with an rbfdot kernel, GP, PCA, PLS, LASSO, ElasticNet, Ridge, LR, and BLUP. For each collection of 100 simulation datasets, indexed from 1 to 100, we calculated the CCs, rCCs, and NRMSEs for every EBV prediction method. For instance, Figure 5.12 plots the CCs of these 10 methods on the 100 completely recessive simulation datasets without background noise. In the plot, the horizontal lines show the average CCs of these methods on the 100 datasets, and the methods are sorted accordingly top-down in decreasing performance. Please refer to Section 5.3 for comparisons of rCCs and NRMSEs, as

we focus on the CC measurement of performance during our experiments.

- completely recessive ($\alpha = 0, \beta = 0$)

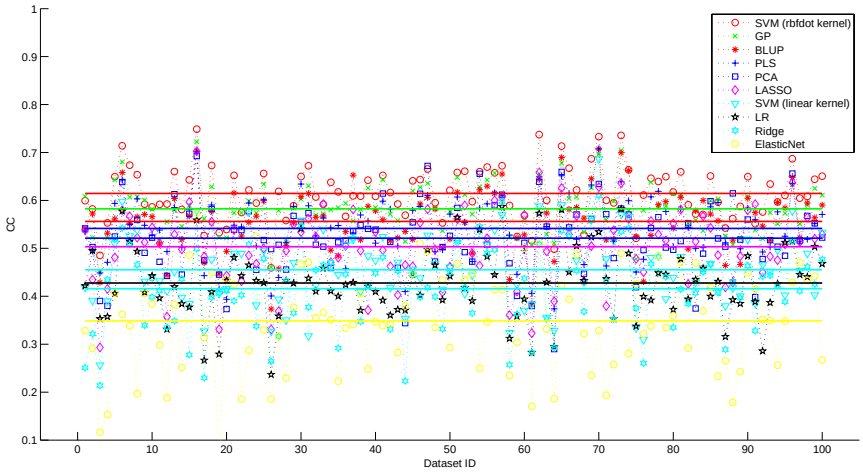

Figure 5.12: Performance of EBV prediction algorithms for the completely recessive model by CC measurement, SVM (rbfdot kernel), GP and BLUP are the top 3 methods for this model.

- partially recessive ($\alpha = 0.25, \beta = 0$)

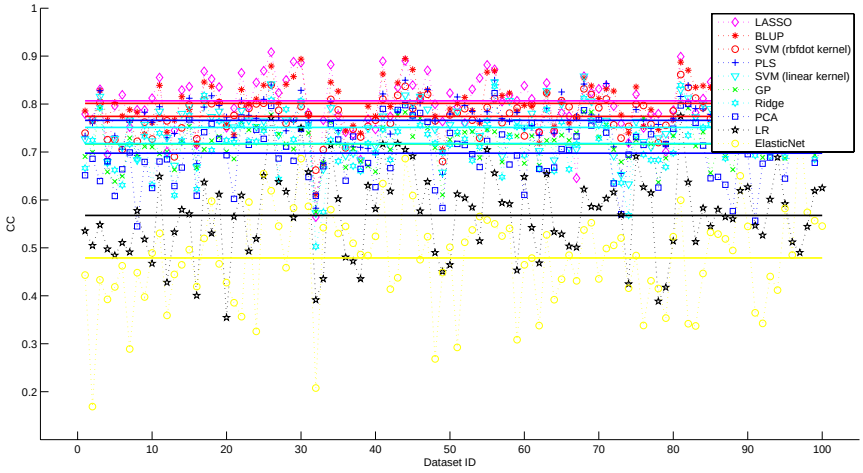

Figure 5.13: Performance of EBV prediction algorithms for the partially recessive model by CC measurement, LASSO, BLUP and SVM (rbfdot kernel) are the top 3 methods for this model.

- co-dominant ($\alpha = 0.5, \beta = 0$)

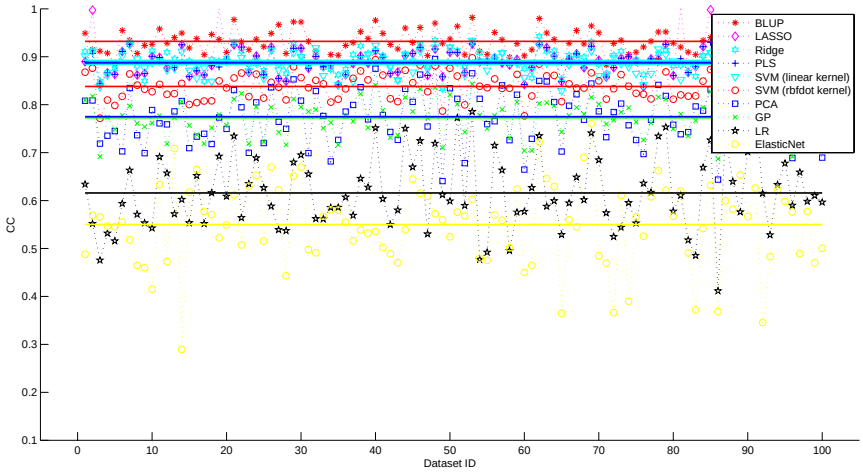

Figure 5.14: Performance of EBV prediction algorithms for the co-dominant model by CC measurement, BLUP, LASSO and Ridge are the top 3 methods for this model.

- partially dominant ($\alpha = 0.75, \beta = 0$)

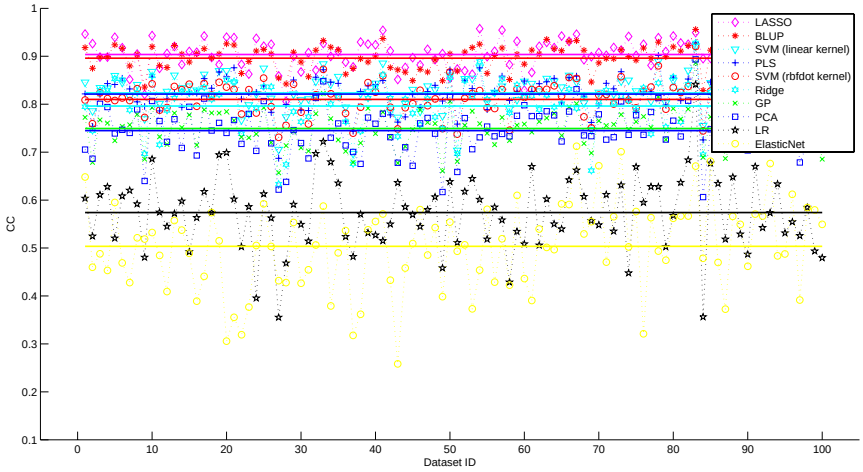

Figure 5.15: Performance of EBV prediction algorithms for the partially dominant model by CC measurement, LASSO, BLUP and SVM (linear kernel) are the top 3 methods for this model.

- completely dominant ($\alpha = 1, \beta = 0$)

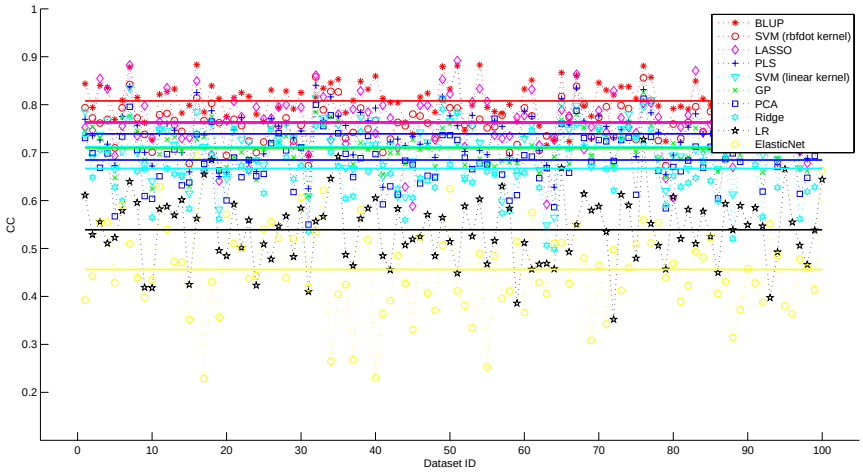

Figure 5.16: Performance of EBV prediction algorithms for the completely dominant model by CC measurement, BLUP, SVM (rbfdot kernel) and LASSO are the top 3 methods for this model.

- co-dominant model with background noise level 0.1 ($\alpha = 0.5, \beta = 0.1$)

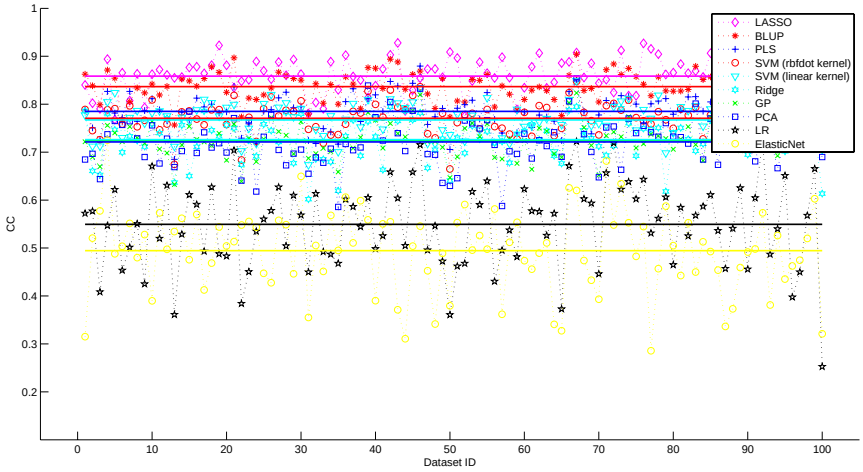

Figure 5.17: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.1 model by CC measurement, LASSO, BLUP and PLS are the top 3 methods for this model.

- co-dominant model with background noise level 0.3 ($\alpha = 0.5, \beta = 0.3$)

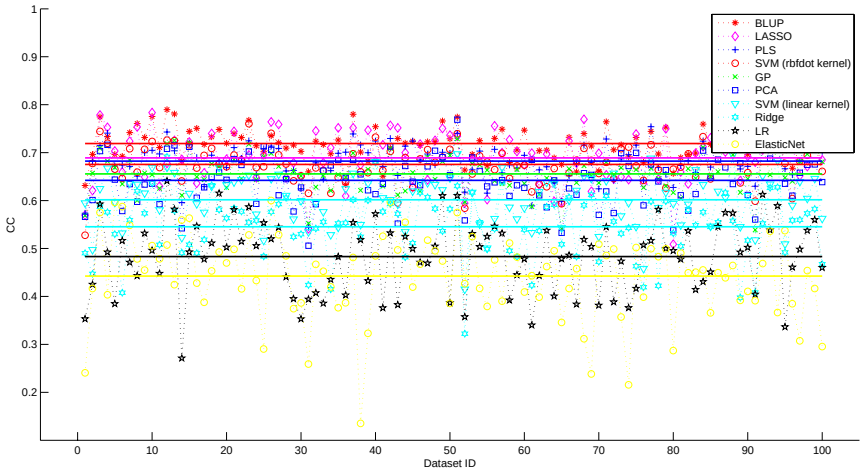

Figure 5.18: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.3 model by CC measurement, BLUP, LASSO and PLS are the top 3 methods for this model.

- co-dominant model with background noise level 0.5 ($\alpha = 0.5, \beta = 0.5$)

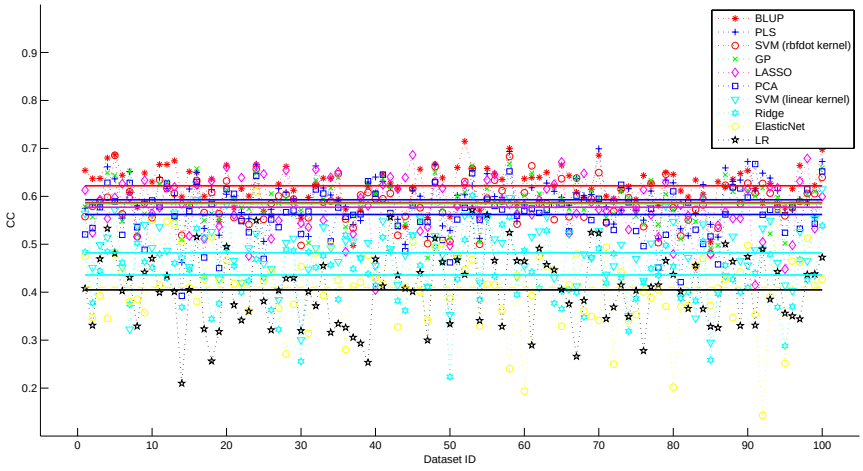

Figure 5.19: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.5 model by CC measurement, BLUP, PLS and SVM (rbfdot kernel) are the top 3 methods for this model.

- co-dominant model with background noise level 0.7 ($\alpha = 0.5, \beta = 0.7$)

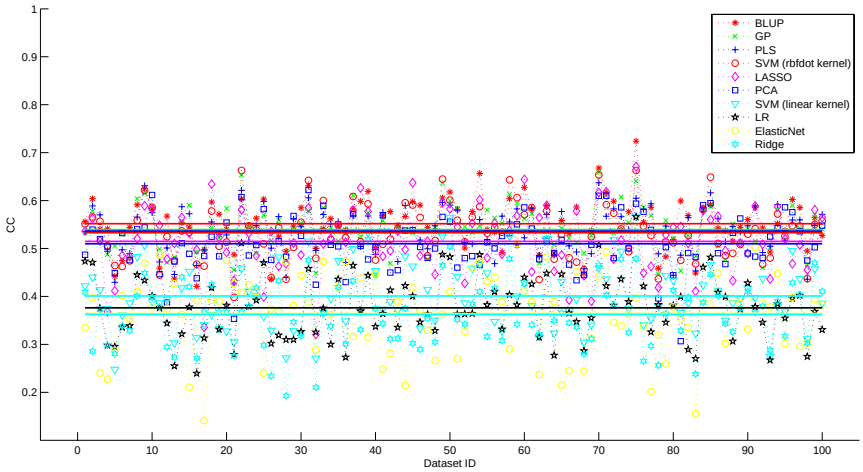

Figure 5.20: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.7 model by CC measurement, BLUP, GP and PLS are the top 3 methods for this model.

- co-dominant model with background noise level 0.9 ($\alpha = 0.5, \beta = 0.9$)

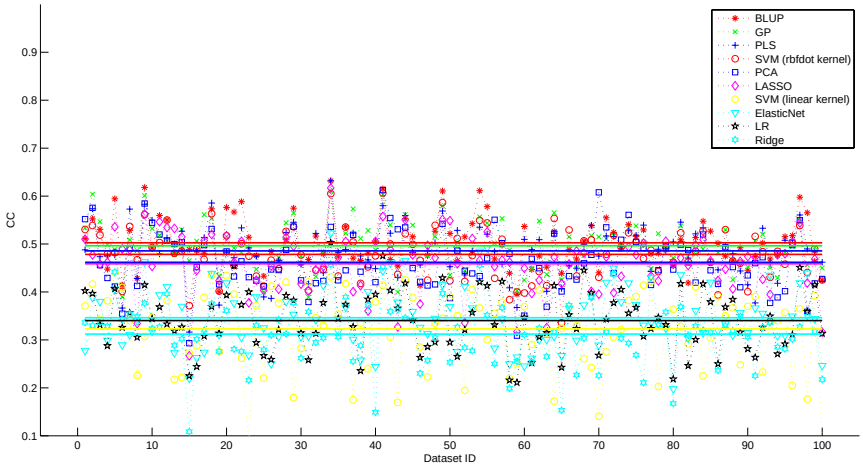

Figure 5.21: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.9 model by CC measurement, BLUP, GP and PLS are the top 3 methods for this model.

Table 5.19 lists the average CCs, rCCs, and NRMSEs of the 10 methods on the 10 types of simulation datasets. For each type, the top 3 performing methods are highlighted. From the table, one can see that the traditional genomic selection method BLUP performed quite well on all types of simulation datasets, though not always the best. In general, we conclude that, besides BLUP, SVM with an rbfdot kernel, PLS, and LASSO have better performance than the others. It is especially

interesting to notice that for the co-dominance simulation datasets without background noise, an average CC of 0.9 can be reached by several methods. Even for the co-dominance simulation datasets with background noise level 0.5, the achieved accuracy (a CC value of 0.622) is better than the best performance achieved on the dairy dataset (a CC value of 0.57).

