The Pennsylvania State University

The Graduate School

College of Medicine Public Health Sciences

STATISTICAL MODELS FOR HIGH DIMENSIONAL SCREENING OF GENETIC AND EPIGENETIC EFFECTS

A Dissertation in

Biostatistics

by

Kirk Gosik

Doctor of Philosophy

February 2017

The dissertation of Kirk Gosik was reviewed and approved* by the following:

Rongling Wu

Distinguished Professor of Public Health Sciences and Statistics

Thesis Advisor, Chair of Committee

Vernon Chinchilli

Distinguished Professor and Chair of Public Health Sciences

Lan Kong

Associate Professor of Public Health Sciences

James Broach

Distinguished Professor and Chair of Biochemistry and Molecular Biology

Abstract

Knowledge about how changes in gene expression are encoded by expression quantitative trait loci (eQTLs) is a key to construct the genotype-phenotype map for complex traits or diseases. Traditional eQTL mapping is to associate one transcript with a single marker at a time, thereby limiting our inference about a complete picture of the genetic architecture of gene expression. Here, I present innovative applications of variable selection approaches to systematically detect main effects and interaction effects among all possible loci on differentiation and function of gene expression and other phenotypes of interest. Forward-selection-based procedures were particularly implemented to tackle complex covariance structures of gene-gene interactions. Simulation studies were performed on each of the models to assess the computational properties of each model. Applications of the models were also performed on real datasets. The first was a reanalysis of a published genetic and genomic dataset collected in a mapping population of Caenorhabditis elegans, gaining new discoveries on the genetic origin of gene expression differentiation, which could not be detected by a traditional one-locus/one-transcript analysis approach. The next dataset was of Mei Tree growth, analyzing the genetic control of the height and diameter during the developmental process. The underlying genotypes and epistasis that impact the process of these developments were considered as candidates for the selection of the procedure.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

There are several techniques used for studying genetics and mapping the results. Some of the more populor techniques include cross-breeding experiments or, in the case of humans, the examination of family histories, known as pedigrees. More recently, CRISPR/Cas9 can be used to mimic mitotic recombination to help map out genes as well. (**?**)

Construction of genetic maps are a variety of techniques used to show relative positions between genes or other sequence features of the genome and the phenotype that is controlled by such sequences. Genes are very useful markers but they are by no means ideal. One problem, especially with larger genomes such as those of vertebrates and flowering plants, is that a map based entirely on genes is not very detailed.(**?**) Genes have long areas of non-coding regions between them and therefore result in large gaps from gene to gene. This is further complicated because not every gene has allelic forms that can be easily or conventiently distinguished. With these considerations in

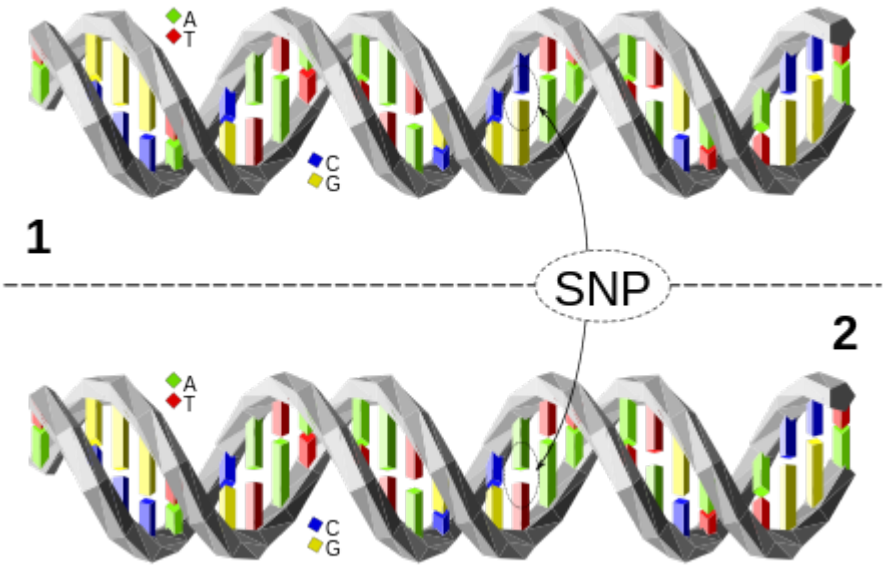

Figure 1.1: SNP Picture

mind gene maps may not be comprehensive enough and other markers may be needed.

