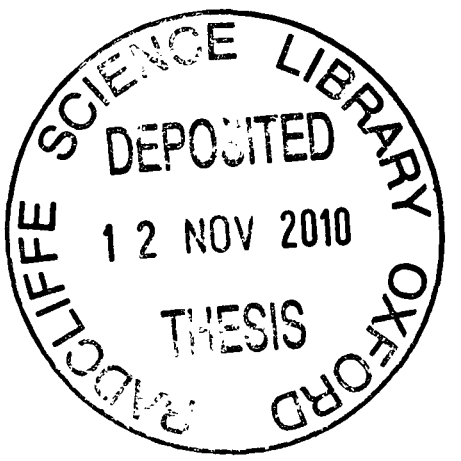
Statistical methodology for QTL mapping and genome-wide association studies

Submitted for the degree of Doctor of Philosophy

Alexander Antonyuk

Wolfson College, The University of Oxford

2009



Statistical methodology for QTL mapping and genome-wide association studies

Alexander Antonyuk, Wolfson College, the University of Oxford

Submitted for the degree of Doctor of Philosophy, Trinity 2008

Abstract

This work deals with statistical tests of association between genetic markers and disease phenotypes. The main criterion used for comparing the tests is statistical power. First we consider animal models and then data from association studies of humans. For the animal section, we analyse a dataset from a prominent mouse experiment which developed a heterogeneous stock of mice via multiple crosses. This stock is characterised by small distances between recombinants which allows fine mapping of genetic loci, but also by uncertainty in haplotypes. We start by highlighting the disadvantages of the currently used approach to deal with this uncertainty and suggest a method that has greater statistical power and is computationally efficient. The method applies the EM algorithm to the broad class of exponential family distributions of phenotypes. We also develop a Bayesian version of the method, for which we extend the widely used IRLS algorithm to maximisation of the weighted posterior.

Then we move on to genome-wide association studies (GWAS), where two situations are considered: known and unknown minor allele frequency. First we develop an innovative Bayesian model with the optimal prior for the known population MAF. We demonstrate that not only it is more powerful than any frequentist test considered (the size of the advantage depends on prevalence of the disease and MAF), but also that the frequentist tests change ranking in terms of power. A remarkable property of the frequentist tests, the advantage of discarding part of the data to gain power, is highlighted.

The second chapter on GWAS considers the currently more common situation of the unknown MAF, when the Armitage test is known to be the most powerful frequentist method. We show that the suggested model is more powerful in the broad selection of settings considered, including the three different allele effect models: additive, dominant and recessive.

For both known and unknown MAF cases we point out that the parameters are constrained and demonstrate how to gain power by taking this constraint into account.

Acknowledgments

I would like to thank my supervisor professor Chris Holmes for his support, enthusiasm and sharing of interesting ideas; MRC Harwell Mammalian Genetics Unit for providing my PhD studentship.

I am very grateful to Caroline for her patience and loving support, this project would not be possible without her.

I also thank Alexandre Pintore for our inspiring discussions on statistics and many other topics.

Contents

1 Introduction **4**

1.1 QTL mapping 4

1.1.1 Quantitative traits 4

1.1.2 Mouse strain crosses 5

1.1.3 Visualisation of phenotype variation 8

1.2 Strategies for finding QTLs 10

1.2.1 Segregating populations and linkage 10

1.2.2 Interval mapping (IM) 13

1.2.3 Mapping multiple QTLs 15

1.2.4 Composite interval mapping (CIM) 16

1.2.5 Multiple-QTL mapping (MQM) 19

1.2.6 Multiple interval mapping (MIM) 22

1.2.7 Regression mapping 25

1.2.8 Bayesian methods 26

1.3 Genome-wide association studies 31

1.3.1 Study designs 31

2 Suggested methodology for QTL mapping in HS stock **39**

2.1 Current approach 39

2.1.1 Background 39

2.1.2 Details of the HAPPY model 42

2.1.3 Current statistical model for HS 45

2.2 Alternative method 48
2.3 The EM algorithm 50
2.4 EM algorithm for finding maximum-a-posteriori 53
2.4.1 IRLS for logistic regression 54
2.4.2 Logistic-Normal Bayesian model 56
2.4.3 Weighted log-posterior via IRLS 57
2.4.4 IRLS for the Bayesian Normal-Normal model 57
2.5 Estimation Results 58
2.5.1 Normal distribution 59
2.5.2 Binomial distribution 59
2.5.3 Computational efficiency 61
2.5.4 EM convergence 61
2.5.5 Convergence under no collinearity 63
2.5.6 Empirical power estimation 63
2.6 Sensitivity to errors in HAPPY estimates 65
2.7 Testing procedures 68
2.7.1 Likelihood ratio test 71
2.7.2 Overview of mixture models 73
2.7.3 F-like test 79
2.7.4 Conjecture 80
2.8 Future work 81

3 Association studies when allele frequencies are known 84
3.1 Introduction 85
3.2 Statistical models 86
3.3 Results 93
3.4 Applications of the method 97
3.5 Discussion 103

4 Hypothesis testing of association when MAF is not known **106**

4.1 Statistical models 107

4.2 A decision-theoretic protocol for setting priors 114

4.3 Results 117

4.4 Application to a colorectal cancer association study 118

4.5 Comparison to similar models 122

4.6 Discussion 124

Appendices **129**

A Notation **130**

B Computer programmes **131**

C Detailed results for Chapter 3 **135**

Chapter 1

Introduction

In this chapter we introduce concepts and terms used in QTL mapping and genome-wide association studies, and describe some of the most popular methods for testing for association between genetic variants and phenotypes. Although there has been some success in locating the genes responsible for traits of interest, the progress has been slow (Korstanje and Paigen (2002)). The difficulty is that, despite the fact that the overall effect which can be attributed to genetic background is often large, each individual locus may only have a small effect on the phenotype and this effect can easily be lost in environmental noise. A solution to this problem is to use laboratory animals which we introduce in the next section.