(α, β)	Measure	SVM (linear)	SVM (rbfdot)	GP	PCA	PLS	LASSO	Elastic Net	Ridge	LR	BLUP
$(0, 0)$	CC	0.455	**0.615**	**0.583**	0.521	0.542	0.503	0.349	0.416	0.428	**0.556**
	rCC	0.452	**0.603**	**0.566**	0.508	0.530	0.485	0.459	0.416	0.430	**0.552**
	NRMSE	1.076	**0.789**	**0.835**	**0.867**	0.871	0.894	1.736	1.066	1.092	0.879
$(0.25, 0)$	CC	0.751	**0.774**	0.717	0.698	0.766	**0.807**	0.479	0.718	0.568	**0.801**
	rCC	0.744	**0.771**	0.717	0.690	0.758	**0.793**	0.581	0.710	0.581	**0.801**
	NRMSE	0.675	**0.645**	0.754	0.716	0.649	**0.648**	1.169	0.830	0.987	**0.601**
$(0.75, 0)$	CC	**0.823**	0.810	0.749	0.745	0.821	**0.904**	0.503	0.796	0.574	**0.896**
	rCC	**0.811**	0.803	0.746	0.730	0.808	**0.893**	0.605	0.782	0.591	**0.891**
	NRMSE	**0.567**	0.604	0.730	0.666	0.572	**0.497**	1.746	0.698	1.113	**0.443**
$(1, 0)$	CC	0.712	**0.764**	0.710	0.685	0.740	**0.762**	0.456	0.667	0.539	**0.808**
	rCC	0.700	**0.752**	0.702	0.669	0.726	**0.753**	0.579	0.654	0.554	**0.802**
	NRMSE	0.737	**0.652**	0.757	0.732	**0.683**	0.709	1.725	0.790	1.160	**0.599**
$(0.5, 0)$	CC	0.885	0.838	0.773	0.775	0.887	**0.891**	0.550	**0.892**	0.616	**0.932**
	rCC	0.875	0.833	0.772	0.761	0.875	**0.880**	0.643	**0.881**	0.634	**0.928**
	NRMSE	0.467	0.570	0.711	0.627	**0.459**	**0.443**	1.417	0.635	1.026	**0.360**
$(0.5, 0.1)$	CC	0.767	0.771	0.724	0.721	**0.785**	**0.859**	0.495	0.726	0.550	**0.837**
	rCC	0.751	0.763	0.720	0.705	**0.770**	**0.843**	0.601	0.708	0.560	**0.827**
	NRMSE	0.653	0.648	0.749	0.693	**0.625**	**0.588**	1.179	0.779	1.084	**0.549**
$(0.5, 0.3)$	CC	0.602	0.675	0.656	0.642	**0.682**	**0.689**	0.443	0.545	0.483	**0.719**
	rCC	0.650	0.699	0.668	0.646	**0.701**	**0.739**	0.585	0.600	0.496	**0.748**
	NRMSE	0.895	**0.738**	0.792	0.772	**0.746**	0.790	1.830	0.919	1.152	**0.707**
$(0.5, 0.5)$	CC	0.482	**0.587**	0.585	0.562	**0.593**	0.577	0.410	0.436	0.405	**0.622**
	rCC	0.461	**0.570**	0.568	0.537	**0.570**	0.557	0.520	0.417	0.402	**0.600**
	NRMSE	1.060	**0.815**	0.836	0.837	**0.824**	0.868	1.617	1.022	1.216	**0.804**
$(0.5, 0.7)$	CC	0.401	0.533	**0.539**	0.510	**0.536**	0.515	0.364	0.362	0.376	**0.552**
	rCC	0.388	0.514	**0.522**	0.487	**0.515**	0.497	0.470	0.353	0.370	**0.535**
	NRMSE	1.198	**0.856**	**0.861**	0.874	0.874	0.902	1.230	1.130	1.220	**0.865**
$(0.5, 0.9)$	CC	0.346	0.478	**0.497**	0.462	**0.486**	0.459	0.323	0.312	0.340	**0.503**
	rCC	0.337	0.461	**0.479**	0.439	**0.466**	0.443	0.444	0.304	0.336	**0.486**
	NRMSE	1.294	**0.897**	**0.880**	0.905	0.909	0.925	1.140	1.190	1.266	**0.903**

Table 5.19: The average CCs, rCCs, and NRMSEs of 10 methods for EBV prediction on 10 types of simulation datasets defined by (α, β), where α denotes the dominance model and β denotes the background noise ratio.

We also examined the effects of dominance model and background noise level on the EBV prediction. As BLUP is the best method according to its average performance on the simulation datasets, we use BLUP to observe the caused effects.

Essentially, for the first five types without background noise, we simulated 100 genotype datasets and then applied the dominance models to assign cattle their EBVs. Afterwards, we ran BLUP on the datasets for EBV prediction, to collect for each genotype dataset the five CCs. These CCs are plotted in Figure 5.22, where the horizontal lines are the average values of $0.928, 0.891, 0.801, 0.802$, and 0.552 for co-dominant, partially dominant, completely dominant, partially recessive and completely recessive, respectively. It is clear that BLUP worked quite well on co-dominance and partial-

dominance models, while not ideally on the completely recessive model. Another observation is that the algorithm performance on the co-dominant model and the partially dominant model are quite close, and the algorithm performance on the completely dominant model and the partially recessive model are quite close.

A similar experiment was set up to examine the effect of background noise level, where the dominance model was fixed at co-dominance. Again, BLUP was run on each of the 100 genotype datasets of six levels of background noise, at $0, 0.1, 0.3, 0.5, 0.7$ and 0.9 respectively, for EBV prediction to collect the six CCs. These CCs are plotted in Figure 5.23, where the horizontal lines are the average values of $0.928, 0.827, 0.701, 0.600, 0.535$, and 0.486 for $\beta = 0, 0.1, 0.3, 0.5, 0.7$, and 0.9, respectively. We can see the algorithm performance decrease as the background noise level increases, as one can expect.

Effects of Dominance Model

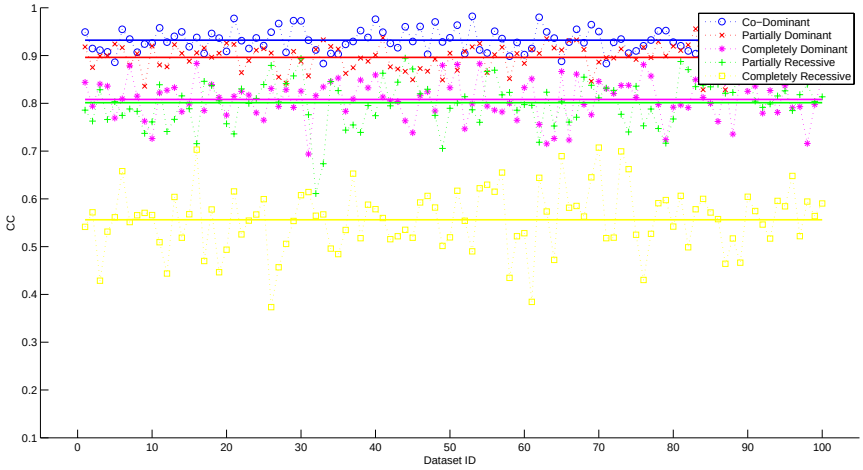

Figure 5.22: The CCs of BLUP EBV prediction on 100 simulation datasets from different dominance models. The average CCs are $0.928, 0.891, 0.801, 0.802$, and 0.552 for co-dominant, partially dominant, completely dominant, partially recessive and completely recessive, respectively.

Effects of Noise Level

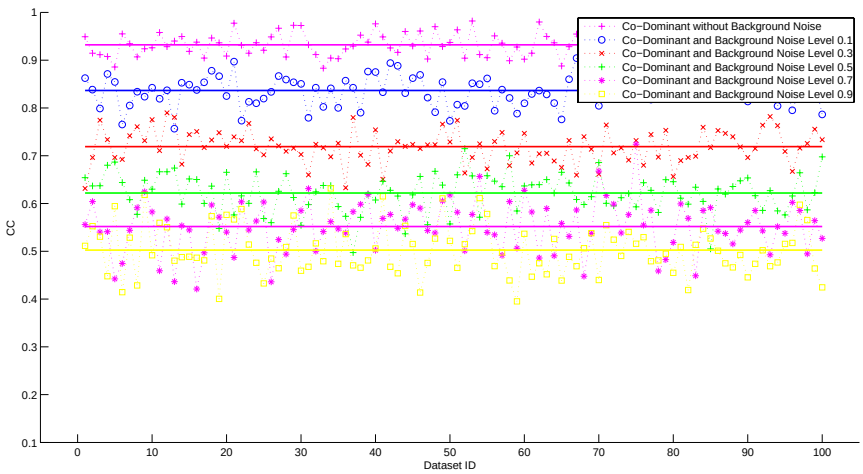

Figure 5.23: The CCs of BLUP EBV prediction on 100 simulation datasets with different levels of background noise. The average CCs are $0.928, 0.827, 0.701, 0.600, 0.535$, and 0.486 for $\beta = 0, 0.1, 0.3, 0.5, 0.7$, and 0.9, respectively.

5.2.6 Co-dominance Representation *vs.* Binary Representation

We examined the effect of the SNP genotype encoding scheme on the EBV prediction for the two real datasets. Since SVM with a linear kernel performed quite well, we used it for examining the effect of the encoding scheme. The results are summarized in Table 5.20, where one can see that there does not seem to be much difference between the schemes for the dairy dataset. However, binary representation generally performs better than co-dominant representation on the beef dataset.

We also examined the effect of SNP genotype encoding scheme on EBV prediction using BLUP and the simulated datasets. The average CCs of BLUP on the 100 simulation datasets for all 10 types are summarized in Table 5.21, where one can see that the co-dominance representation performed the best in all cases except the completely recessive and the completely dominant models (without background noise). There could be multiple reasons for this phenomenon, one of which is that BLUP internally assumes the co-dominance model for EBV regression. Nevertheless, the two exceptional cases suggest that the co-dominance representation should be used with caution, although the binary representation might be helpful in some rare cases.

Comparing the results on the two real datasets with the results on the simulation datasets, one conclusion is that probably none of the eight traits in the two real datasets follows the extreme cases of the completely recessive or the completely dominant models. A tentative conclusion is that using either representation is fine, and thus the co-dominance representation is preferred as it is easier to use.

Real Datasets

Trait	Co-dominance Representation			Binary Representation		
	CC	rCC	NRMSE	CC	rCC	NRMSE
FY	0.401	0.393	0.991	**0.450**	**0.478**	0.977
FP	**0.450**	**0.462**	0.955	0.380	0.366	**0.918**
MY	**0.506**	**0.420**	**0.872**	0.409	0.389	0.928
PY	0.401	0.322	1.031	**0.408**	**0.398**	**0.916**
PP	**0.450**	**0.504**	**0.890**	0.421	0.380	0.902
ADG	0.504	**0.470**	0.879	**0.527**	0.459	0.854
BW	0.541	0.512	**0.852**	**0.562**	**0.523**	0.889
RFI	0.701	**0.646**	**0.713**	**0.730**	0.611	0.724

Table 5.20: EBV prediction results on all 8 traits from the two real datasets by SVM with a linear kernel, using two SNP genotype encoding schemes. Bold text indicates the encoding scheme with better performance.

Simulation Datasets

(α, β)	Co-dominance Representation			Binary Representation		
	CC	rCC	NRMSE	CC	rCC	NRMSE
$(0, 0)$	0.556	0.552	0.879	**0.833**	**0.824**	**0.521**
$(0.25, 0)$	**0.801**	**0.801**	**0.601**	0.782	0.783	0.751
$(0.75, 0)$	**0.896**	**0.891**	**0.443**	0.831	0.827	0.635
$(1, 0)$	0.808	0.802	0.599	**0.893**	**0.887**	**0.495**
$(0.5, 0)$	**0.932**	**0.928**	**0.360**	0.769	0.765	0.795
$(0.5, 0.1)$	**0.837**	**0.827**	**0.549**	0.684	0.676	0.791
$(0.5, 0.3)$	**0.719**	**0.748**	**0.707**	0.588	0.574	0.998
$(0.5, 0.5)$	**0.622**	**0.600**	**0.804**	0.492	0.473	1.211
$(0.5, 0.7)$	**0.552**	**0.535**	**0.865**	0.448	0.429	1.243
$(0.5, 0.9)$	**0.503**	**0.486**	**0.903**	0.397	0.383	1.408

Table 5.21: EBV prediction results on all 10 types of simulation datasets by BLUP, using two SNP genotype encoding schemes. Bold text indicates the encoding scheme with better performance.

5.2.7 Experimental Results of the Bagging EB Feature Selection Method

Real Datasets

On the two real datasets, we applied again the bagging EB method for selecting SNPs of significant effects. Two other feature selection methods, M5 and correlation-based, were also used. We ran two EBV prediction methods, LR and SVM with a linear kernel, to examine the effect of SNP selection. For ease of presentation, "LR + Bagging EB" is denoted as LR^b, and similarly "LR + M5" ("LR + correlation-based") is denoted as LR^m (LR^c, respectively). Table 5.22 contains all the detailed performance results for both LR and SVM, with or without the SNP selection. One can see that none of the three feature selection methods improved the results much, though occasionally marginal improvements were seen. However, for the 3 traits in the beef dataset, the algorithm accuracy does not decrease much after using the feature selection method, but the number of SNPs used in the regression model decreases substantially from the original 5000 SNPs to between 50 and 100 SNPs,

which is beneficial for us as we can detect the location of actual genes which affect the traits more precisely, or design assays that are less expensive.

Trait	Measure	SVM	SVM^b	SVM^c	SVM^m	LR	LR^b	LR^c	LR^m
FY	CC	**0.463**	0.420	0.431	0.276	0.401	0.371	**0.423**	0.211
	rCC	**0.465**	0.403	0.429	0.294	0.383	0.366	**0.418**	0.209
	NRMSE	0.941	**0.928**	0.928	0.968	1.027	1.024	**0.926**	1.013
FP	CC	0.388	**0.451**	0.334	0.293	0.334	**0.453**	0.304	0.289
	rCC	0.376	**0.437**	0.326	0.213	0.317	**0.440**	0.303	0.290
	NRMSE	0.995	**0.938**	0.983	1.017	1.042	**0.958**	1.001	1.046
MY	CC	**0.475**	0.263	0.326	0.201	**0.482**	0.277	0.355	0.239
	rCC	**0.423**	0.331	0.353	0.217	**0.475**	0.333	0.343	0.245
	NRMSE	**0.897**	0.985	0.988	0.989	**0.911**	0.953	0.972	1.019
PY	CC	**0.434**	0.354	0.261	0.248	**0.417**	0.279	0.293	0.217
	rCC	**0.434**	0.334	0.279	0.243	**0.391**	0.277	0.287	0.218
	NRMSE	0.952	**0.941**	0.953	0.981	**0.967**	1.060	0.999	1.042
PP	CC	**0.409**	0.181	0.200	0.193	**0.410**	0.200	0.213	0.127
	rCC	**0.371**	0.192	0.195	0.187	**0.373**	0.187	0.210	0.131
	NRMSE	**0.993**	1.122	1.073	1.157	**1.010**	1.132	1.055	1.150
ADG	CC	**0.504**	0.354	0.377	0.416	**0.461**	0.356	0.393	0.306
	rCC	**0.470**	0.310	0.344	0.377	**0.471**	0.319	0.356	0.304
	NRMSE	**0.879**	1.034	1.105	0.941	**0.943**	1.067	0.997	1.308
BW	CC	**0.541**	0.469	0.424	0.460	**0.483**	0.426	0.375	0.340
	rCC	**0.512**	0.425	0.477	0.416	0.380	**0.411**	0.331	0.321
	NRMSE	**0.852**	0.887	0.947	0.945	**0.888**	1.001	1.017	1.278
RFI	CC	**0.701**	0.657	0.624	0.691	**0.620**	0.555	0.538	0.515
	rCC	**0.646**	0.544	0.510	0.568	**0.505**	0.458	0.427	0.465
	NRMSE	**0.713**	0.770	0.807	0.728	**0.854**	0.887	0.893	1.056

Table 5.22: The average CCs, rCCs, and NRMSEs of SVM and LR, with or without using a SNP selection method, for EBV prediction on 8 traits of the two real datasets.

Simulation Datasets

EB is a shrinkage method that can be used for identifying SNPs having significant effects on the target trait. For each of the 100 datasets for one type of simulation, we used the bagging EB method to select SNPs in the training dataset, and then we ran the LR method to fit an EBV predictor and tested it on the testing dataset. Using the 10-fold cross validation scheme, we collected the performance statistics of such an approach, called "LR + Bagging EB". For comparison, we also collected the performance statistics for LR on the 100 datasets. Table 5.23 lists the detailed performance results for both LR and "LR + Bagging EB", where one can see that the bagging EB method did improve the EBV prediction. For example, even for the simulation type $(0.5, 0.9)$, i.e. co-dominant model with background noise level 0.9, the bagging EB method improved the LR EBV prediction performance more than 20%; for the co-dominant model without background noise, the LR EBV prediction reached a CC 1, which is 40% higher than using the entire set of SNPs.