According to brown, mapped features that are not genes are called DNA markers. As with gene markers, a DNA marker must have at least two alleles to be useful. There are three types of DNA sequence feature that satisfy this requirement: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and single nucleotide polymorphisms (SNPs). (**?**) The genetic markers that have been emphaised in this work are single nucleotide polymorphisms. Attempting to be at the highest levels of resolution for identifying quantitative traits, using SNPs are the most specific case. This will give exact location of the nucletid that may be impacting the genetic control over the phenotype.

There are several goals to genetic mapping and association studies that identify certain regions of the genome that contain genes involved in specifying a quantitative trait, refered to as quantitative trait loci (QTLs). One main goal is to estimate the genetic effects of these loci. The relationship between the genetic effects of QTLs and the phenotypic value of quantitative traits can be described

by a linear model (**?**, **?**). Typically, because of the high throughput nature of the data there are a large number of markers across the whole genome, and most of the markers may have very little or next no effect on the phenotype under study. The models can be very sparse, with most cases, the number of genetic markers or variables is bigger than the sample size, especially when interactions among markers are considered. This makes a model is oversaturated and furher model selection techniques may be required to capture the necessary information. **?**

1.2 Some Exisiting Methods

Numerous methods exist and are being developed to measure and find quantitative trait loci (QTL) effects. These methods can broadly fall into three main categories. These categories are Least-Square methods, maximum likelihood and Bayesian approaches. (**?**) Each method has advantages and considerations that you would need to be aware before conducting analyses to find QTL effects from the given markers. A brief discussions on a few of the methods are given to highlight some areas of consideration and how the methods proposed can handle such considerations.

Marker Regression would fall in the category of Least Squares approaches. If looking at one marker analysis general t-test and ANOVA procedures can be used to analyze the relationship. It is not recommended however for use in general practice because you do not know how dense the markers are measured. QTL interval mapping would be preferred in such an analysis because the methods take account for missing genotype data that may not have been measured. When estimating a QTL position through maximum likelihood methods, like interval mapping, positions of other possible QTLs could affect the detection of the true position. Neighboring QTLs could possibly flatten the likelihood in instances where there are multiple QTLs on the same chromosome. This would make

an effect look less significant at a given location than it actually is. Another possibility is that in the search over the interval you may find an area where the likelihood could reach a peak but could be a "ghost" QTL. This is where an effect is observed because a neighboring QTL is skewing the results at the particular position you are looking in and the result is a false discovery of the position. Marker Regression has been shown to improve interval mapping, which is call Composite Interval Mapping. This is where the QTL position found is also combined in a linear regression where the covariates are the other markers in the dataset. By including the markers as covariates the other position in the chromosome are accounted for in the analysis and false discovery is reduced.

The analysis of interval mapping and single marker analyses has shown to be effective but it limits our inference to one marker at a time as a possible loci that controls a trait. Using Marker Regression however you can incorporate multiple markers in a single analysis to test for possible QTL for a given trait. It is cautioned that running such an analysis is only an approximate test because the null hypothesis is there is no difference between the marker levels and therefore a non-mixture distribution but the alternative is a mixture of distributions. The assumptions regression would make of the errors within the marker type to be normally distributed may not be entirely met if the QTL's fall between the marker regions. However **?** have shown that a direct regression of phenotypes on marker types, provides the same information about location of QTL-effects without having to step to all positions on the interval. With this information using the entire marker set in a regression analysis would provide a nice, computationally efficient way to map out the genetic architecture of a trait.

1.3 Chapter Overview

The main goal of this paper is to propose an improved selection procedures which use regression techniques to approach high scale variable selection problems such as the ones arising in epistatic analysis

The variable selection procedure for QTLs mapping can be seen as one of deciding which subset of variables have effects on phenotypes, and identifying out all possible effects of those markers.