1.1 QTL mapping

1.1.1 Quantitative traits

Traits of interest to geneticists are more often quantitative than qualitative - the difference lies in measure of quantity rather than the kind of variation (i.e. Mendelian traits). Many medical conditions, such as diabetes and high blood pressure are the result of deviation form the normal level of the phenotype. These quantitative traits may have a complex genetic and environmental basis, their continuous values being the result of the action of one or more genes known as quantitative trait loci (QTLs)

situated at different positions in the genome combined with random variation from the environment, and possibly interactions between these genetic and environmental factors. Identifying the genes responsible for quantitative traits is the objective of QTL mapping and can provide insight into how complex diseases arise and can suggest new methods for treatment.

Animal models such as rats and mice have contributed to significant progress in our understanding of the biological mechanisms involved in human disease and have recently been used to define the genetic basis of these disorders. A major advantage of using animals rather than humans is that both environmental factors and genetic relationships can, if required, be controlled. Although many of the same problems that affect studies in humans are present in animal studies, the latter have several unique aspects, which make their use very attractive (Mott (2006)).

In the past 15 years, more than 2,000 QTLs have been identified in crosses between inbred strains of mice and rats, but less than 1% characterised at molecular level according to Flint et al. (2005). However, new resources, such as chromosome substitution strains and the proposed Collaborative Cross, together with new analytical tools, including probabilistic ancestral haplotype re-construction (PAHR) in outbred mice, YinYang crosses and In-Silico analysis of sequence variants in many inbred strains, could make QTL cloning tractable (Flint *et al.* (2005)).

Thus, further advances in QTL mapping will rely at least in part on effective use of PAHR. This chapter will examine modeling of the association between phenotypes and haplotypes following this type of reconstruction.

1.1.2 Mouse strain crosses

The goal of breeding a population for association mapping is to create marker loci in disequilibrium with QTL. It will be shown how such disequilibrium may create marker-trait associations. Perhaps the most straightforward way of engineering disequilibrium is through crosses between inbred lines. Inbred lines are strains of animals that are homozygous at all loci, i.e. have identical haplotypes. Below, we

review the existing strategies for QTL mapping based on inbred strain crosses.

- F2 intercrosses and backcrosses

  Two parental inbred strains are crossed to produce an F1, the starting point for the two most common mapping strategies. Intercrossing the F1 generates an F2 and backcrossing to one or other of the parental strains produces a backcross. Most of the examples for the methods outlined below are given either for a backcross or F2 populations.

- Recombinant inbred (RI) lines

  A pair of inbred strains is intercrossed and their progeny is inbred to generate a panel of inbred animals, each with a different combination of the progenitor genomes. One advantage of using RI lines is that they need to be genotyped only once, and the marker data can be used for all further experiments. Mapping is carried out by correlating the strains that are present at the markers with phenotypic differences.

- Congenics

  Congenics are most popularly used in fine-mapping QTLs in rodents. It is possible to ensure that animals have a particular genomic region from one strain and the rest of their genome from the other by repeatedly back-crossing one strain onto another; further intercrossing means that the genomic segment becomes homozygous and the mouse is fully inbred. This technique must be used with care. According to Flint *et al.* (2005), unpredictable results may occur when incorrect assumptions are made about the distribution of chromosome segments, the population structure, the marker spacing and the selection strategy.

- Advanced intercross lines

  Two parental strains are crossed to produce an F1 that is intercrossed to produce an F2. Subsequent generations are produced by intercrossing, and thus animals in an advanced intercross line accumulate new recombinants. A

pseudo-random breeding protocol is used to reduce loss of genetic variation in the population. An appropriate advanced intercross line might take 5 years or more to develop, but can then be used to map many loci. Advanced intercross lines are a particular case of heterogeneous stock (see below), where only two lines are used.

- Heterogeneous stock (HS)

  Advantages of using HS in QTL mapping are given in Section 2.1; here we only briefly describe what they are and how they are produced. To generate a heterogeneous stock one takes a number of inbred lines (8 in our case) and applies a series of intercrosses to them. The intercrosses are carried out in such a way that the resulting population of mice has average distance between recombinants of less than 2 centimorgans. It takes many generations to produce this fine-grain mosaic of recombinants (Figure 1.1, courtesy of William Valdar). The dataset we analysed here came from the Northport heteroge-

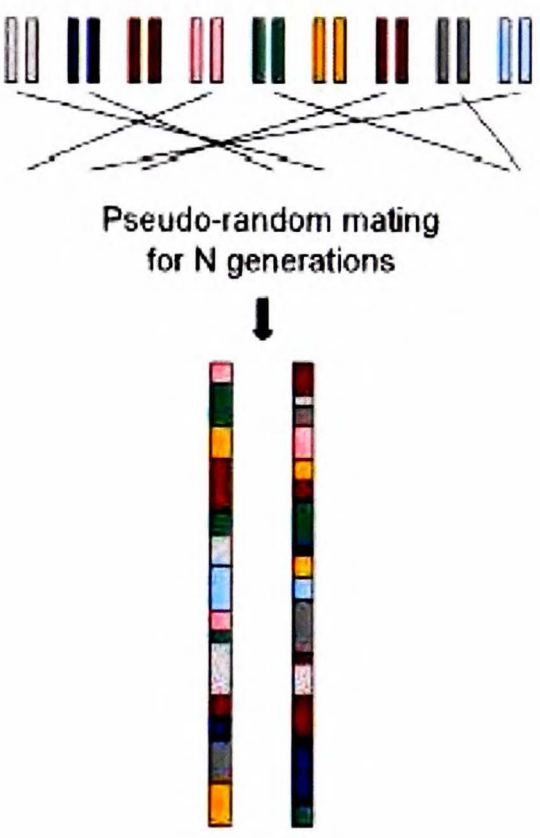


Figure 1.1: The HS breeding scheme.

neous stock, which has passed its fortieth generation, and is derived from the A/J,AKR/J,BALB/cJ,C3H/HeJ, C57BL/6J,CBA/J,DBA/2J and LP/J

strains. Compared with other crosses, analysing HS is more complicated because of the larger number of strains involved. PAHR is used to estimate probabilities of descending from progenitor strains.