- completely recessive ($\alpha = 0, \beta = 0$)

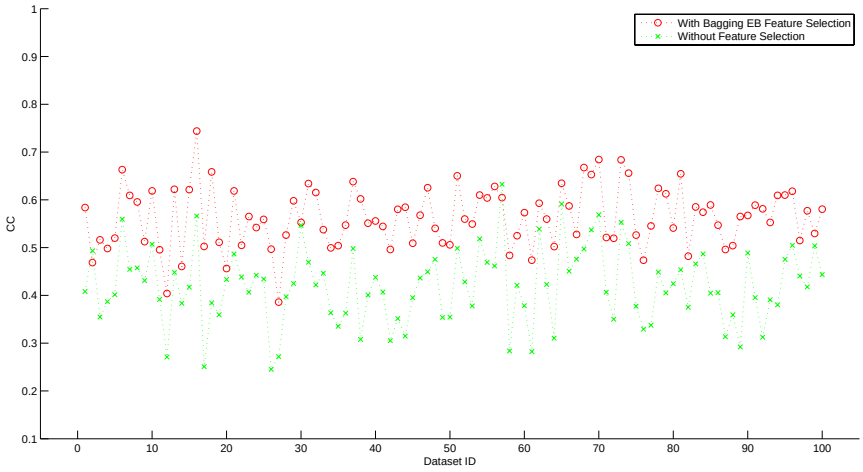

Figure 5.24: Algorithm performance with and without bagging EB feature selection method for the completely recessive model by CC measurement.

- partially recessive ($\alpha = 0.25, \beta = 0$)

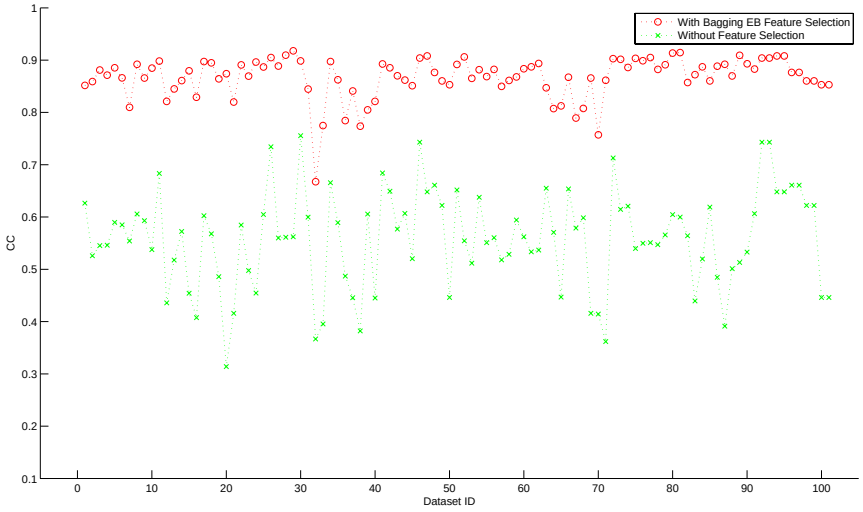

Figure 5.25: Algorithm performance with and without bagging EB feature selection method for the partially recessive model by CC measurement.

- co-dominant ($\alpha = 0.5, \beta = 0$)

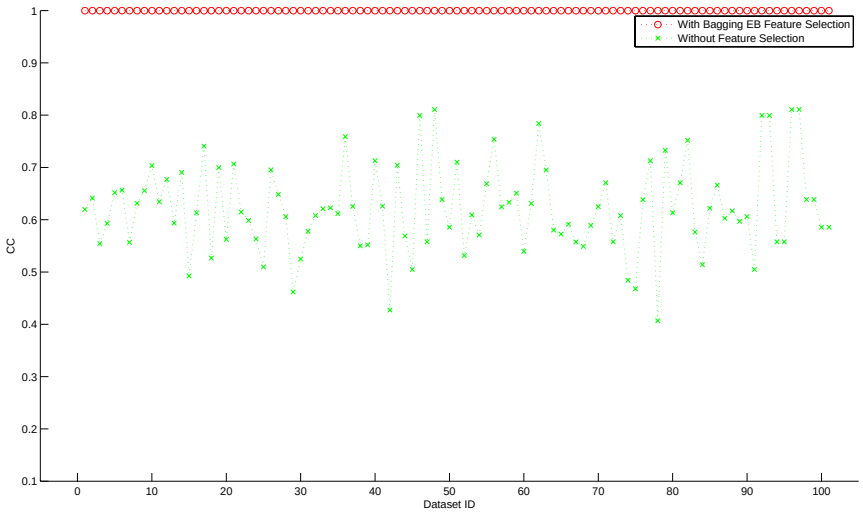

Figure 5.26: Algorithm performance with and without bagging EB feature selection method for the co-dominant model by CC measurement.

- partially dominant ($\alpha = 0.75, \beta = 0$)

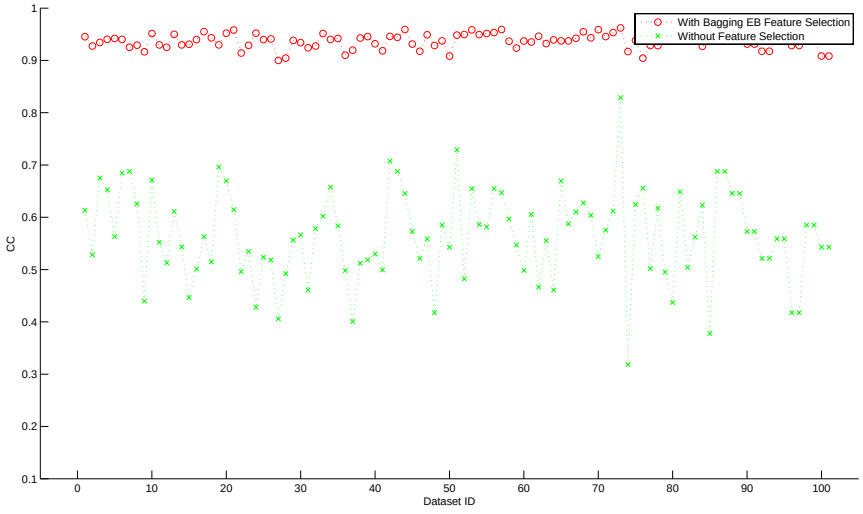

Figure 5.27: Algorithm performance with and without bagging EB feature selection method for the partially dominant model by CC measurement.

- completely dominant ($\alpha = 1, \beta = 0$)

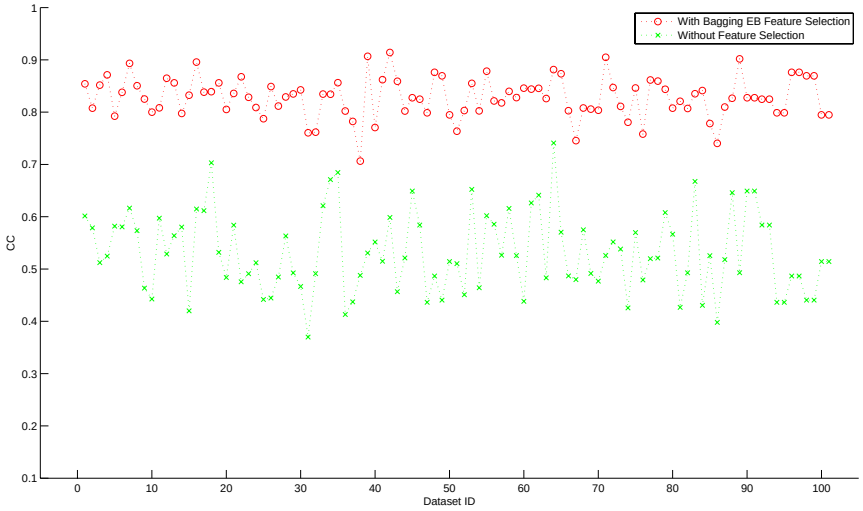

Figure 5.28: Algorithm performance with and without bagging EB feature selection method for the completely dominant model by CC measurement.

- co-dominant model with background noise level 0.1 ($\alpha = 0.5, \beta = 0.1$)

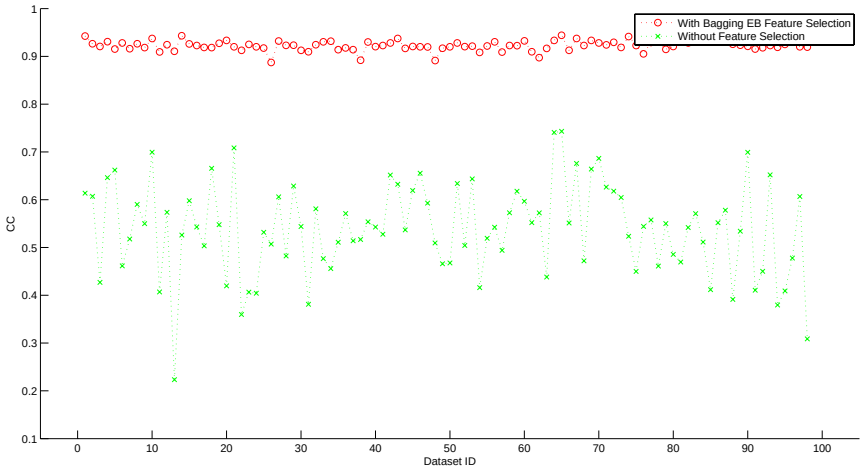

Figure 5.29: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.1 by CC measurement.

- co-dominant model with background noise level 0.3 ($\alpha = 0.5, \beta = 0.3$)

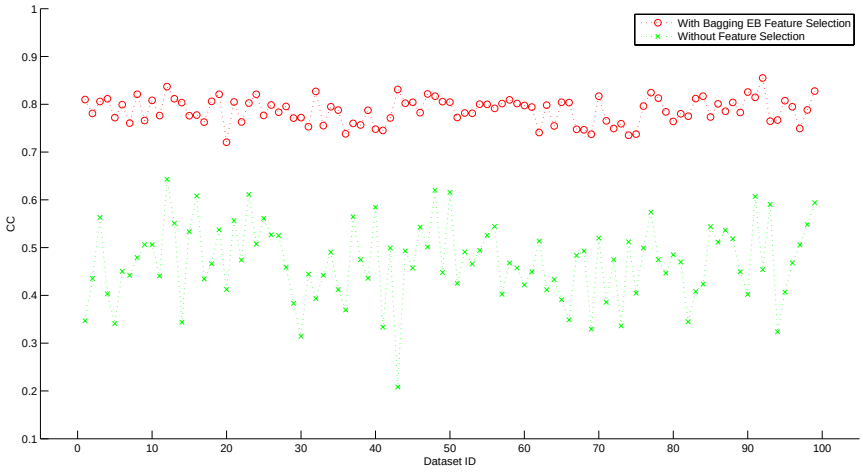

Figure 5.30: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.3 by CC measurement.

- co-dominant model with background noise level 0.5 ($\alpha = 0.5, \beta = 0.5$)

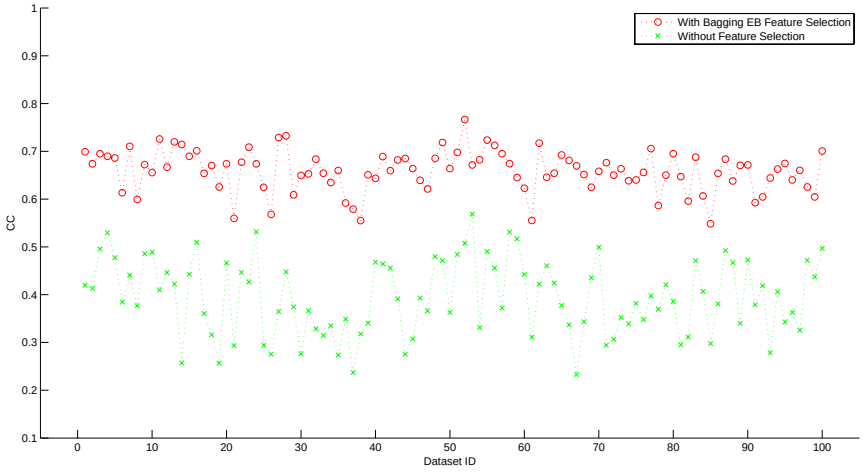

Figure 5.31: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.5 by CC measurement.

- co-dominant model with background noise level 0.7 ($\alpha = 0.5, \beta = 0.7$)

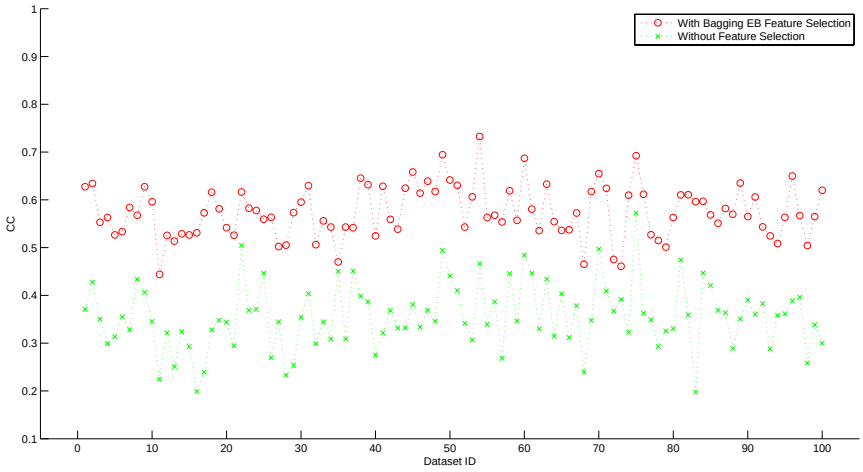

Figure 5.32: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.7 by CC measurement.

- co-dominant model with background noise level 0.9 ($\alpha = 0.5, \beta = 0.9$)

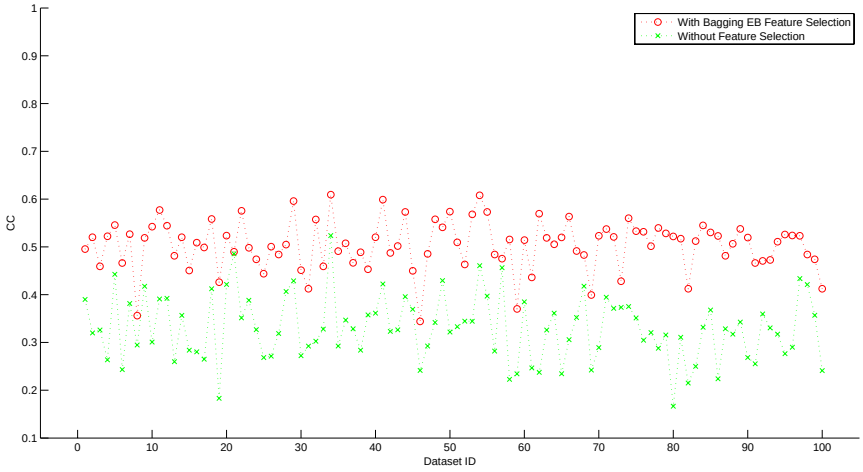

Figure 5.33: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.9 by CC measurement.

(α, β)	LR			LR + Bagging EB		
	CC	rCC	NRMSE	CC	rCC	NRMSE
$(0, 0)$	0.420	0.425	1.103	**0.563**	**0.556**	**0.888**
$(0.25, 0)$	0.554	0.571	1.006	**0.865**	**0.861**	**0.512**
$(0.75, 0)$	0.565	0.585	1.126	**0.938**	**0.932**	**0.351**
$(1, 0)$	0.533	0.551	1.163	**0.827**	**0.820**	**0.576**
$(0.5, 0)$	0.618	0.638	1.023	**1.000**	**1.000**	**0.000**
$(0.5, 0.1)$	0.538	0.551	1.103	**0.923**	**0.915**	**0.390**
$(0.5, 0.3)$	0.469	0.474	1.168	**0.787**	**0.771**	**0.629**
$(0.5, 0.5)$	0.394	0.392	1.229	**0.659**	**0.638**	**0.777**
$(0.5, 0.7)$	0.356	0.351	1.253	**0.575**	**0.559**	**0.853**
$(0.5, 0.9)$	0.330	0.327	1.282	**0.504**	**0.487**	**0.909**

Table 5.23: EBV prediction results on 10 types of simulation datasets by LR, with or without using the bagging EB method for SNP selection.

5.3 Additional Results

5.3.1 Comparison of Algorithm Performance on Real Datasets

Dairy Dataset

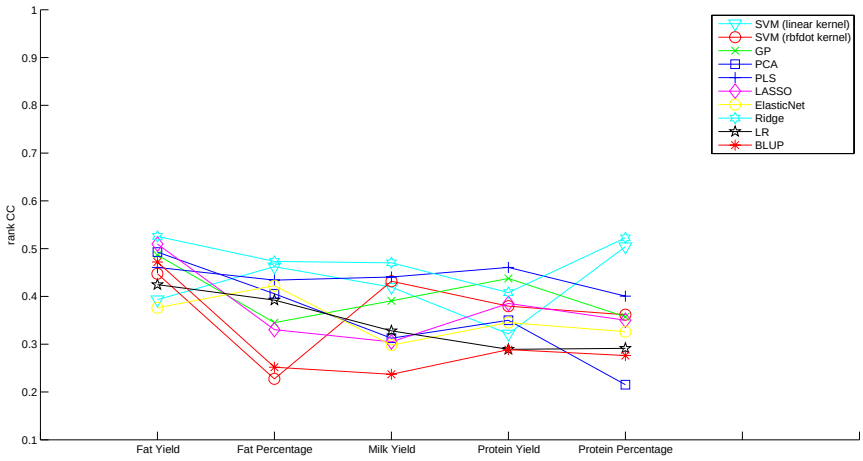

Figure 5.34: Performance of EBV prediction algorithms on dairy dataset by rCC measurement, which shows that for fat yield trait, Ridge is the best method; for fat percentage trait, Ridge is the best method; for milk yield trait, Ridge is the best method; for protein yield trait, PLS is the best method and for protein percentage trait, Ridge is the best method.