1.3.1 HighDeQTL

In this chpater we introduce the iform procedure, originally proposed by (hao and zhang 2014). This includes the algorithm and how it compares to forward selection. From there it is adapted to use for a genetic mapping studies. The properies of this will be explored in detail. Simulation studies were conducted to assess the properties. These were also compared to other models to get a sense of the utility and advantages that come with the new selection procedure. After the comparisons and simulations a real world application is performed. In the application a data set using C Elegans

The model perfromed well but feedback was given on… consider weak heredity

Needs more flexibility

a lot of important genetic variance and heritability were discovered by the inclusion epistatic effects. These effects boost the overall variance explained and predictive power of the final model selected.

1.3.2 Higher Order Epistasis

Higher order epistasis is important but under studied because of practical limitations not because of biological limitations

from chapter 03 The theoretical models of high-order epistasis have well been established by mathematical biologists (**?**; **?**). These models provided a foundation to interpret high-order epistasis from a biological standpoint. A few statistical models have been derived to estimate and test high-order epistasis in case-control designs (**?**) and population-based mapping settings

By extending the order of interaction effects the properties would still need to be studied. The thereotrical properties that held for the iform procedure were inspected again. Simulation studies were performed to assess practical applications. Several scenarios and comparison models were considered to extensively look at what properites were being met and which were not. Then an application to Mei tree growth was conducted in order to see the real world application of such a selection procedure. Different growth parameters were prviously fit and these were used as the phenotype. Interesting and more predictive implications came out of the model when considering higher order epistasis throughout the selection procedure.

The static growth parameters were interesting but using the entire growth curve would be more beneficial to capture all relevant information.

1.3.3 iForm Funcional Mapping

In order to use all relevant information of the repeated measure data would take additional computation burden to the selection and the modeling but it would also give more power and flexibility to

the modeling that would not be present otherwise. It is important to use all relevant information in order to make the most accurate prediction about the data.

Using a growth curve model to assist in fitting the data would help ease some of the comutational burden. The selection of genetic effects however would be very simplistic as some additive shift to the curve. This simplistic view may not be the most accurate and therefore more complicated structures could produce better results. As we have seen in Chapter 03, the genetic effects over time seem to vary among different genotypes and the interaction effects follow more of a non-linear pattern. In order to allow for more flexible, non-linear modeling with not expending too much computation burden, Legendre Orthogonal Polynomials were considered to model the genetic effects. These polynomials have various form and would allow for the genetic effect to follow different patterns but also not induce unnecessary correlation between predictors included in the model.

These benefits do come at a computational cost but by including the repeated measures portion of the data it enabled for the use of more data and therefore increase the power of the modeling.

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter **??**. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))

plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure **??**. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table **??**.