1.1.3 Visualisation of phenotype variation

The first step after collecting phenotypic data, as advocated in Chapter 15 of Balding *et al.* (2003), should be visualising the data. The authors warn of possible pitfalls in interpretation. One example is given (Figure 1.2), where a unimodal and approximately normal distribution can be interpreted as a distribution of a trait encoding many genes of small action, but in reality is a single additive major gene from an F_2 population with a mixture of genotypes in the ratio $1 : 2 : 1$.

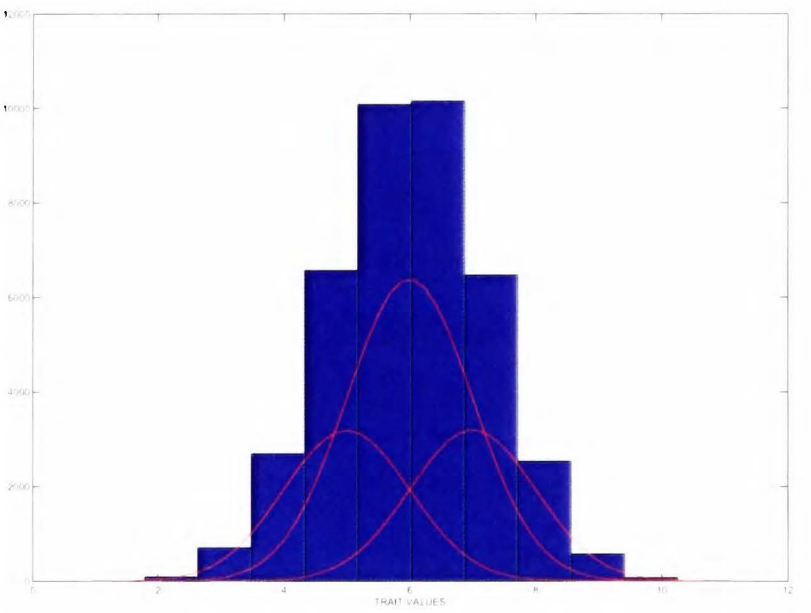


Figure 1.2: Mixture distribution. Lines show distribution of the components and the bars are histogram of the mixture.

In another example we observe a skewed unimodal distribution of an F_2 population and might be inclined to conclude that there is a dominant major gene action (Figure 1.3). Supposing individuals from two groups have the same mean phenotypes and underlie the main body of the distribution. The other group underlie a "shoulder" of the distribution and we make the assumption that the errors are symmetric or even normal. In reality the histogram was generated under a polygenic

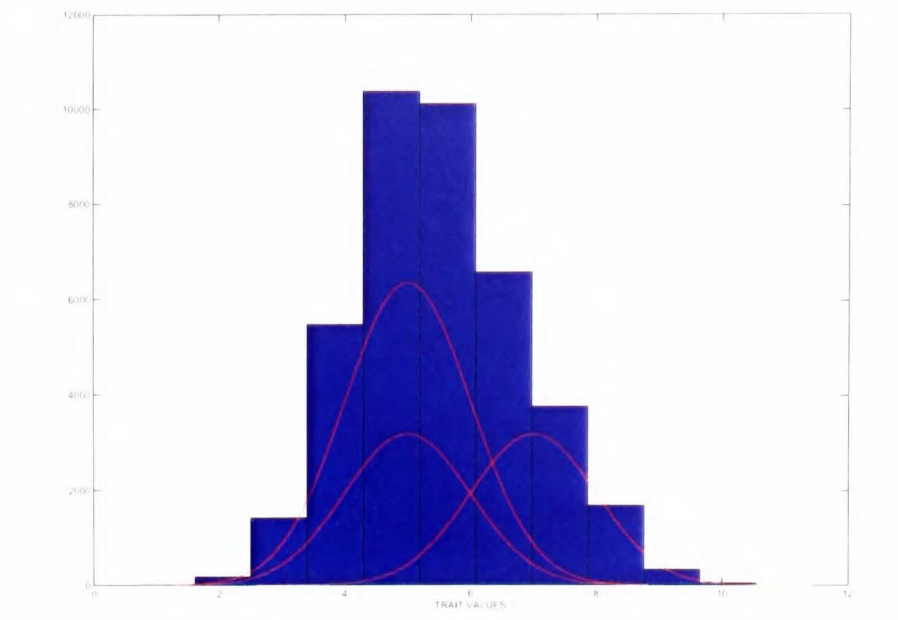


Figure 1.3: Mixture and its components for the second example.

model with a skewed residual error distribution.

Many researchers are aware of the pitfalls and transform the data but it has to be done with care because we may lose power for detecting dominant gene action. To check the type of distribution it is best to look at the parental data and use standard statistical techniques to help find an optimal data transformation.

Choice of mixing distribution becomes more critical when the model becomes more complex (e.g. if models include gene interactions). It is known, for instance, that significant interactions can arise as statistical artefacts: the interactions try to compensate for faulty model assumptions. Still, distribution checks are only occasionally reported in genetic analysis. In the same chapter Balding *et al.* (2003) also suggest plotting a mixture distribution (once it has been chosen) on top of the histogram of the phenotype, once parameters has been estimated. It can give a visualisation of how well the model fits the data.