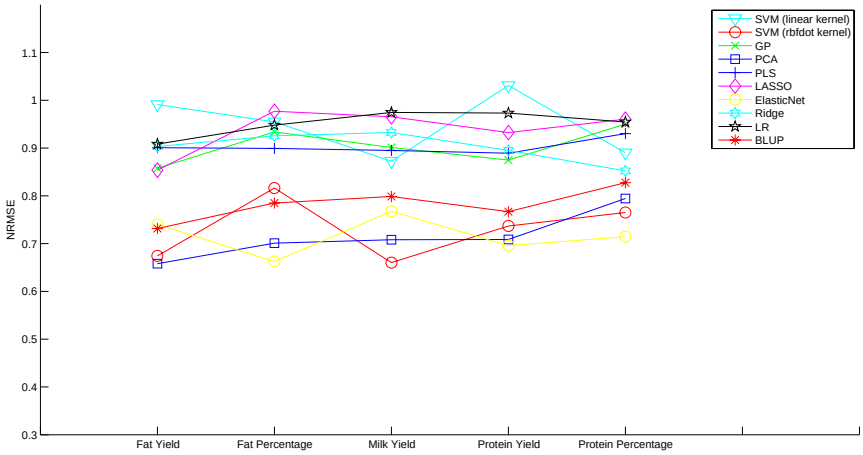

Figure 5.35: Performance of EBV prediction algorithms on dairy dataset by NRMSE measurement, which shows that for fat yield trait, PCA is the best method; for fat percentage trait, ElasticNet is the best method; for milk yield trait, SVM with rbfdot kernel is the best method; for protein yield trait, ElasticNet is the best method and for protein percentage trait, ElasticNet is the best method.

Beef Dataset

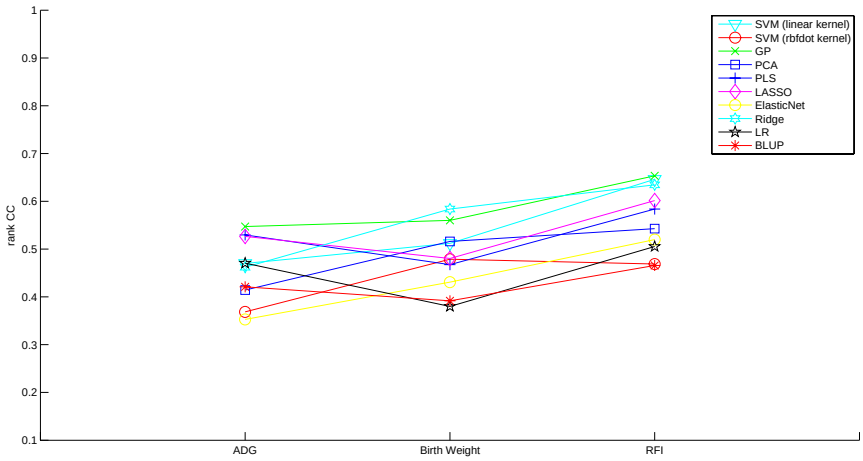

Figure 5.36: Performance of EBV prediction algorithms on beef dataset by rCC measurement, which shows for ADG trait, GP is the best method; for birth weight trait, Ridge is the best method and for RFI trait, GP is the best method.

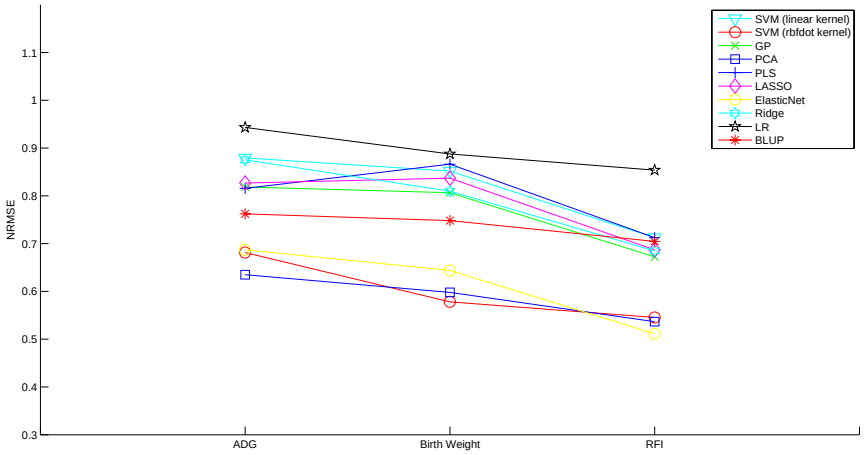

Figure 5.37: Performance of EBV prediction algorithms on beef dataset by NRMSE measurement, which shows that if using this measurement, for ADG trait, PCA is the best method; for birth weight trait, SVM with rbfdot kernel is the best method and for RFI trait, ElasticNet is the best method.

5.3.2 Comparison of Algorithm Performance on Simulation Datasets

In the legend of each of the following figures, the regression methods are sorted by their performance. There are a total of 100 datasets for each simulation model. The horizontal line for each regression method is the average performance of that method on the 100 datasets.

- completely recessive ($\alpha = 0, \beta = 0$)

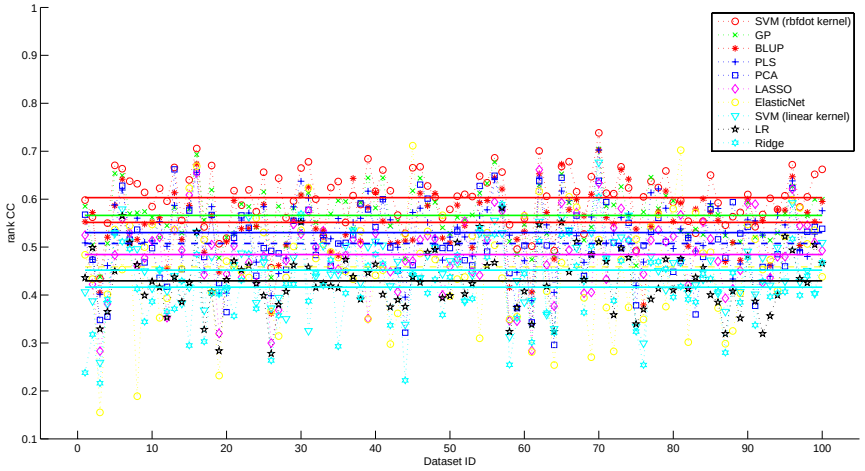

Figure 5.38: Performance of EBV prediction algorithms for the completely recessive model by rCC measurement, SVM (rbfdot kernel), GP and BLUP are the top 3 methods for this model.

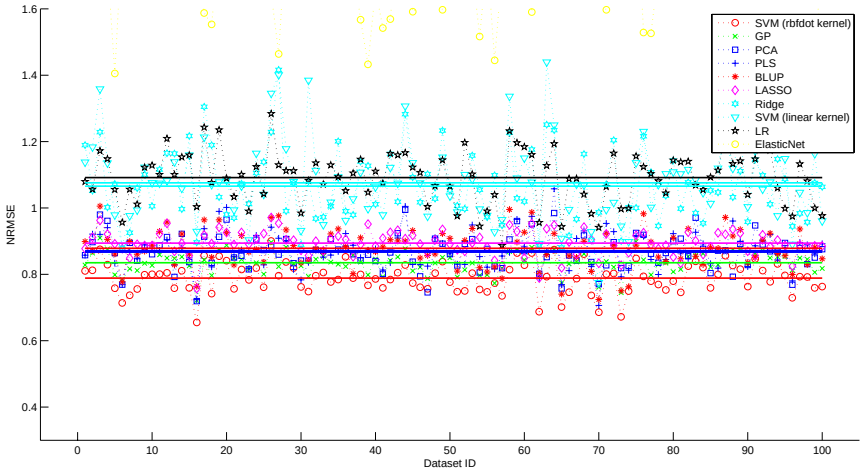

Figure 5.39: Performance of EBV prediction algorithms for the completely recessive model by NRMSE measurement, SVM (rbfdot kernel), GP and PCA are the top 3 methods for this model.

- partially recessive ($\alpha = 0.25, \beta = 0$)

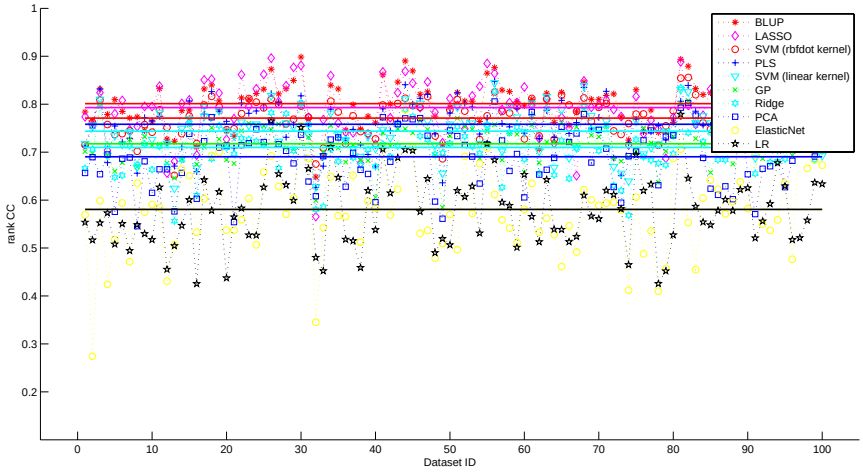

Figure 5.40: Performance of EBV prediction algorithms for the partially recessive model by rCC measurement, BLUP, LASSO and SVM (rbfdot kernel) are the top 3 methods for this model.

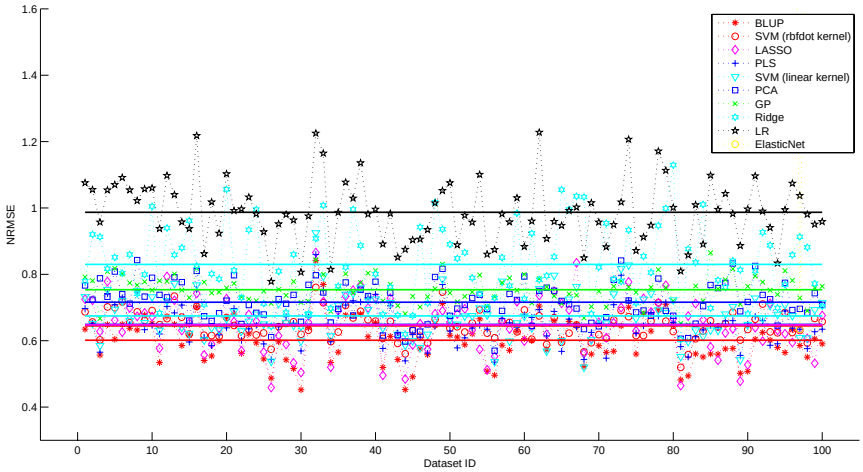

Figure 5.41: Performance of EBV prediction algorithms for the partially recessive model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and LASSO are the top 3 methods for this model.

- co-dominant ($\alpha = 0.5, \beta = 0$)

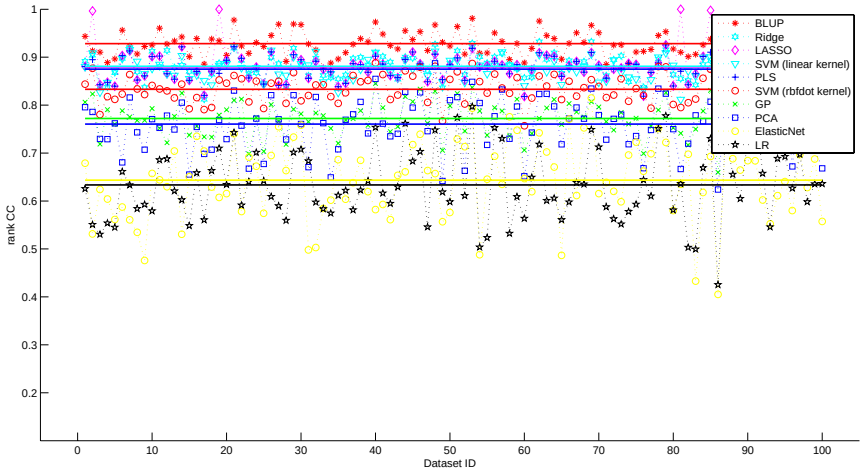

Figure 5.42: Performance of EBV prediction algorithms for the co-dominant model by rCC measurement, BLUP, Ridge and LASSO are the top 3 methods for this model.

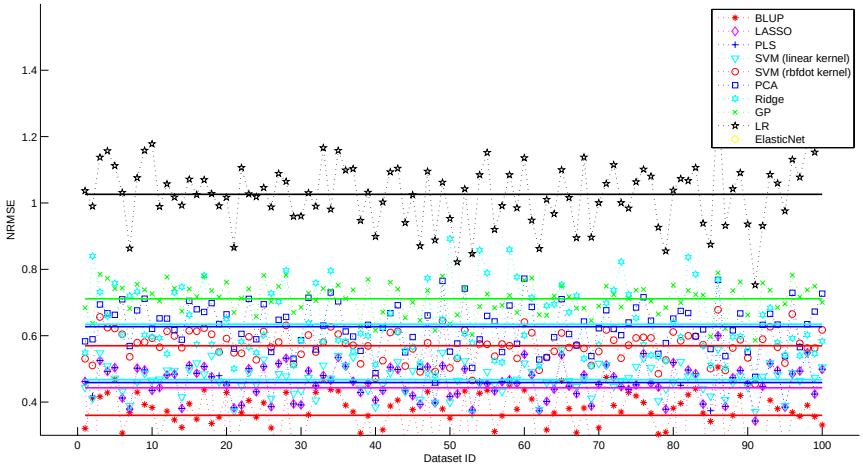

Figure 5.43: Performance of EBV prediction algorithms for the co-dominant model by NRMSE measurement, BLUP, LASSO and PLS are the top 3 methods for this model.

- partially dominant ($\alpha = 0.75, \beta = 0$)

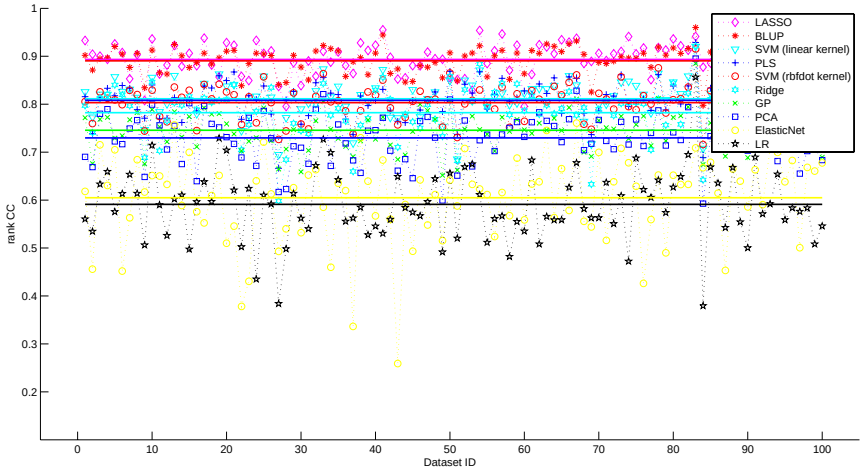

Figure 5.44: Performance of EBV prediction algorithms for the partially dominant model by rCC measurement, LASSO, BLUP and SVM (linear kernel) are the top 3 methods for this model.

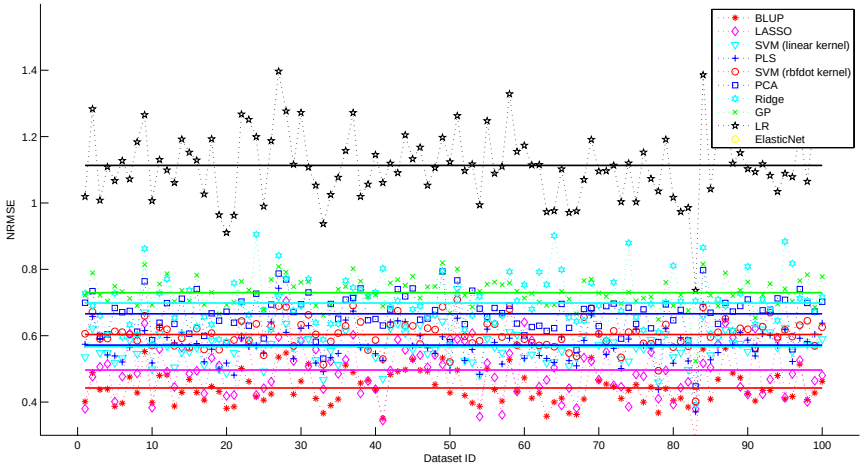

Figure 5.45: Performance of EBV prediction algorithms for the partially dominant model by NRMSE measurement, BLUP, LASSO and SVM (linear kernel) are the top 3 methods for this model.