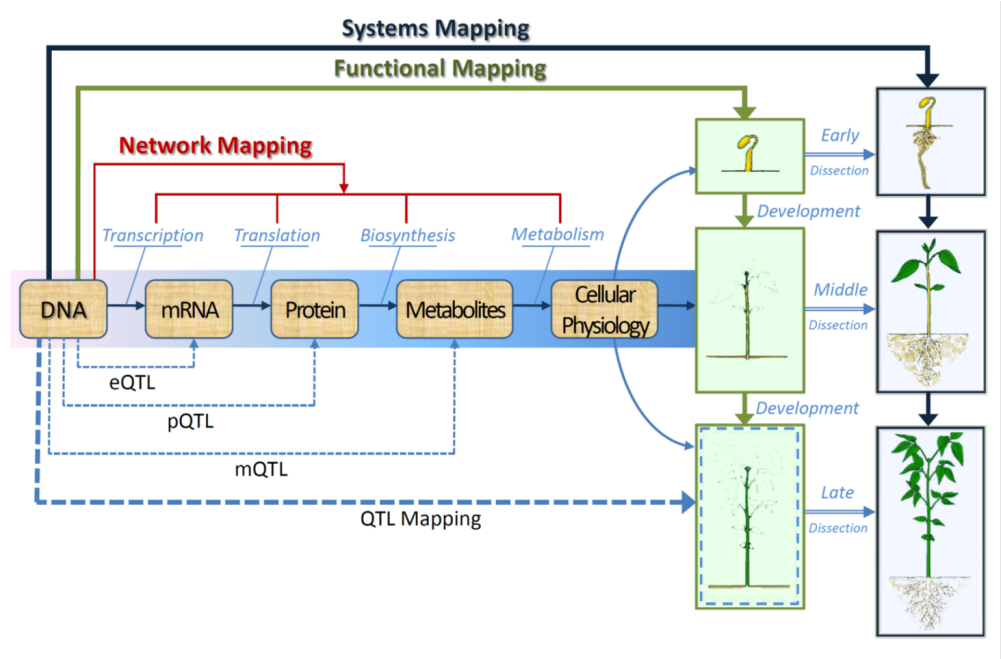

Figure 1.2: Systems Map

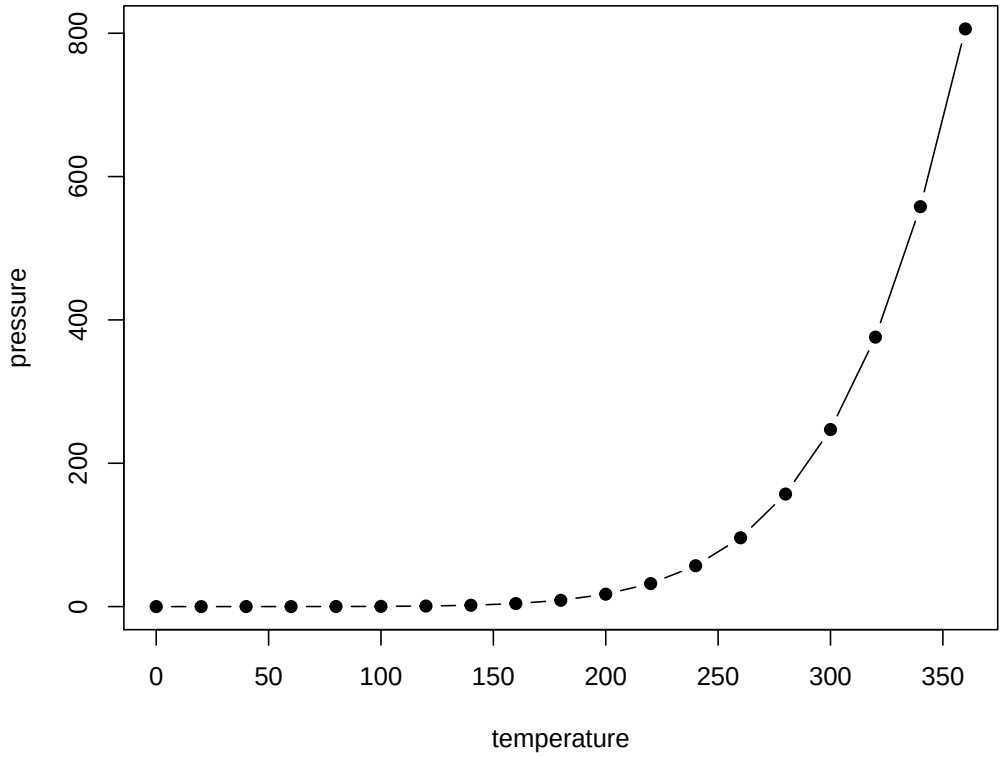

Figure 1.3: Here is a nice figure!

Table 1.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (**?**) in this sample book, which was built on top of R Markdown and **knitr** (**?**).

Chapter 2

High Dimensional eQTL

2.1 Motivation

Since activation or inhibition of gene expression causes change in phenotypic formation, the identification of expression quantitative trait loci (eQTLs) that regulate the pattern of gene expression is essential for constructing a precise genotype-phenotype map (**?**; **?**; **?**). With the advent and development of various biotechnologies, it has become possible that genome-scale marker and expression data can be generated, providing an important fuel to systematically study the biological function of any types of cellular components in an organism (**?**; **?**; **?**). Several genome-wide association studies (GWAS) have been initiated to map a complete set of eQTLs for the abundance of genome-wide transcripts whose expression levels are related to biological or clinical traits (**?**;**?**; **?**). Statistical analysis and modeling are playing an increasing role in mapping and identifying the underlying eQTLs from massive amounts of observed data (**?**; **?**; **?**; **?**).