1.2 Strategies for finding QTLs

1.2.1 Segregating populations and linkage

We begin with a simple but illustrative example for one QTL linked to a marker due to Rapp (2000). Consider the four individuals from a segregating population of rats with genotypes shown in (1.1) and phenotype measurements. Let A, B, C, D, and E be loci on five different chromosomes that influence the phenotype. Knowing a quantitative phenotype of an individual does not yield unique information about the genotype at any given single locus, because the phenotype may be a net effect of the four loci. Assuming there are two alleles at each locus such that an allele with a subscript 1 (A_1, B_1, C_1, etc.) lowers the phenotype by 5 units of measurement (minus allele) and an allele with subscript 2 (A_2, B_2, C_2, etc.) increases the phenotype by 5 units (plus allele) around an overall mean of 150 units. As a first approximation we assume that the effects at all loci are additive. For example, individual 1 has eight subscripts 1, $8 * (-5) = -40$ units, and two subscripts 2, $2 * 5 = +10$ units, so the net effect of the plus and minus effects is $-40 + 10 = -30$ units about an overall mean of 150 units yields 120 units for individual 1. The phenotype for the other individuals is calculated similarly.

Individual	*Genotype*	*Phenotype*
1	$A_1A_1B_1B_2C_1C_1D_1D_2E_1E_1$	120
2	$A_2A_2B_1B_1C_1C_1D_1D_1E_1E_1$	120
3	$A_1A_2B_2B_2C_1C_2D_1D_1E_1E_2$	150
4	$A_1A_1B_2B_2C_2C_2D_2D_2E_2E_2$	180

(1.1)

Clearly, knowing that individuals 1 and 2 have the same phenotype does not define the genotype at, for example, QTL A which is A_1A_1 in individual 1 and A_2A_2 in individual 2. Similarly, individuals 1 and 4 have a markedly different phenotype

but have the same genotype, A_1A_1, at QTL A. Apart from these difficulties we do not know locations of QTL and can only observe genetic markers along the genome.

Microsatellite markers are commonly used and are inherited as simple Mendelian traits. It is logical to determine which genetic markers cosegregate with the phenotype to find the approximate location of the QTL. We now demonstrate how this cosegregation works.

Consider a marker locus M, which is a microsatellite marker locus on the same chromosome and closely linked to QTL A. Suppose the alleles M_1 and M_2 at locus M and alleles A_1 and A_2 at QTL A are organized as follows in two inbred parental strains P_1 and P_2. Let A1 be a minus allele that lowers the phenotype and A2 be a plus allele that increases the phenotype.

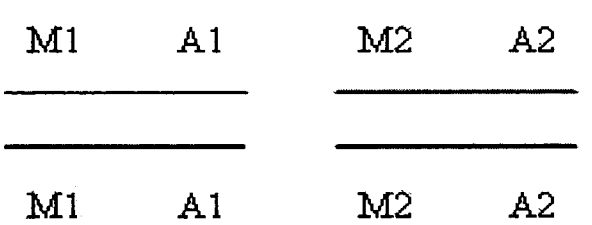


Figure 1.4: A marker linked to a QTL.

In Figure 1.4 each line represents one chromosome of a pair of chromosomes. Thus marker allele M_1 is linked to the QTL A minus allele A_1, and marker allele M_2 is linked to the QTL A plus allele A_2. Because P_1 and P_2 are inbred strains, both the marker and QTL loci are homozygous for their respective alleles.

P_1 is crossed to P_2 and the F_1 rats are intercrossed to produce a large F_2 population:

$$
\begin{array}{ccc}
P_1 & \times & P_2 \\
(M_1M_1) & \downarrow & (M_2M_2) \\
 & F_1 & \\
 & (M_1M_2) & \\
F_1 & \times & F_2 \\
(M_1M_2) & \downarrow & (M_1M_2) \\
 & F_2 & \\
(1M_1M_1: & 2M_1M_2 & : 1M_2M_2)
\end{array}
\tag{1.2}
$$

The F_2 rats are phenotyped and genotyped at the microsatellite marker locus using polymerase chain reaction (PCR). The marker alleles will segregate in Mendelian fashion $1M_1M_1 : 2M_1M_2 : 1M_2M_2$. Phenotypes of the rats in the three marker classes can be compared by a one-way ANOVA. In our example there will be phenotypical differences among the marker classes because the marker is linked to a QTL. In the extreme case where locus M is very closely linked to QTL A, the marker class M_1M_1 will consist of rats with genotype A_1A_1 at the QTL, M_1M_2 will represent rats with QTL genotype A_1A_2, and M_2M_2 will represent rats with QTL genotype A_2A_2. Because alleles at QTL on other chromosomes (loci B, C, D, and E in the example above) will assort at random with respect to M (and A), there will be no consistent enrichment of the alleles on other chromosomes in one marker class at M compared with another marker class at M, and the plus and minus phenotype effects at the QTL other than at locus A will cancel each other out. The method just described is called single marker association (SMA).

The segregation of alleles at the other QTL will, however, increase the phenotypical variance within a marker class at M, thus making the detection of the trait effects associated with M more difficult statistically.

As presented above, the marker M was very close to QTL A. In general, this will

not be the case. As the chromosomal map distance between M and A increases, the chances for recombination by chromosomal cross over between M and A increases. This will degrade the accuracy with which the genotype at M predicts the genotype at A, and consequently, the phenotypes cosegregating with M will decrease as the distance between M and A increases.

Some of the approaches presented below require a genetic map function, i.e. a relation $r = M(d)$ between recombination fractions r and genetic map distances d. The recombination fraction between two loci is the proportion of a chromosome that is recombinant. The genetic map distance between two loci is the average number of exchange points that occur along the chromosome between the loci. One of the widely used map functions is Haldane's map function (Haldane (1919)). It is a simple and easy to use function, however the Poisson process underlying this map function often does not fit recombination data (Speed (2005)). Assumption that the points of exchange form Poisson process leads implies no *crossover interference*, which is said to exist when the probability of having an exchange point in an interval depends on the occurrence of exchange points in another disjoint interval. Such interference is in general observed (Speed (2005)), but there are map functions used in genetics that account for interference (for example Kosambi map function). For a more detailed overview of map functions see Speed (2005).