- completely dominant ($\alpha = 1, \beta = 0$)

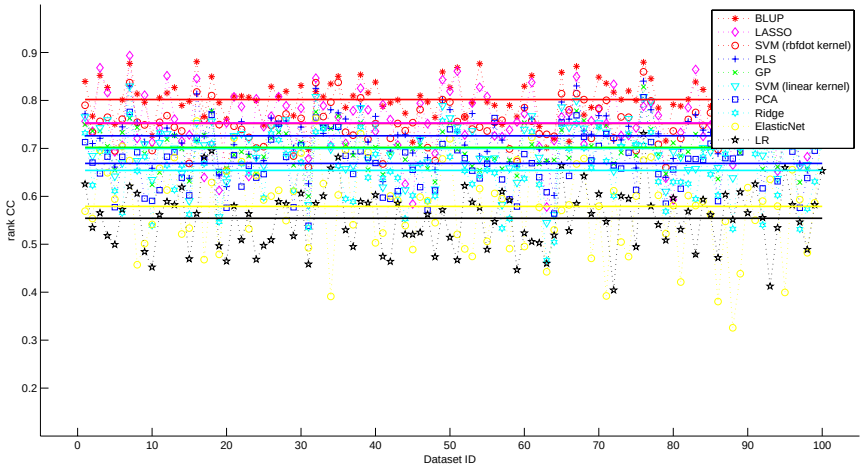

Figure 5.46: Performance of EBV prediction algorithms for the completely dominant model by rCC measurement, BLUP, LASSO and SVM (rbfdot kernel) are the top 3 methods for this model.

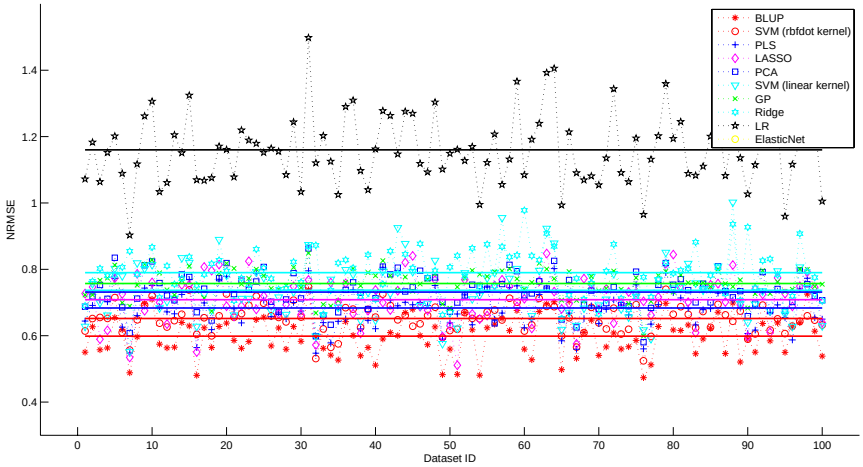

Figure 5.47: Performance of EBV prediction algorithms for the completely dominant model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and LASSO are the top 3 methods for this model.

- co-dominant model with background noise level 0.1 ($\alpha = 0.5, \beta = 0.1$)

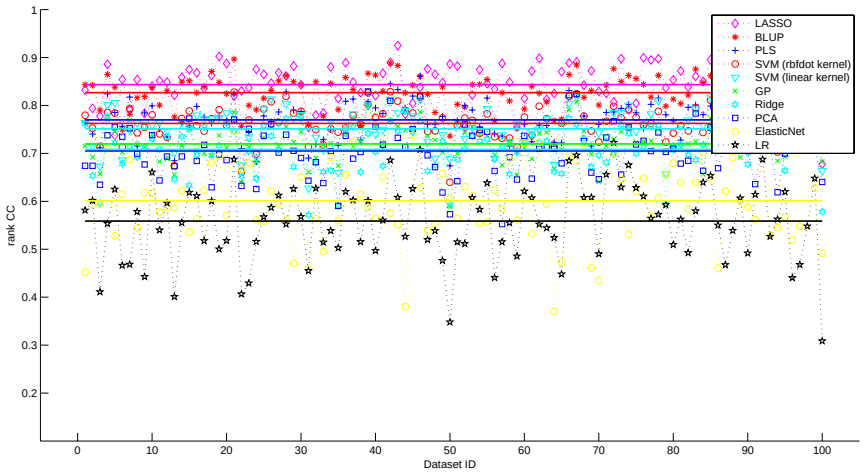

Figure 5.48: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.1 model by rCC measurement, LASSO, BLUP and PLS are the top 3 methods for this model.

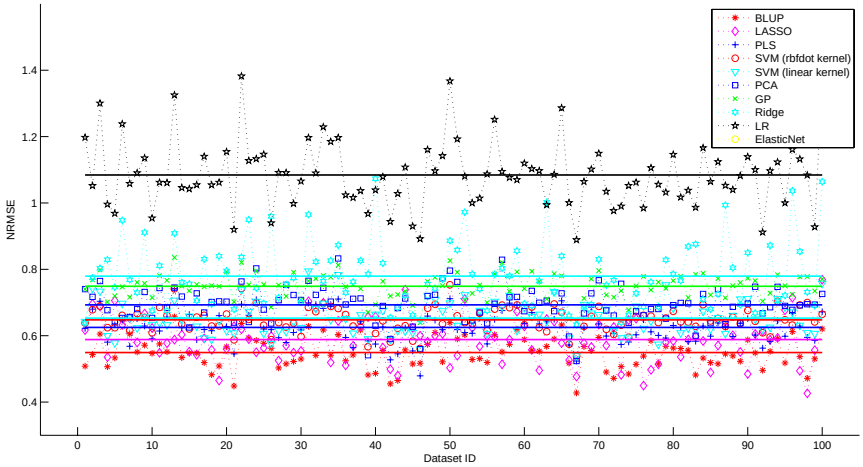

Figure 5.49: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.1 model by NRMSE measurement, BLUP, LASSO and PLS are the top 3 methods for this model.

- co-dominant model with background noise level 0.3 ($\alpha = 0.5, \beta = 0.3$)

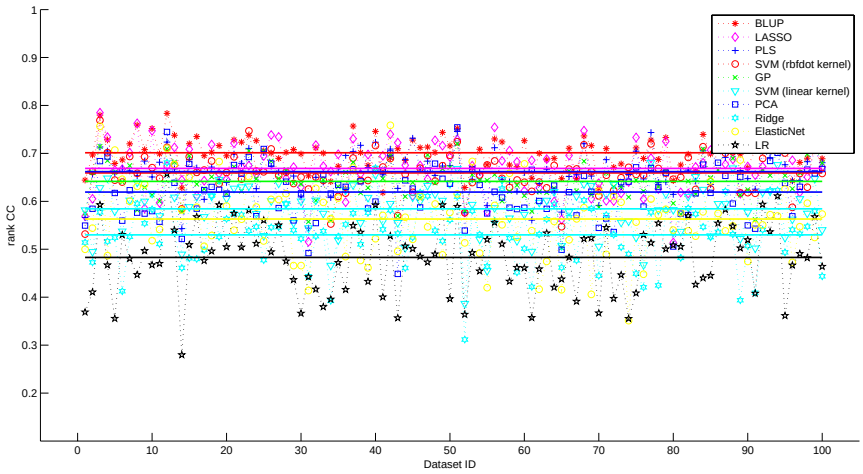

Figure 5.50: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.3 model by rCC measurement, BLUP, LASSO and PLS are the top 3 methods for this model.

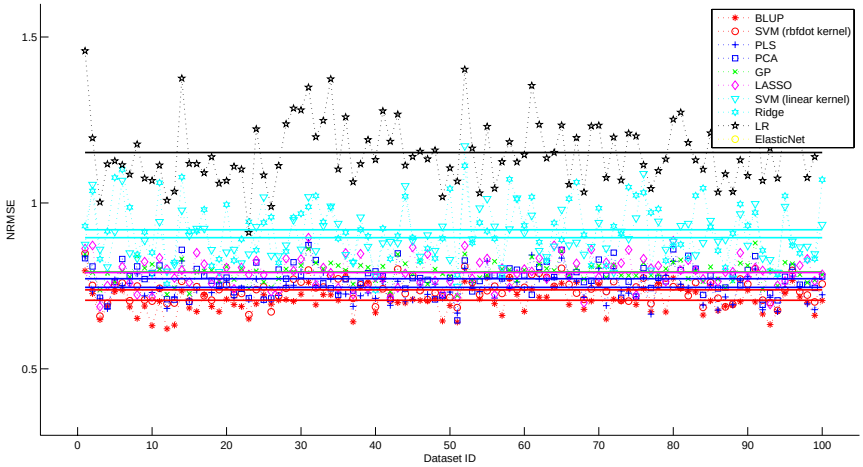

Figure 5.51: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.3 model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and PLS are the top 3 methods for this model.

- co-dominant model with background noise level 0.5 ($\alpha = 0.5, \beta = 0.5$)

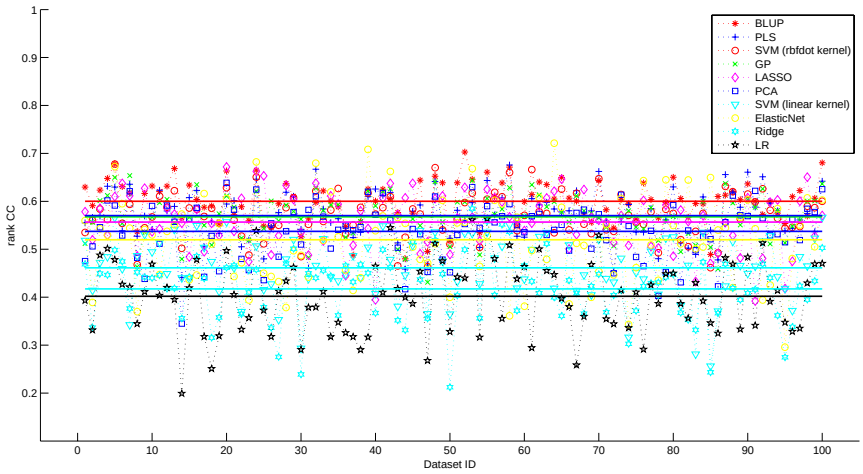

Figure 5.52: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.5 model by rCC measurement, BLUP, PLS and SVM (rbfdot kernel) are the top 3 methods for this model.

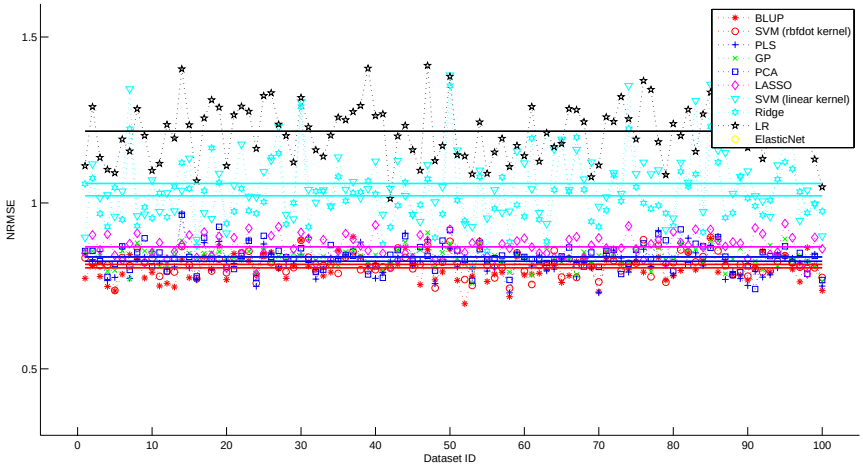

Figure 5.53: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.5 model by NRMSE measurement, BLUP, SVM (rbfdot kernel) and PLS are the top 3 methods for this model.

- co-dominant model with background noise level 0.7 ($\alpha = 0.5, \beta = 0.7$)

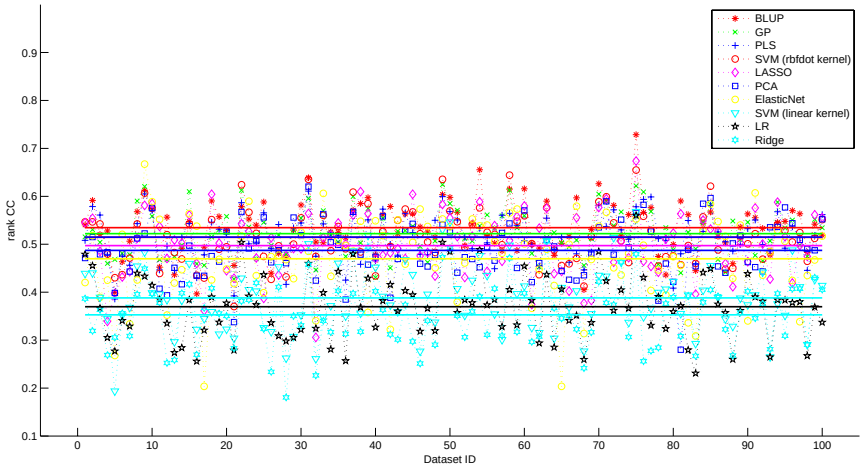

Figure 5.54: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.7 model by NRMSE measurement, BLUP, GP and PLS are the top 3 methods for this model.

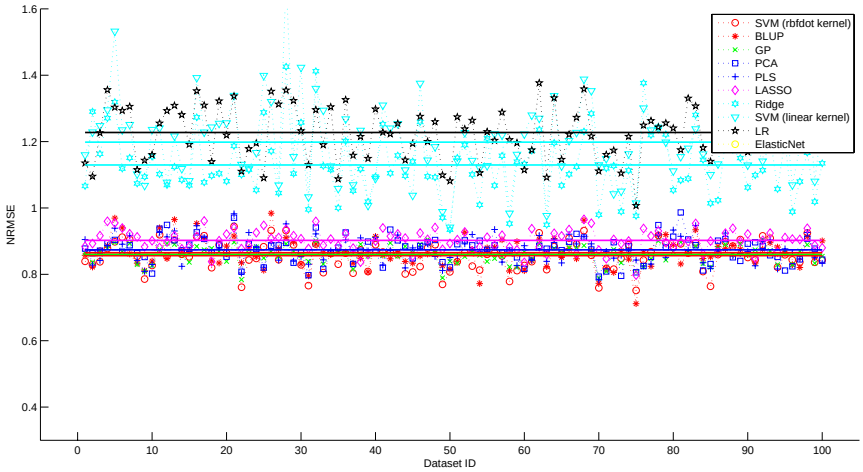

Figure 5.55: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.7 model by NRMSE measurement, SVM (rbfdot kernel), BLUP and GP are the top 3 methods for this model.

- co-dominant model with background noise level 0.9 ($\alpha = 0.5, \beta = 0.9$)

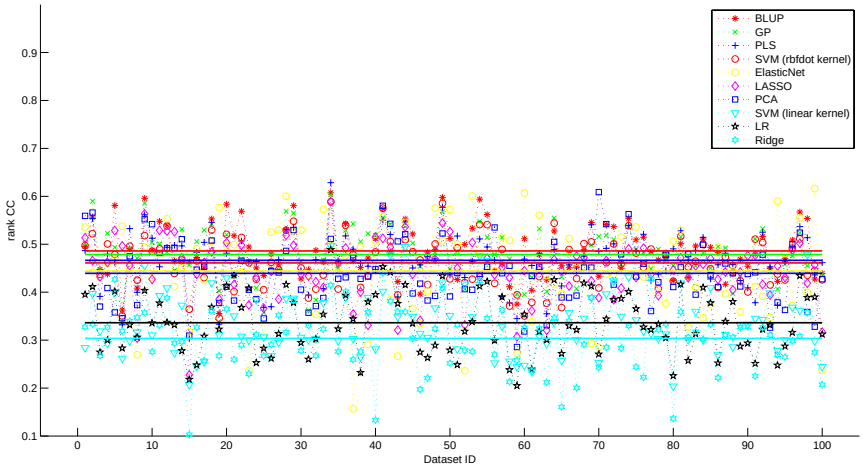

Figure 5.56: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.9 model by NRMSE measurement, BLUP, GP and PLS are the top 3 methods for this model.