A typical eQTL mapping approach is to associate a gene transcript with a single marker such

as single nucleotide polymorphism (SNP). By analzying the significance of all these markers one by one adjusted for multiple testing, one can count significant loci that contribute to variation of expression by the gene. This marginal approach based on a simple regression model has been instrumental for the identifcation of eQTLs in a variety of organisms (**?**; **?**). However, there are two major limitations for the results by such a marginal analysis: First, it does not take into account the dependence of different markers, thus a significant association detected by one marker may be due to the other markers that are linked with it. The marginal marker analysis cannot separate the confounding effect of eQTLs due to marker-marker dependence or linkage (**?**). Second, an eQTL may act through its interaction with other eQTLs and environmental factors. Because of their paramount importance in affecting complex diseases and traits, gene-gene interactions, or epistatic effects, and gene-environment interactions have been studied intensively in modern biological and medical research (**?**; **?**; **?**; **?**)

These two limitations can be overcome by analyzing all markers and their pairwise interactions simultaneously through formulating a high-dimensional regression model. Although it can infer a complete picture of the genetic architecture of gene expression, this endeavor is highly challenged by the curse of dimensionality, i.e., the number of predictors far exceeds the number of observations. The past decade has witnessed the tremendous development of variable selection models for high-dimensional data analysis, such as LASSO (**?**), SCAD (**?**), Dantzig selector (**?**), elastic net (**?**), minimax concave penalty (MCP) (**?**) among others. Many methods possess favorable theoretical properties such as model selection consistency (**?**) and oracle properties (Fan and Lv 2011). When the number of predictors is much larger than the number of observation, sure screening is a more realistic goal to achieve than oracle properties or selection consistency (**?**; **?**). Sure screening assures that all important variables are identified with a probability tending to one, hence achieving effective dimension reduction without information loss and providing a reasonable starting point

for low-dimensional methods to be applied.

More recently, Hao and Zhang (**?**) extended variable selection approaches to jointly model main and interaction effects from high-dimensional data. Based on a greedy forward approach, their model can identify all possible interaction effects through two algorithms iFORT and iFORM which have been proved to possess sure screening property in an ultrahigh-dimensional setting. In this article, we implement and reform Hao and Zhang's model to map the genetic architecture of eQTL actions and interactions for gene expression profiles. This model is modified to accommodate to the feature of a genetic mapping or GWAS design in which molecular markers as genetic predictors are discrete although some additional continuous predictors can also be considered. We expand Hao and Zhang's regression model to include discrete components. Also, for an F2 or a natural population with three genotypes at each locus, we need to estimate a total of eight genetic effects for a pair of markers, which are additive and dominant effects at each locus, and additive $\times$ additive, additive $\times$ dominant, dominant $\times$ additive and dominant $\times$ dominant effects between the two loci (**?**). Thus, if the number of markers is p, a total number of predictors including all main and two-way interaction terms is $2p^2$. For a typical moderate-sized mapping study, in which several thousands of markers are genotyped on a few hundred individuals, consideration of pair-wise genetic interactions will quickly make the dimension of predictors an ultrahigh one.

By modeling all markers jointly at one time under an organizing framework, the modified model can detect all possible significant eQTLs and their epistasis. An eQTL can be either a cis-QTL, coming from the same physical location as the gene expression, or a trans-QTL, coming from other areas of the genome. Our model can more precisely discern these two different types of eQTLs and their interactions than traditional marginal analysis. By reanalyzing a published data collected in a mapping population of C. elegans (**?**), the new model has validated previous results by the

marginal approach, meanwhile obtained new discoveries on the genetic origin of gene expression differentiation, which could not be detected in a traditional way.