1.2.2 Interval mapping (IM)

If phenotype linkage to a single marker has been found, a problem arises: where is the QTL in relation to the marker and how large is the effect of the QTL on the phenotype? A particular effect could arise from QTL very closely linked to the marker or it could arise from a QTL of larger effect at a further distance from the marker locus. The twofold problem of proving the existence of a QTL and of finding its map position has received a great deal of attention from statistical geneticists.

The advent of molecular genetic marker maps for various organisms produced by using molecular biology techniques has greatly facilitated the systematic mapping

and analysis of individual QTLs. Lander and Botstein (1989) proposed a method called interval mapping. They used one marker interval at a time to construct a putative QTL for testing by performing a likelihood ratio test (LRT) at every position in the interval. With a fine-scale genetic marker map throughout the genome, IM can be performed at any position covered by markers to produce a continuous LRT statistical profile along chromosomes. The position with the significantly largest LRT statistic in a chromosome region is an estimate of QTL position. Lander and Botstein (1989) only consider backcrosses for simplicity, and signified two inbred strains used for a backcross by A and B. Then, at a given marker, genotype AA means that both alleles were inherited from strain A, BB both from strain B, and AB one from A and the other from B.

Using Lander and Botstein's notation, let x_i mean an unknown genotype of individual i from a backcross population, at a given genetic location. The phenotype y_i and genotype are assumed to be related by the equation

$$y_i = \mu + \beta x_i + \epsilon, \tag{1.3}$$

where x_i is encoded as a (0,1) indicator variable equal to the number of alleles that increase the phenotype, $\epsilon \sim N(0, \sigma^2)$; μ, β and σ^2 are unknown parameters. The likelihood function proposed in IM is

$$L(\mu, \beta, \sigma^2) = \prod_i^n [G_i(0)L_i(0) + G_i(1)L_i(1)], \tag{1.4}$$

where $L_i = z((y_i - (\mu + \beta x)), \sigma^2)$ is the likelihood function for the individual i, $z \sim N(0, \sigma^2)$, assuming that $x_i = g$ and $G_i(g)$ denotes the probability that $x_i = g$ conditional on the genotypes and positions of the flanking markers. For computing $G_i(g)$ one need to define a map function. Lander and Bonstein (1989) did not assume any specific function in the general description of the method. However,

they provided one example of computing $G_i(g)$ for a map function that assumes no interference: if at both markers the genotypes are AA and recombination fractions between the putative QTL and the markers are θ_1 and θ_2, then the probability of the QTL genotype being AB is $\theta_1\theta_2$.

Then finding the maximum likelihood solution for parameters μ, β and σ^2 from (1.4) can be regarded as a linear regression problem with missing data: we only have probability distributions for the missing genotypes. To find solutions the authors proposed using the EM algorithm, implemented in their software package MAPMAKER-QTL. This approach also applies directly to F_2 crosses and recombinant inbred strains with necessary modifications provided in Lander and Botstein (1989).

The LOD scores mentioned above are computed as follows. We find MLEs from the full model, $(\hat{\mu}, \hat{\beta}, \hat{\sigma^2})$ and the constraint model with $(\beta = 0)$, $(\hat{\mu_0}, \hat{\beta}_0, \hat{\sigma_0^2})$. Then

$$LOD = log_{10} \frac{L(\hat{\mu}, \hat{\beta}, \hat{\sigma^2})}{L(\hat{\mu_0}, \hat{\beta}_0, \hat{\sigma_0^2})}. \tag{1.5}$$

Even though it models one gene at a time, this approach has been successfully applied to detect multiple genes that underlie a wide range of complex and quantitative traits (Sen and Churchill (2001)).

1.2.3 Mapping multiple QTLs

The approach of IM considers one QTL at a time in the model for QTL mapping. Therefore, IM can bias identification and estimation of QTL when multiple QTL are linked. Multiple QTL models are an improvement over single QTL models because of their ability to separate linked QTL on the same chromosome and to detect interacting QTL that may otherwise be undetected. Such methods provide increased power to detect QTL and can eliminate biases in estimates of effect size and location that can be introduced by using an inappropriate single QTL model. Jansen (1993) provides examples of more efficient detection and more accurate mapping resulting

from using multiple QTLs.

Many approaches have been suggested for mapping multiple QTL. Jansen (1993) and Zeng (1993a) independently proposed the idea of combining IM with multiple regression analysis in mapping. They both attempt to reduce the multidimensional search for QTL to a series of one-dimensional searches. This is achieved by conditioning on markers outside a region of interest to account for the effects of other QTL. Zeng named his method composite interval mapping (CIM), and Jansen's method is called multiple-QTL mapping (MQM). Multiple interval mapping (MIM) proposed by Kao *et al.* (2001) extend interval mapping directly to the case of multiple QTL.

1.2.4 Composite interval mapping (CIM)

The method attempts to increase reliability and accuracy of QTL mapping by adequately separating the effects of multiple linked QTL in testing and estimation. The gist of the method is to perform an interval test, in which a test statistic is constructed to be unaffected by QTLs located outside a defined interval. It is achieved by using the properties of multiple regression analysis described in Zeng (1993b). Even though multiple regression analysis has been used before for mapping QTL, its theoretical properties in relation to QTL mapping were only analysed in that paper and independently by Rodolphe and Lefort (1993). Most importantly, Zeng proposed using these properties to construct a new method that improves accuracy. We briefly describe these properties and then the method itself.