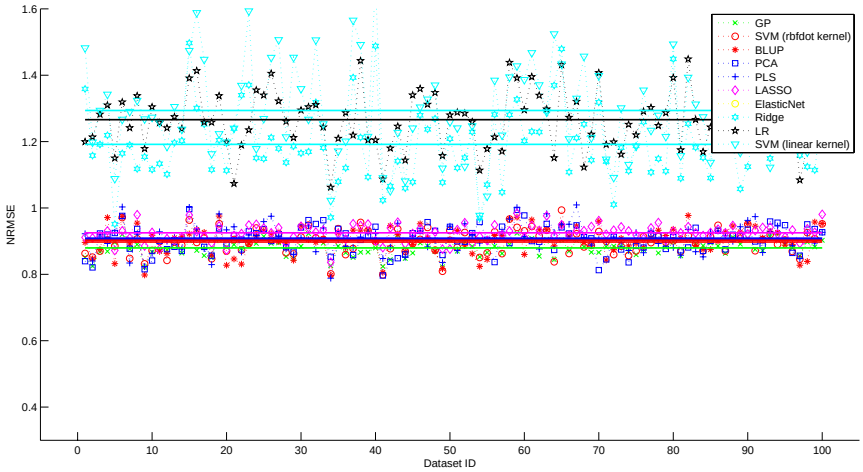

Figure 5.57: Performance of EBV prediction algorithms for the co-dominant model with background noise level 0.9 model by NRMSE measurement, GP, SVM (rbfdot kernel) and BLUP are the top 3 methods for this model.

Effects of Dominance Model

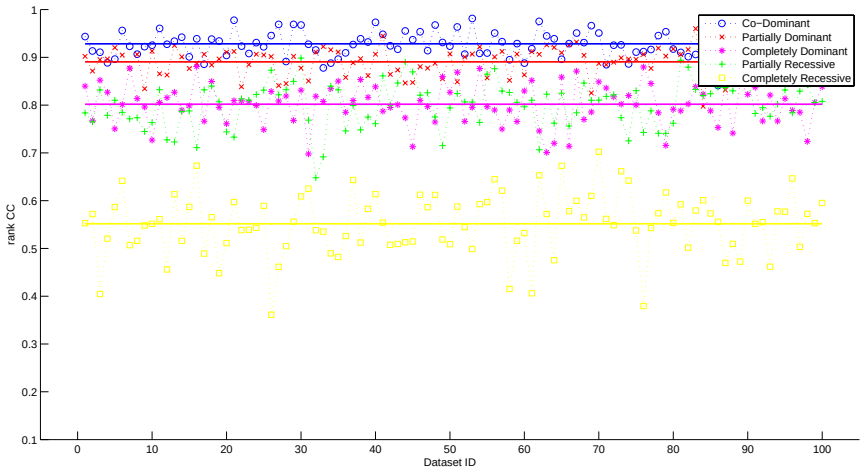

Figure 5.58: Performance of EBV prediction algorithms on 5 dominance models by rCC measurement.

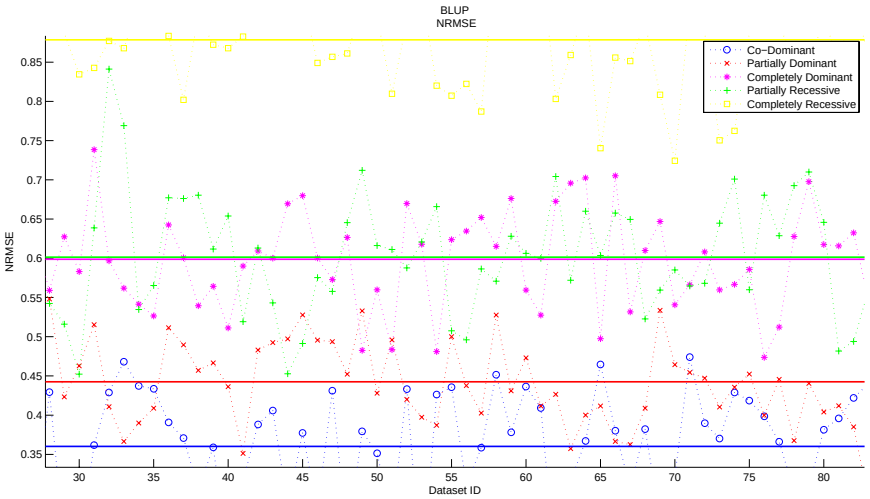

Figure 5.59: Performance of EBV prediction algorithms on 5 dominance models by NRMSE measurement.

Effects of Noise Level

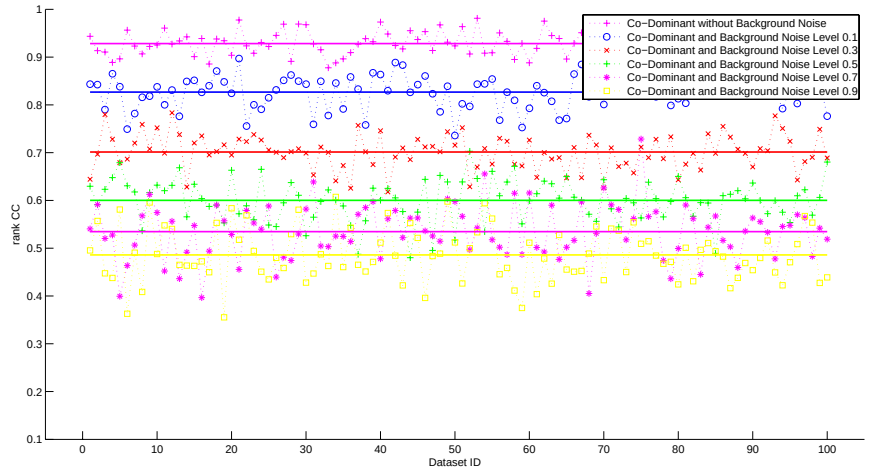

Figure 5.60: Performance of EBV prediction algorithms on co-dominant model with 5 noise levels by rCC measurement.

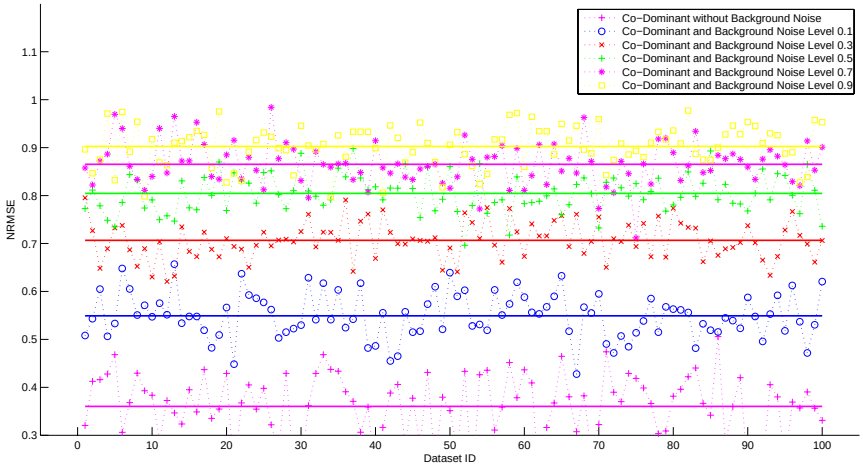

Figure 5.61: Performance of EBV prediction algorithms on co-dominant model with 5 noise levels by NRMSE measurement.

5.3.3 Experimental Results of the Bagging EB Feature Selection Method

Simulation Datasets

- completely recessive ($\alpha = 0, \beta = 0$)

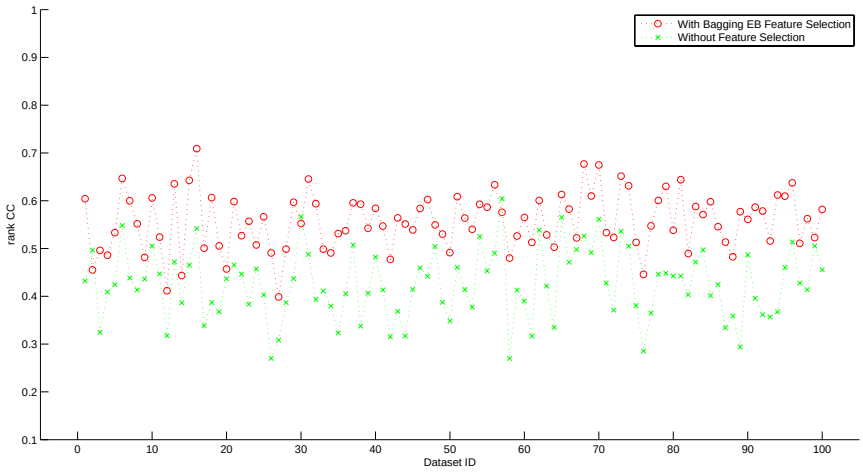

Figure 5.62: Algorithm performance with and without bagging EB feature selection method for the completely recessive model by rCC measurement.

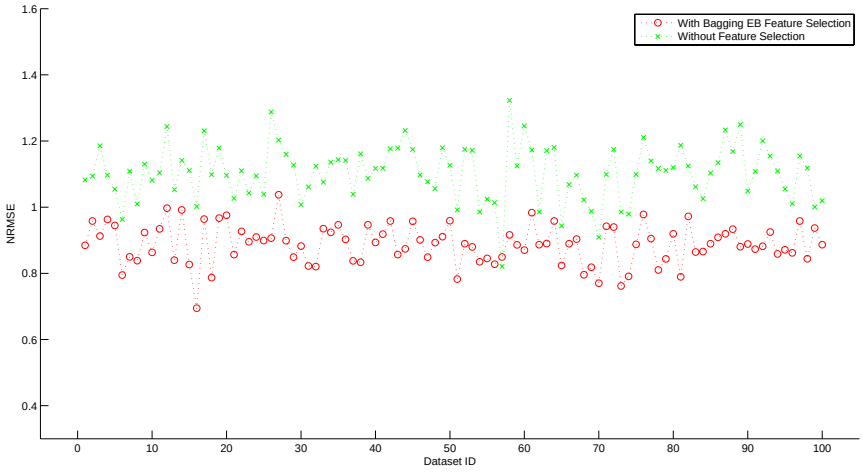

Figure 5.63: Algorithm performance with and without bagging EB feature selection method for the completely recessive model by NRMSE measurement.

- partially recessive ($\alpha = 0.25, \beta = 0$)

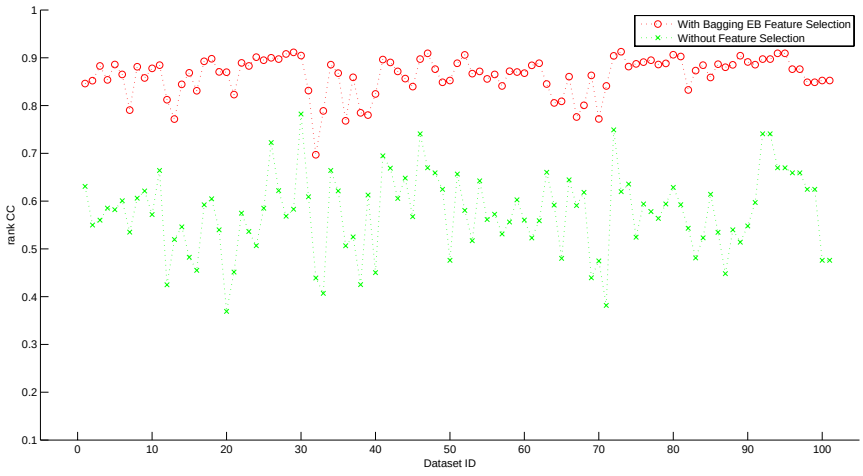

Figure 5.64: Algorithm performance with and without bagging EB feature selection method for the partially recessive model by rCC measurement.

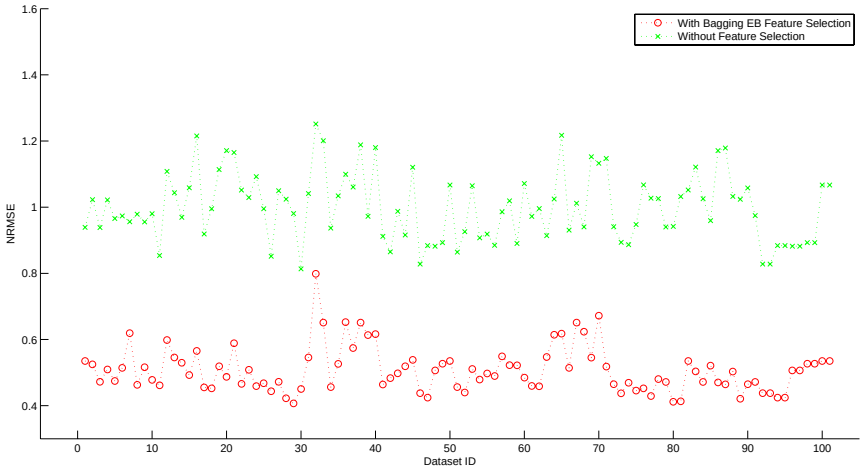

Figure 5.65: Algorithm performance with and without bagging EB feature selection method for the partially recessive model by NRMSE measurement.

- co-dominant ($\alpha = 0.5, \beta = 0$)

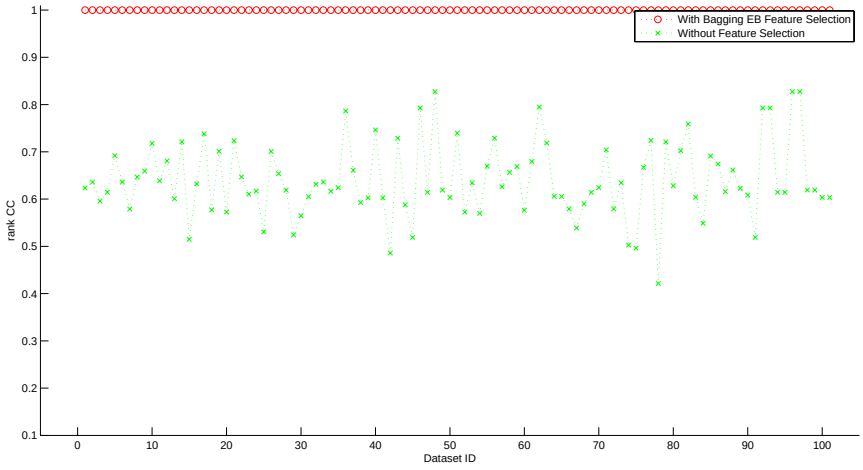

Figure 5.66: Algorithm performance with and without bagging EB feature selection method for the co-dominant model by rCC measurement.

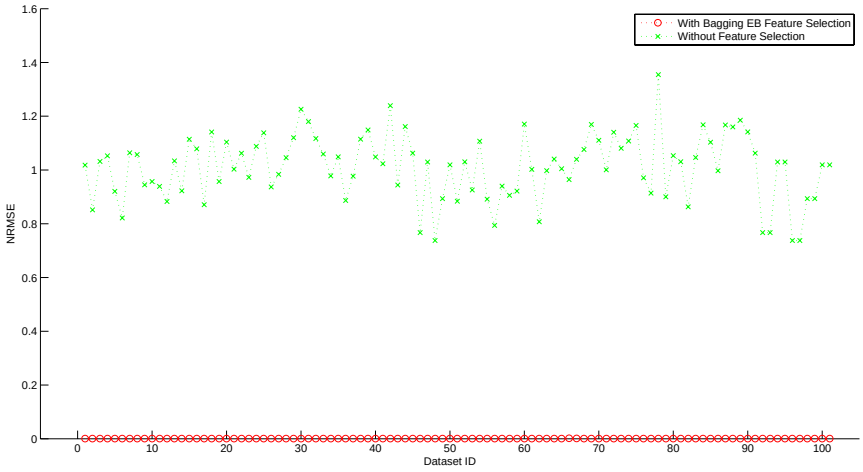

Figure 5.67: Algorithm performance with and without bagging EB feature selection method for the co-dominant model by NRMSE measurement.

- partially dominant ($\alpha = 0.75, \beta = 0$)

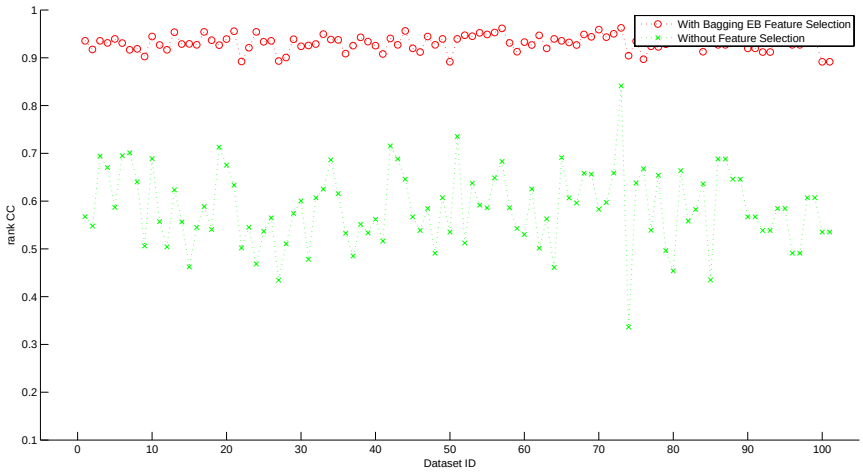

Figure 5.68: Algorithm performance with and without bagging EB feature selection method for the partially dominant model by rCC measurement.