2.2 Methods

2.2.1 Experimental design

Consider an experimental population for genetic studies of complex traits, such as the backcross and F2 initiated from two inbred lines, full-sib family derived from two outcrossing parents, or random samples drawn from a natural population. These types of populations are used specifically for different species. Although they have different levels of complexities for statistical modeling, the genetic dissection of different populations underlies a similar principle. For the purpose of simplicity, we consider a backcross design in which there are only two genotypes at each marker.

Suppose the backcross contains n progeny, each of which is genotyped by p markers, such as single nucleotide polymorphisms (SNPs), distributed over different chromosomes. The number of SNPs p should be large enough to completely cover the entire genome at an adequate depth so that we can possibly capture all possible genetic variants. An increasing body of evidence suggests that significant SNPs associated with complex traits or diseases are more likely to be eQTLs (**?**). Hence the identification of eQTLs is an important first step toward the genetic dissection of end-point phenotypes. For this reason, we assume that genome-wide gene transcripts have been available for the assumed study population. Assume that all progeny are recorded for the same organ by microarray, leading to expression abundance data of m gene transcripts. We purport to identify all possible genetic variants including main effects and interaction effects of SNPs that contribute to

each gene transcript.

2.2.2 Adaptation of iFORM procedure

Hao and Zhang **?** formulated an interaction forward selecting procedure under the marginality principle (iFORM). The marker and gene transcript data of the study population can be denoted as $(X_i, Y_i)(i = 1, \ldots, n)$ which are independent and identically distributed copies of (X,Y), where $X = (X_1, \ldots, X_p)^T$ is a p-dimensional predictor vector and Y is the response, expressed by a linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \ + \beta_p X_p + \epsilon(\#eq : lin - mod) \tag{2.1}$$

The 's are the coefficients for the genetic effects of each marker. Like most genome-wide datasets, the number of markers here grossly outnumbers the number of observations, $p >> n$. Therefore, selection procedures would need to be implemented in order to fit a linear regression model such as (1). We are already at the point of high-dimensional data but if we want to include epistatic effects between different markers as predictors as well it would increase the amount of predictors by $(p^2 + p)/2$. The resulting linear model would grow to be,

$$Y = \beta_0 + \beta_1 X_1 + \ + \beta_p X_p + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \ + \gamma_{pp} X_p^2 + \epsilon(\#eq : lin - mod2) \tag{2.2}$$

where 's are the coefficients for the epistatic effects for all the quadratic and two-way interactions between the markers. For convenience we will assume that the markers and the transcripts are

standardized before running the selection procedure. Therefore, E(X_ij)=0,Var(X_ij)=1,E(Y_i)=0 and Var(Y_i)=1 for i=1,…,n; j=1,…,p. Also, the quadratic and two-way interaction effects will be centered which we will write as Z_i=(…,X_ik X_il-E(X_ik X_il),…)ˆT. By doing so we would eliminate the need for an intercept in regression model (2). This would reduce the model to the form,

$$Y = X^T\beta + Z^T\gamma(\#eq : lin - mod3) \tag{2.3}$$

Some notations that will be used to define the elements of **?** iFORM procedure are as follows. $P_1 = 1, 2, \ldots, p$ $P_2 = (k, l) : 1 \leq k \leq l \leq p$. which are the index sets for the linear and two-way interactions terms, respectively. The significant main effects for the markers and their interaction effects are $T_1 = j : \beta_j \neq 0, j \in P_1, T_2 = (j, k) : \beta_{jk} \neq 0, (j, k) \in P_2$. For any model M, |M| will be used to denote the number of predictors contained in the model. The true model size would be indicated by $|T_1| = p_0$ and $|T_2| = q_0$ or together would be |T|=d_0=p_0+q_0. For the procedure, three sets will be used throughout. The sets are M for the model set, C for the candidate set of predictors and S for the solution set of predictors currently selected in the model.

There are two principles that are used in the selection procedure when considering interactions as candidates for selection into the final model. The first is considering the principle of marginality. The principle states that it is inappropriate to model interaction terms when the main effects contributing to the interaction have either not been included in the model or are deleted because their effects become marginal by the inclusion of the interaction effect. The second principle important to the procedure is the heredity principle. The strong case of the principle states that an interaction effect should not be considered unless both the contributing main effects are in the