Consider a sample of n individuals from a backcross population with recorded quantitative phenotypes. Suppose we have m ordered markers and that we use the following model for our analysis:

$$y_i = \mu + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i, \tag{1.6}$$

where x_{ij} is the type of the jth marker in the ith individual. In the backcross design they can be of two types: with both alleles from the strain to which we

backcross, or one allele from each line. β_j is the partial regression coefficient of phenotype y_i on the j-th marker conditional on other markers. Zeng established a number of properties of the multiple regression model, which he used to construct his test. One of the most important of them is that if we assume additivity of QTL effects between loci, the expected partial regression coefficient does not depend on QTLs located outside the interval bracketed by the two neighbouring markers. It only depends on the QTLs within the interval. This property allows one to construct a conditional interval test based on the partial regression coefficient.

Zeng (1993b) showed that the coefficients in (1.6) are biased estimates of the relevant QTL effect and proposed using the model (for the marker interval $(k, k+1)$):

$$y_i = \mu + \beta^* x_i^* + \sum_{j \neq k, k+1} \beta_j x_{ij} + \epsilon_i. \tag{1.7}$$

Markers k and $k+1$ are used as indicators of the genotype of the putative QTL within the interval and we test for a QTL on a marker interval $(k, k+1)$. x_i^* is an indicator variable taking values 1 or 0 with probability depending on the position being tested and genotypes of markers k and $k+1$; β^* is the effect of the putative QTL, which is the difference of effects between homozygote and heterzygote; x_{ij} are known covariates for the j-th marker; β_j are partial regression coefficients of phenotype y on the j-th marker conditional on all other markers.

The likelihood function is

$$L_1 = \prod_{i=1}^{n} [p_i(1) f_i(1) + p_i(0) f_i(0)] \tag{1.8}$$

where $p_i(1)$ is a prior probability of $x_i^* = 1$ and $f_i(1)$ and $f_i(0)$ is a normal density function for y_i with means $\mu + \beta^* + \sum_{j \neq k, k+1} \beta_j x_{ij}$ and $\mu + \sum_{j \neq k, k+1} \beta_j x_{ij}$. The probability for assigning values of x_i^* depend on the recombination fractions between marker k and the putative QTL, and between markers k and $k+1$. Double

recombination within the marker interval is ignored, and there is no assumed map function (see Table 1 in Zeng (1993b) for more details).

Maximum likelihood estimates of the parameters are found by differentiating the function and setting the derivatives equal to zero:

$$\begin{aligned} \hat{\beta}^* &= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'\hat{\mathbf{P}}/\hat{c} \\ \hat{\mathbf{B}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{P}}\hat{\beta}^*) \\ \hat{\sigma^2} &= [(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) - c\hat{\beta}^{*2}]/n, \end{aligned} \tag{1.9}$$

where $\hat{\mathbf{B}}$ is a vector of ML estimates of β_j's, $\mathbf{X}$ is a matrix of x_{ij}'s, $\hat{\mathbf{P}}$ is a vector of ML estimates of the posterior probability, with elements $\hat{P}_i$ specifying maximum likelihood estimates of the posterior probability of $x_i^* = 1$:

$$\hat{P}_i = (p_i(1)\hat{f}_i(1)/[p_i(1)\hat{f}_i(1) + p_i(0)\hat{f}_i(0)], \tag{1.10}$$

$\hat{c} = \sum_{i=1}^{n} \hat{P}_i$. Zeng (1993b) also provides a formula and a table of values for ML estimates of p_i via $\hat{P}_i$.

Estimates of the parameters in 1.9 can be found via the expectation/ conditional maximisation (ECM) algorithm (Meng and Rubin (1993)). It consists of one E-step, equation 1.10, and three CM-steps which are the three equations in 1.9. The advantage of this algorithm over the full EM algorithm (simultaneous maximisation of $\hat{\beta}^*$ and $\hat{\mathbf{B}}$ in the M-step) is that $\mathbf{X}'\mathbf{X}$ does not need to be updated (Zeng (1993b)).

We test $H_0 : \beta_* = 0$ versus $H_1 : \beta_* \neq 0$. The likelihood function for the null hypothesis that we use for the LR test statistic is

$$L_0 = \prod_{i=1}^{n} f_i(0). \tag{1.11}$$

Property 3 of the multiple regression model given in Zeng (1993b) states that conditioning on *linked* markers in the multiple regression analysis may increase sample variance and entail a loss in the statistical power. Hence the method is a trade-off between precision and efficiency of mapping, and in practical applications of the method finding a balance is a major consideration. Epistasis is ignored by the method.

1.2.5 Multiple-QTL mapping (MQM)

Jansen (1993) developed a method where we model the simultaneous likelihood of the trait and of multiple markers and QTLs. Genotypes in the method are multilocus. Their probabilities are unconditional, which is in contrast to most methods for inbred line analysis. In interval mapping, for example, one usually calculates the likelihood of the trait conditional on the two flanking markers. If the marker genotypes can be missing then we use the nearest informative marker in the same direction. Such an approach is good when there is one QTL but works less well for multiple-QTL models. The authors claim that only using the unconditional multilocus likelihoods can form the basis for direct comparison of the fit of several competing models.