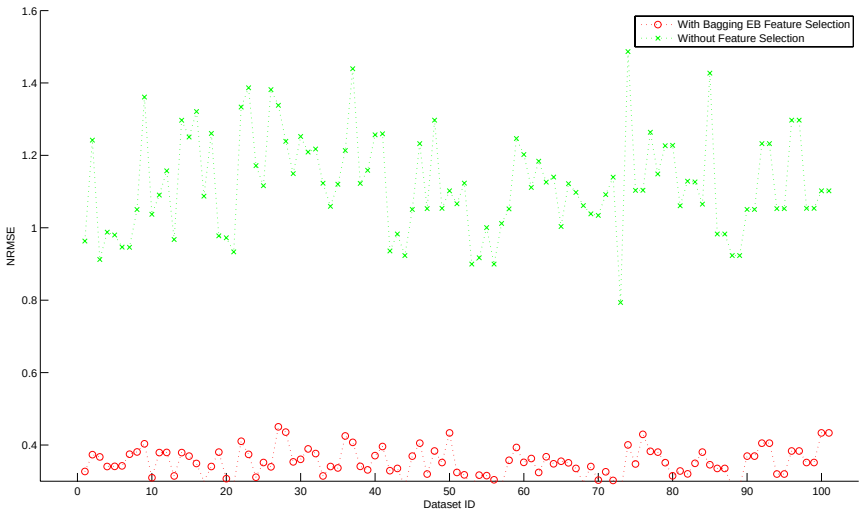

Figure 5.69: Algorithm performance with and without bagging EB feature selection method for the partially dominant model by NRMSE measurement.

- completely dominant ($\alpha = 1, \beta = 0$)

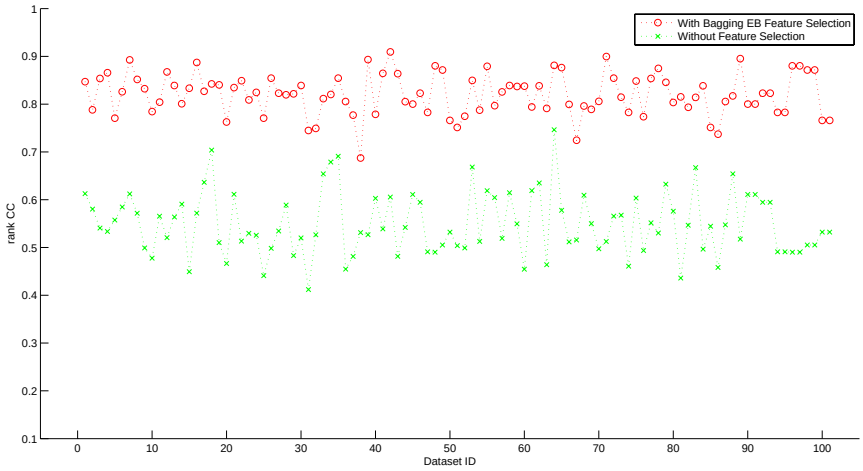

Figure 5.70: Algorithm performance with and without bagging EB feature selection method for the completely dominant model by rCC measurement.

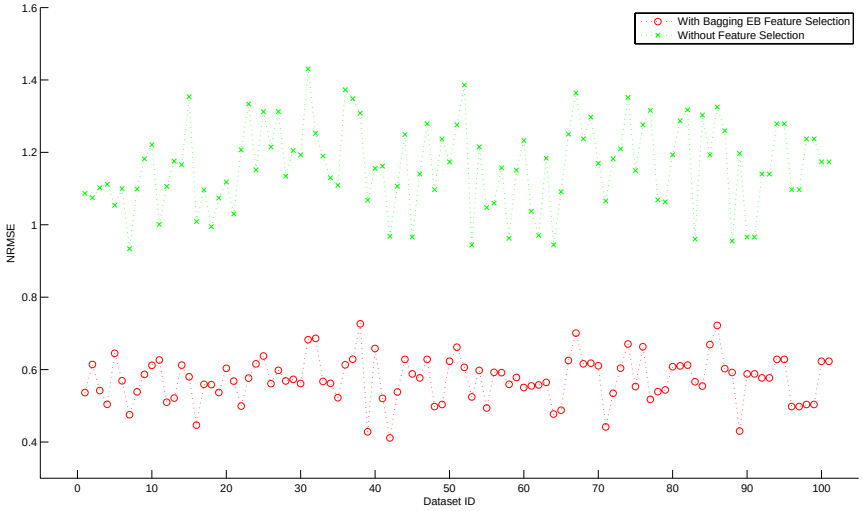

Figure 5.71: Algorithm performance with and without bagging EB feature selection method for the completely dominant model by NRMSE measurement.

- co-dominant model with background noise level 0.1 ($\alpha = 0.5, \beta = 0.1$)

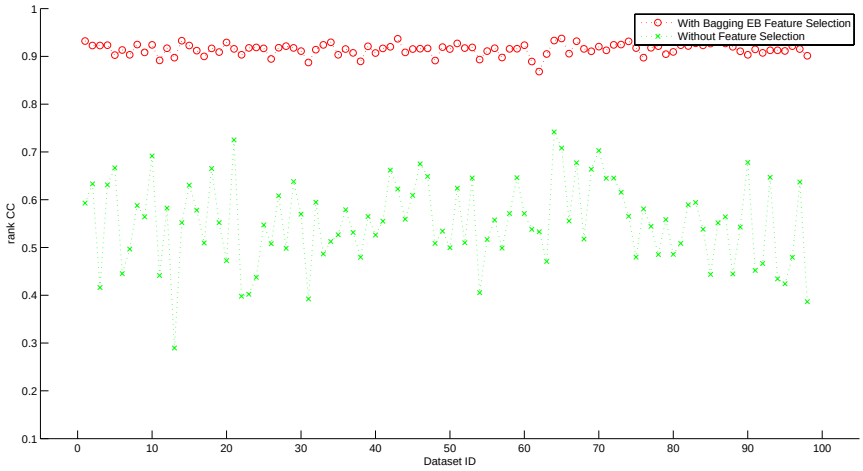

Figure 5.72: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.1 by rCC measurement.

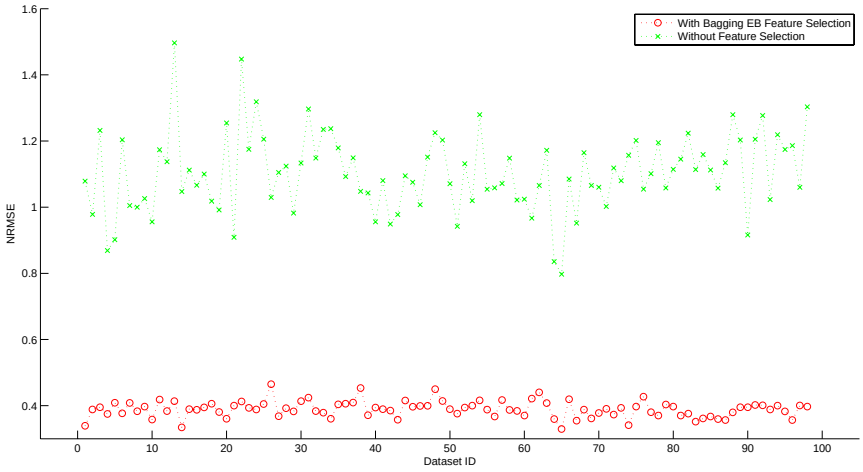

Figure 5.73: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.1 by NRMSE measurement.

- co-dominant model with background noise level 0.3 ($\alpha = 0.5, \beta = 0.3$)

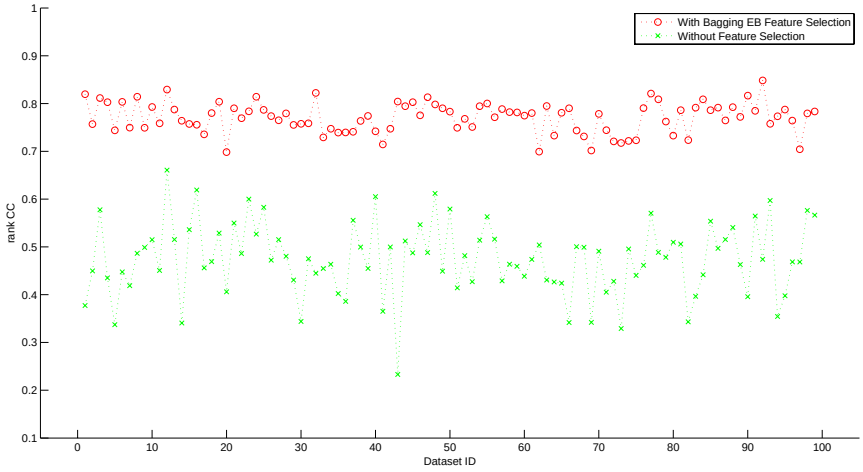

Figure 5.74: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.3 by rCC measurement.

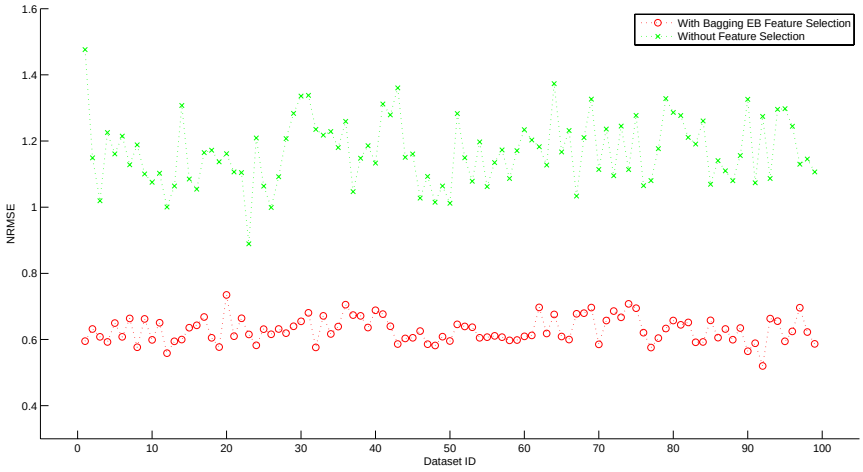

Figure 5.75: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.3 by NRMSE measurement.

- co-dominant model with background noise level 0.5 ($\alpha = 0.5, \beta = 0.5$)

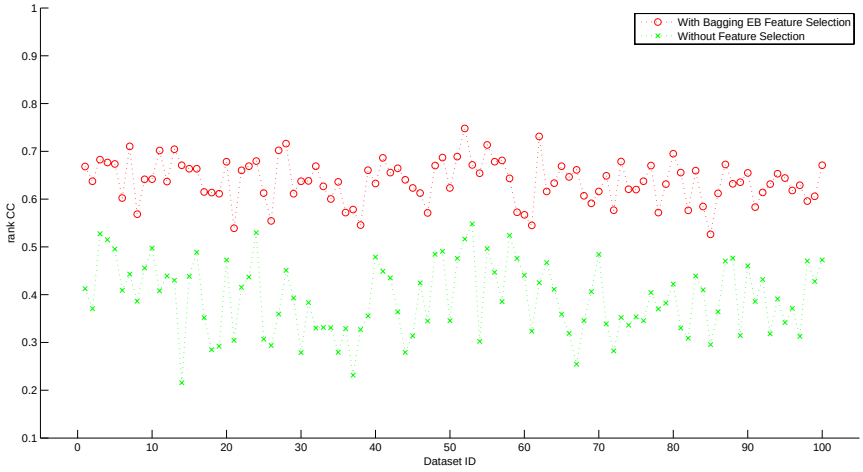

Figure 5.76: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.5 by rCC measurement.

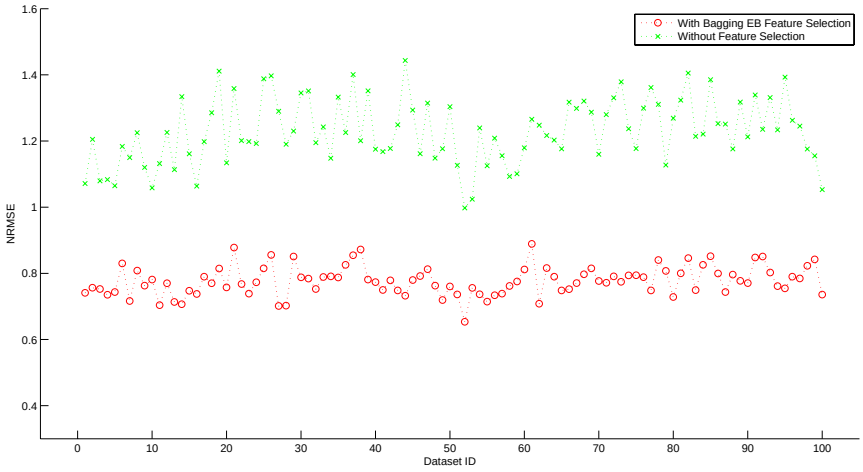

Figure 5.77: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.5 by NRMSE measurement.

- co-dominant model with background noise level 0.7 ($\alpha = 0.5, \beta = 0.7$)

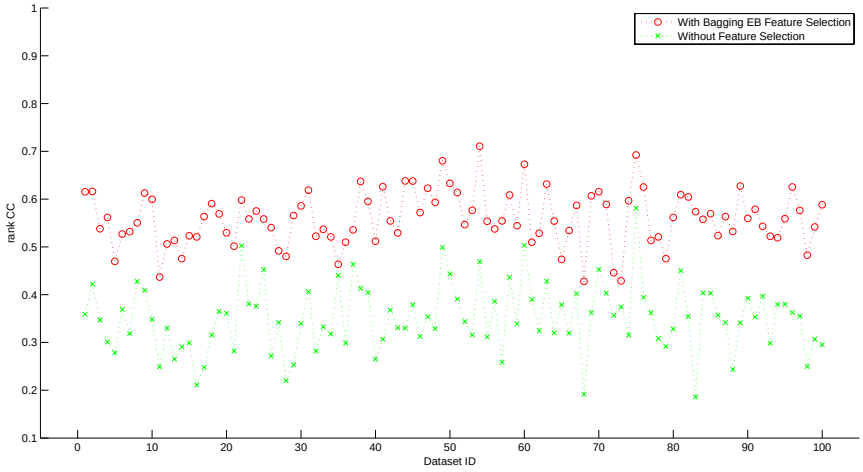

Figure 5.78: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.7 by rCC measurement.

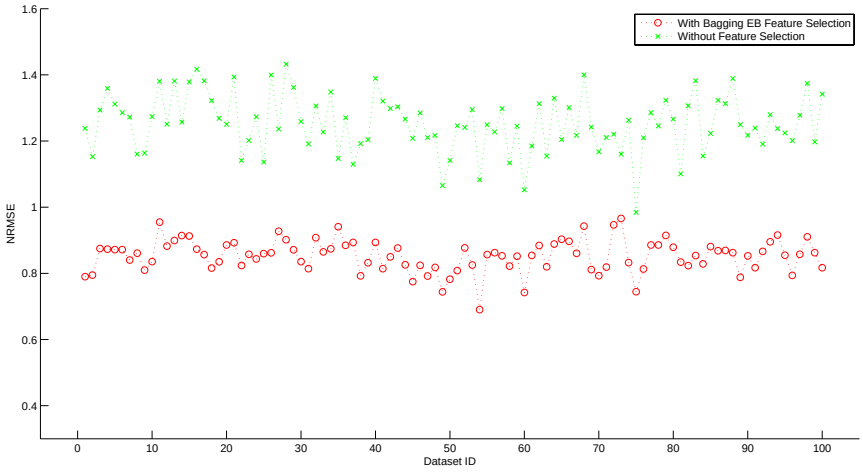

Figure 5.79: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.7 by NRMSE measurement.

- co-dominant model with background noise level 0.9 ($\alpha = 0.5, \beta = 0.9$)

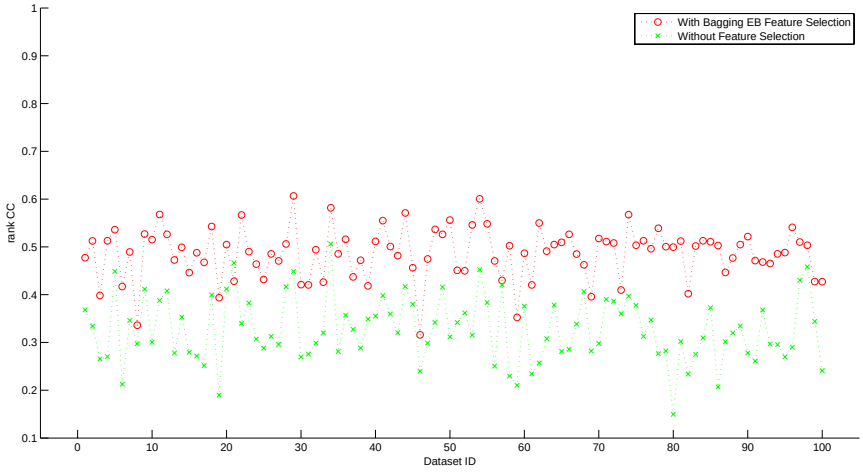

Figure 5.80: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.9 by rCC measurement.