The method is best demonstrated by an example which we take from Jansen (2001). Consider an F_2 population and five loci, with measurements $AADUA$, where A and B define allele from the first and second strains respectively, D means "not B", U is unknown genotype (for all cases). Assume that the 4th loci is a QTL with additive allele effect and the rest are markers; all loci are located on a linkage map. The idea then is to complete the observed data for the missing locus information. This gives us 6 different complete genotypes with corresponding component densities:

$$
\begin{array}{ll}
AAAAA & f_A(y_i) \\
AAAHA & f_H(y_i) \\
AAABA & f_B(y_i) \\
AAHAA & f_A(y_i) \\
AAHHA & f_H(y_i) \\
AAHBA & f_B(y_i)
\end{array} \tag{1.12}
$$

where H is a heterozygous locus. The component densities to each multilocus genotype can be defined, taking into account that at the moment we have only one explanatory variable - the 4th locus:

$$
\begin{aligned}
f_A(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i-\mu_A)^2}{2\sigma^2}) \\
f_B(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i-\mu_B)^2}{2\sigma^2}) \\
f_H(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y_i-\mu_H)^2}{2\sigma^2}) \\
\mu_H &= 0.5(\mu_A+\mu_B).
\end{aligned} \tag{1.13}
$$

If we know the frequencies of recombination between the loci, q, r, s, t, we can calculate multilocus genotype probabilities. For instance, for the first genotype it is $1/4(1-q)^2(1-r)^2(1-s)^2(1-t)^2$: $1/4$ for genotype A at the first locus, times $(1-q)^2$ for no recombination between the first locus and the second locus, etc. Now we can construct mixture density for the i-th individual

$$
\begin{aligned}
f_{mix}(y_i, AADUA) &= [P(AAAAA) + P(AAHAA)]f_A(y_i) \\
&+ [P(AAAHA) + P(AAHHA)]f_H(y_i) \\
&+ [P(AAABA) + P(AAHBA)]f_B(y_i) \qquad (1.14)
\end{aligned}
$$

Then the simultaneous mixture likelihood of all observations is

$$
\prod_{i=1}^{n} f_{mix}(y_i, AADUA). \qquad (1.15)
$$

The dataset is augmented as illustrated below, where for each individual in the dataset we write down three components of the likelihood. The model then becomes the standard linear weighted regression, which can be written in matrix notation for the i-th individual:

$$
\begin{pmatrix} y_i \\ y_i \\ y_i \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix}, weights \begin{pmatrix} P(A|y_i) \\ P(H|y_i) \\ P(B|y_i) \end{pmatrix}, \qquad (1.16)
$$

where $P(A|y_i)$ denotes probability of allele A at the fourth loci, which is a QTL. The model can be extended with extra QTLs in the marker interval. For example, we can add an extra locus with U score, then the probability density function for the i-th individual becomes:

$$
f_{mix}(y_i, AUADUA) = \sum_{\substack{u=A,H,B \\ v=A,H,B}} [P(AuAAAvA) + P(AuAAHvA)]f_{uv}(y_i). \qquad (1.17)
$$

For all individuals, and for any number of QTLs and markers, the general regression model is $y^{(a)} = X^{(a)}\beta + e^{(a)}$ with the weight matrix $W^{(a)}$ (superscript (a)

stresses that the matrices are augmented). The parameters are estimated via the EM algorithm:

- E-step. Update weights using Bayes formula

$$P(A|y_i) = \frac{[P(AAAAA) + P(AAHAA)]f_A(y_i)}{f_{mix}(y_i)}. \tag{1.18}$$

- M-step. Update estimates of β and σ^2 using weighted regression formulae

$$\begin{aligned} \hat{\beta} &= (X'WX)^{-1}X'WY \\ \hat{\sigma^2} &= \frac{1}{n}(Y - X\hat{\beta})'W(Y - X\hat{\beta}). \end{aligned} \tag{1.19}$$

Data augmentation in the example above gave rise to six different complete genotypes. The approach allows multiplying other individuals by any other value, depending on how much genetic information is missing.

For many QTLs the method becomes computationally unfeasible.

The methods of Zeng (1993a) and Jansen (1993) both proposed procedures that combine interval mapping with multiple regression. According to Zeng (1993a), which was published after Jansen (1993), there are some critical differences between the two: the emphasis in Zeng (1993a) is to control the precision of mapping as much as possible; Jansen (1993) tests multiple markers simultaneously to indicate possible multiple QTLs; Zeng (1993a) provides theoretical and statistical properties of the method.

1.2.6 Multiple interval mapping (MIM)

The methods extends the previous QTL mapping techniques to a multiple QTL model in a way that the trait loci can be directly controlled in the model. MIM uses multiple marker intervals simultaneously to construct multiple putative QTL. Compared to earlier methods such as IM and CIM, MIM is seen as more powerful and precise in detecting QTL as demonstrated by an example in Kao *et al.* (1999). The

method allows to analyse epistatic QTL and estimate individual genotypic values. Also, Mayer (2005) carried out an extensive comparative analysis of MIM with *regression mapping* (described later in Section 1.2.7). The analysis showed that for *linked QTL* MIM generally outperformed regression mapping with regard to power, accuracy of position, effect estimates, and estimation of the residual variance.

The main ideas of the method are: to separate choosing intervals with putative QTL from parameter estimation; to use a joint probability for multiple QTL and then factorise; to use Cockerham's genetic model to model epistases; treat the mixture likelihood as a missing data problem.

Let us go through the first step, parameter estimation. Consider m intervals I_j bounded by markers. Each interval contains a QTL at an unknown position p_j. We first show the likelihood function for the MIM method and then explain how it is constructed:

$$L(\theta|Y,X) = \prod_{i=1}^{n} \left[\sum_{j=1}^{2^m} p_{ij} \phi(\frac{y_i - \mu_{ij}}{\sigma}) \right], \tag{1.20}$$

where ϕ is standard normal density, Y are phenotypes, p_{ij} are the conditional probabilities of QTL genotypes (not to confuse with QTL positions p_j). X, the design matrix, and μ_{ij} are described later. The summation is done over 2^m possible multiple genotypes of the QTLs. There are only two possible genotypes at each position because we consider a backcross population. Mixing proportions p_{ij} are probabilities of each multilocus genotype calculated by using the factorised joint multi-QTL probability distribution. If there is no crossing-over interference the conditional distributions of the individual putative QTL genotypes, given the flanking marker genotypes, are independent:

$$Pr(Q_1, ..., Q_m|I_1, ..., I_m) = \prod_{j=1}^{m} P(Q_j|I_j), \tag{1.21}$$

where Q_j is the genotype of the putative QTL in interval I_j.