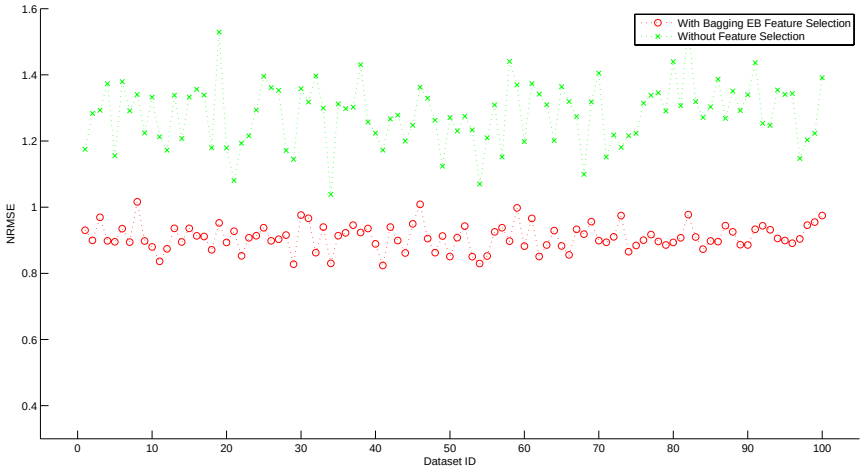

Figure 5.81: Algorithm performance with and without bagging EB feature selection method for the co-dominant model with background noise level 0.9 by NRMSE measurement.

Chapter 6

Conclusions

6.1 Conclusions

Genetic improvement of livestock populations can be achieved though detection of genetic markers linked to QTL. With the completion of the bovine genome sequence assembly, SNP assays spanning the whole bovine genome and research work on large scale genetic evaluation using genomic selection in cattle has become possible. Since DNA extraction is not restricted by age or gender, genomic selection can alleviate some of the limitations of quantitative genetic selection. Genomic selection promises to allow prediction of accurate EBVs based on genotypic information in newborn individuals without phenotypic records. Breeding program design with genomic selection will reduce the generation interval greatly and shift the structure of cattle breeding industry. In this thesis, we demonstrated the use of genotype data for QTL mapping and EBV prediction, by traditional methods and several major machine learning algorithms. We focus on the evaluation of a bagging EB method in terms of its ability to select a subset of markers for accurate breeding value prediction. The evaluation was performed using several simulated and real datasets consisting of genotypes and phenotypes.

One of the major objectives of this thesis is to verify whether machine learning methods will help in genomic selection. Results on simulation datasets and real datasets showed that machine learning methods work well even when the noise level (environmental effect) is high. Machine learning methods outperform the traditional genomic selection method on both of the real datasets, which is a sign that we should start to use machine learning methods instead of traditional genomic selection method in future breeding selection.

Another conclusion from the experiments is that, the bagging EB method can do a good job in detecting the SNP markers associated with the phenotypic traits, which makes it possible that genes affecting those phenotypic traits can be located. All the experiment results indicate that the bagging EB method can serve as a method for QTL mapping on perhaps any type of datasets.

When using the SNP markers detected by the bagging EB method to predict breeding values, the prediction accuracy improved dramatically on the simulation datasets. However, the bagging

EB method failed as a feature selection method to help improve prediction accuracy on two real datasets.

One possible reason is that the real datasets contain too much noise. Due to the relatively small sample size as compared with the number of SNPs, only a few or no samples are available per SNP marker, which makes the feature selection methods fail to discover the true interactions between SNPs and phenotypic traits. The more phenotypic records are available, the more observations there will be per haplotype or SNP marker, and the more accurate the detection of QTL associated SNP markers or haplotypes.

Another possible reason might be that in real datasets, it is the haplotypes instead of single SNP marker that are associated with the phenotype traits. In our simulation model, we simulated QTL as they are in LD with single SNP markers. That might be why the bagging EB method makes a great success in the simulation datasets. However, in real datasets, it would be more reasonable to use haplotypes instead of SNP markers to do the association study with phenotypic traits. Several experiments conducted by Hayes *et al.* on real datasets showed that using marker haplotypes will give better accuracy of QTL mapping than using single markers [24]. Therefore, our future work would be to verify the applicability of haplotypes for genomic selection.

6.2 Future Work

6.2.1 Using Haplotypes

The advantage of haplotypes over single markers in genomic selection is that marker haplotypes may be in greater LD with the QTL alleles than single markers. This is dependent on how accurately *identical by descent* (IBD) chromosome segments are identified using haplotypes, compared to that using single markers. If the haplotype consists of many markers instead of a single marker, the possibility of regenerating identical marker haplotypes by recombination is reduced. As a result, the possibility that identical haplotypes carried by different animals are IBD is increased, then the proportion of QTL variance which is explained by the haplotype effects will increase. Therefore, we can say that the marker haplotypes are more likely to be associated with QTL alleles.

Results from Hayes *et al.* showed that for a real dataset, using marker haplotypes gave better accuracy of QTL mapping than using single markers [24]. They also found that in genomic selection, use of the IBD approach gave greater accuracy of breeding values than using either single marker regression or regression on haplotypes, particularly at low marker densities (or lower LD between adjacent makers). In the experiment of Roos *et al.* the accuracy achieved by the IBD approach was also higher than regression on single markers or markers haplotypes [12]. On the other hand, Calus *et al.* also compared the accuracy of EBV prediction using single markers or marker haplotypes on simulated data [8]. They found that the prediction accuracy of using haplotypes increased at lower marker densities. However, if the LD between adjacent markers was 0.2 or greater, the advantage of

using marker haplotypes is not so obvious [8].

In future research, we would like to verify whether prediction accuracy increases using haplotypes or IBDs instead of using single markers.

6.2.2 Including Non-additive Effects in Simulation Models

Breeding values, by definition, should include only genetic additive effects, which can be passed from one generation to the next. However, it might improve the accuracy of estimating by including dominance and epistatic effects. "Moreover, dominance and epistatic effects can be exploited to produce sets of progeny with maximum genetic merit, through mate selection for example" [31].

Estimates of dominance effects with single markers is straight forward by extending the genetic model to estimate two effects per SNP rather than one. However, the estimation of epistatic effects is more difficult due to the extremely large number of pairwise combinations between hundreds of thousands of markers or haplotypes, and thus is very time consuming.

In the future work, we would like to add the non-additive effects to our simulation models and verify whether machine learning methods can still help to increase the prediction accuracy.

Bibliography

[1] W. Barendse, S. M. Armitage, and L. M. Kossarek. A genetic linkage map of the bovine genome. *Nature Genetics*, 6:227–235, 1994.

[2] V. Blanz, B. Schoolkopf, H. Bulthoff, C. Burges, V. N. Vapnik, and T. Vetter. Comparison of view-based object recognition algorithms using realistic 3D models. In *Proceedings of International Conference on Artificial Neural Networks*, pages 251–256, 1996.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[4] B. Bost, D. de Vienne, F. Hospital, L. Moreau, and C. Dillmann. Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. *Genetics*, 157:1773–1787, 2001.

[5] K. W. Broman and T. P. Speed. A review of methods for identifying QTL in experimental crosses. *Statistics in Molecular Biology and Genetics*, 33:114–142, 1999.

[6] M. P. L. Calus, S. P. W. de Roos, and R. F. Veerkamp. Estimating genomic breeding values from the QTL-MAS workshop data using a single SNP and haplotype/IBD approach. *BMC Proceedings*, 3:S1–S10, 2009.

[7] M. P. L. Calus, T. H. E. Meuwissen, S. P. W. de Roos, and R. F. Veerkamp. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*, 178:553–561, 2008.

[8] M. P. L. Calus and R. F. Veerkamp. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics*, 124:362–368, 2007.

[9] E. A. Carbonell, T. M. Gerig, E. Balansard, and M. J. Asins. Interval mapping in the analysis of nonadditive quantitative trait loci. *Biometrics*, 48:305–315, 1992.

[10] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[11] H. D. Daetwyler, B. Villanueva, P. Bijma, and J. A. Woolliams. Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics*, 124:369–376, 2007.

[12] S. P. W. de Roos, C. Schrooten, E. Mullart, M. P. L. Calus, and R. F. Veerkamp. Genomic selection for fat percentage using markers on BTA14. *Journal of Dairy Science*, 2007.

[13] J. C. M. Dekkers. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *Journal of Animal Science*, 82:E313–E328, 2004.

[14] J. C. M. Dekkers and F. Hospital. The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, 3:22–32, 2002.

[15] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. N. Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9:155–161, 1997.

[16] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *American Statistical Association*, 97:77–87, 2002.

[17] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–451, 2004.

[18] R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399, 1918.

[19] C. Fornell and F. L. Bookstein. Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research*, 19:440–452, 1982.

[20] M. E. Goddard and B. J. Hayes. Genomic selection. *Journal of Animal Breeding and Genetics*, 124:323–330, 2007.

[21] L. Grapes, J. C. M. Dekkers, M. F. Rothschild, and R. L. Fernando. Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics*, 166:1561–1570, 2004.

[22] C. S. Haley and S. A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69:315–324, 1992.

[23] B. Hayes. QTL mapping, MAS, and genomic selection. Technical report, Animal Breeding and Genetics, Department of Animal Science, Iowa State University, 2007.

[24] B. J. Hayes, A. J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M. E. Goddard. Accuracy of marker assisted selection with single markers and marker haplotypes in cattle. *Genetics Research*, 89:215–220, 2007.

[25] S. C. Heath. Markov chain Monte Carlo segregation and linkage analysis of oligogenic models. *The American Journal of Human Genetics*, 61:748–760, 1997.

[26] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231, 1968.

[27] R. C. Jansen. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics*, 85:252–260, 1992.

[28] R. C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135:205–211, 1993.

[29] T. Joachims. *Advances in Kernel Methods: Support Vector Learning*, chapter Making large scale SVM learning practical. MIT Press, 1999.

[30] C. H. Kao, Z. B. Zeng, and R. D. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152:1203–1216, 1999.

[31] B. P. Kinghorn. Mate selection by groups. *Journal of Dairy Science*, 81:55–63, 2007.

[32] D. Kolbehdari, L. R. Schaeffer, and J. A. B. Robinson. Estimation of genome-wide haplotype effects in half-sib designs. *Journal of Animal Breeding and Genetics*, 124:356–361, 2007.

[33] D. Kolbehdari, Z. Wang, J. R. Grant, B. Murdoch, A. Prasad, Z. Xiu, E. Marques, P. Stothard, and S. S. Moore. A whole genome scan to map QTL for milk production traits and somatic cell score in Canadian Holstein bulls. *Journal of Animal Breeding and Genetics*, 126:216–227, 2009.

[34] D. Kolbehdaria, G. B. Jansenc, L. R. Schaeffera, and B. O. Allend. Power of QTL detection by either fixed or random models in half-sib designs. *Genetics Selection Evolution*, 37:601–614, 2005.

[35] L. Kruglyak and E. S. Lander. A nonparametric approach for mapping quantitative trait loci. *Genetics*, 139:1421–1428, 1995.

[36] R. Lande and R. Thompson. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124:743–756, 1990.

[37] E. S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121:185–199, 1989.

[38] N. Long, D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendan. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *Journal of Animal Breeding and Genetics*, 124:377–389, 2007.

[39] P. D. Markel, D. W. Fulker, B. Bennett, R. P. Corley, J. C. DeFries, Erwin V. G., and Johnson T. E. Quantitative trait loci for ethanol sensitivity in the LS SS recombinant inbred strains: interval mapping. *Behavior Genetics*, 26:447–458, 1996.

[40] O. Martinez and R. N. Curnow. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics*, 85:480–488, 1992.

[41] L. K. Matukumalli, C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and Van Tassell C. P. Development and characterization of a high density snp genotyping assay for cattle. *PLoS ONE*, 4:e5350, 2009.

[42] T. H. E. Meuwissen. Genomic selection: the future of marker assisted selection and animal breeding. In *FAO: Biotechnology in Food and Agriculture: Conference 10*, 2003.

[43] T. H. E. Meuwissen. Genomic selection: marker assisted selection on a genome wide scale. *Journal of Animal Breeding and Genetics*, 124:321–322, 2007.

[44] T. H. E. Meuwissen and M. E. Goddard. The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution*, 28:161–176, 1996.

[45] T. H. E. Meuwissen, B. J. Hayes, and Goddard. M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001.

[46] W. M. Muir. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, 124:342–355, 2007.

[47] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using a support vector machine. In *Neural Networks for Signal Processing Proceedings of the 1997 IEEE Workshop*, pages 511–519, 1997.

[48] K. R. Muller, A. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen, and V. N. Vapnik. Predicting time series with support vector machines. In *International Conference on Artificial Neural Networks*, pages 999–1004, 1997.

[49] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing Proceedings of the 1997 IEEE Workshop*, pages 276–285, 1997.

[50] PerkinElmer Inc. `http://las.perkinelmer.com/content/snps/genotyping.asp`.

[51] M. Pirooznia, J. Y. Yang, M. Q Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9:S1–S13, 2008.

[52] N. Piyasatian, R. L. Fernando, and J. C. M. Dekkers. Genomic selection for marker-assisted improvement in line crosses. *Theoretical and Applied Genetics*, 115:665–674, 2007.

[53] C. E. Rasmussen. Gaussian processes in machine learning. Technical report, Institute for Biological Cybernetics, Machine Learning Summer School, Tubingen, 2003.

[54] J. M. Satagopan, B. S. Yandell, Newton M. A., and Osborn T. C. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, 144:805–816, 1996.

[55] L. R. Schaeffer. Strategy for applying genome wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 123:218–223, 2006.

[56] J. Shlens. A tutorial on principal component analysis. Technical report, Institute for Nonlinear Science, University of California, San Diego, 2005.

[57] M. J. Sillanpaa and E. Arjas. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, 148:1373–1388, 1998.

[58] M. Soller and J. S. Beckmann. Genetic polymorphism in varietal identification and genetic improvement. *Theoretical and Applied Genetics*, 67:25–33, 1983.

[59] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21:631–643, 2005.

[60] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[61] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.

[62] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[63] B. Villanueva, R. Pong-Wong, J. Fernandez, and M. A. Toro. Benefits from marker-assisted selection under an additive polygenic genetic model. *Journal of Animal Science*, 83:1747–1752, 2005.

[64] H. Wang, Y. M. Zhang, X. Li, G. L. Masinde, S. Mohan, D. J. Baylink, and S. Xu. Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics*, 170:465–480, 2005.

[65] W. Wei. Breeding value estimation and quantitative trait loci detection by machine learning methods based on high dimensional SNP dataset. Master's thesis, Department of Computing Science, University of Alberta, 2008.

[66] J. C. Whittaker, R. Thompson, and M. C. Denham. Marker assisted selection using ridge regression. *Annals of Human Genetics*, 63:366–366, 1999.

[67] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. *Advances in Neural Information Processing Systems*, 8, 1996.

[68] S. Xu. Theoretical basis of the Beavis effect. *Genetics*, 165:2259–2268, 2003.

[69] S. Xu. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, 63:513–521, 2007.

[70] N. Yi, V. George, and D. B. Allison. Stochastic search variable selection for identifying quantitative trait loci. *Genetics*, 164:1129–1138, 2003.

[71] L. Zhang, H. Li, Z. Li, and J. Wang. Interactions between markers can be caused by the dominance effect of QTL. *Genetics*, 180:1177–1190, 2008.

[72] Y. M. Zhang and S. Xu. A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heredity*, 95:96–104, 2005.

[73] H. H. Zhao, R. L. Fernando, and J. C. M. Dekkers. Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. *Genetics*, 175:1975–1986, 2007.

[74] H. Zou and T. Hastie. *ElasticNet: Elastic Net Regularization and Variable Selection*. R package version 1.03, 2005.

[75] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series*, 67:301–320, 2005.

Index

α (dominance level), 18
β (background noise level), 17

binary representation, 19
BLUP (best linear unbiased prediction), 13

CC (correlation coefficient), 46
co-dominance representation, 19

EB (empirical Bayes), 20
EBV (genomic breeding value), 6
ElasticNet (elastic-net regression), 27

genomic selection, 8
GP (Gaussian process), 24

IBD (identical by descent), 92

LASSO (least absolute shrinkage and selection operator), 26
LD (linkage disequilibrium), 5
linkage mapping, 4

machine learning, 8
MAS (marker assisted selection), 7

NRMSE (normalized root mean square error), 47

PCA (principal component analysis), 24
PLS (partial least square), 25

QTL (quantitative trait loci), 3
QTL mapping, 3
quantitative genetics approaches, 7

rCC (rank correlation coefficient), 47
Ridge (ridge regression), 26

SNP (single nucleotide polymorphism), 5
SVM (support vector machine), 23