The conditioning is on the marker genotypes that bracket the intervals. Each probability in the product depends on the locations p_j and is calculated using Haldane's map function (Haldane (1919)).

Now let us consider variables x_{ij}^s instead of the QTL genotypes Q_j to explain how the mean values μ_{is} change. Coded variable x_{ij}^s takes values $1/2$ and $-1/2$ for QTL genotypes QQ and Qq at interval j, for individual i and corresponding to the multilocus QTL genotype combination s. The mean values in each normal density of (1.20), and therefore for each combination of QTL genotypes, is constructed using Cockerham's genetic model (Cockerham (1954)). The model takes two loci affecting the phenotype and breaks down the phenotypic variance into additive, dominance and epistatic parts, which is suitable for MIM that only considers interactions between two genes. Then the genetic model links main gene effect parameters, β_j, and epistatic effects w_{jk} with the mean value of the phenotype for a particular combination of QTL genotypes:

$$\mu_{is} = \mu + \sum_{j=1}^{m} \beta_j x_{ij}^s + \sum_{j,k}^{m} \delta_{jk}(w_{jk} x_{ij}^s x_{ik}^s), \tag{1.22}$$

δ_{jk} is an indicator variable for epistasis (interaction) between QTL alleles Q_j and Q_k; it is equal 1 when there is epistasis between Q_j and Q_k and $j < k$, otherwise it is equal zero (so that interaction of a gene with itself is excluded, and interactions are not included twice in the summation). The method estimates parameters $\theta = (p_i, ..., p_m, \beta_1, ..., \beta_m, ..., w_{jk}, ..., \sigma)$ via the EM algorithm. Note that the number of components in the mixture model and the number of parameters can be quite large , for $m = 10$ it is 1024 normals and 22 parameters to estimate. To solve the matter the authors propose using the general formulae for finding MLEs and their covariance matrix they derived in another paper, Kao *et al.* (1997).

The second part of the method addresses the problem of finding QTLs to fit in the model. The model selection is done using LR test for hypotheses: $H_0 : \beta_j = 0$

and $H_1 : \beta_j \neq 0$. There is an open issue in the method regarding the critical value of the test statistic to use in the model selection. The authors point out that the issue is complex even for a single QTL model and that for multiple QTL models the critical value also depends on the true model. They propose to use Bonferroni's method (Bonferroni (1936)) until the issue is resolved.

It is proposed to use a 3 steps procedure:

- Step 1:

  Calculate significant values of LRT for entry (SVE) and staying (SVS).

- Step 2:

  For each position within the region covered by markers we calculate LRT and add those that have the statistic exceeding SVE.

- Step 3:

  After k QTL are added to the model we consider models with $k+1$ QTL and select the position with the most significant partial LRT at the SVE level. After $k+1$ QTL are fitted we delete any QTL with LRT lower than SVS.

1.2.7 Regression mapping

Regression mapping was independently developed by Knapp *et al.* (1990), Haley and Knott (1992) and Martinez and Curnow (1992). The method approximates interval mapping very well but requires less computation (Broman and Speed (1999)). Consider a backcross population where the inbred line A has a lower quantitative phenotype than line B : $\mu_A < \mu_B$. Two markers are separated by recombination fraction r, and the putative QTL is at recombination fraction r_l from the left marker. The conditional expected value of the phenotype for an individual given the genotypes at the flanking markers is

$$E(y|x_j, x_{j+1}) = \mu_A + (\mu_B - \mu_A)Pr(Q_j = AB|x_j, x_{j+1}), \tag{1.23}$$

where x_j, x_{j+1} are marker genotypes of the interval, and $Pr(Q_j = AB|x_j, x_{j+1})$

is calculated from the given recombination fractions (for example, using Table 1 in (Broman and Speed (1999)). So in this approach we regress phenotypes on conditional probability of having the genotype AB at the putative QTL, given the marker genotypes. Phenotypes are taken as normally distributed with the mean given in (1.23) and unknown variance, which leads to the LOD score

$$LOD = \frac{N}{2} log_{10} \left(\frac{RSS_0}{RSS} \right), \tag{1.24}$$

where N is the sample size, RSS is the residual sum of squares, $\sum_i (y_i - \hat{y}_i)^2$, and RSS_0 is the residual sum of squares under the null hypothesis, $\sum_i (y_i - \bar{y}_i)^2$. Similarly to the interval mapping, at each locus in the genome we calculate the LOD score, but we only need to estimate a single regression at each locus, instead of running the EM algorithm at each location.

1.2.8 Bayesian methods

The Bayesian framework provides a clear picture of the probabilistic structure of the QTL mapping problem. Unknown quantities, such as QTL location, QTL genotype, phenotypic mean and variance can all be quantified via probability. In this section we give a brief review of the published methods and describe on of them on more details, closely following Sillanpaa and Arjas (1998).

Bayesian methods have been applied to gene mapping by Tai (1989), Thomas and Cortessis (1992), Smith and Roberts (1993), and Stephens and Smith (1993). Satagopan et al. (1996) put forward a Bayesian QTL mapping method, in which a known number of QTLs was presumed to be present in the chromosome under consideration. However, this method failed to take into account the effects in other chromosomes of QTLs. The actual number of QTLs was inferred from separate MCMC runs (Metropolis *et al.* (1953)) each containing different numbers of QTLs.

Stephens and Fisch (1996) and Satagopan and Yandell (1996) considered all chromosomes simultaneously, and treated the number of QTLs as a random variable. Stephens and Fisch (1996) used Metropolis-Hastings (M-H) scheme (Casella

and George (1992)), whereas Satagopan and Yandell (1996) applied Gibbs sampler (Geman and Geman (1984)) in estimation. Stephens and Fisch (1996) simulated six datasets of 10 chromosomes each, with a different heritability in each, and tested several different priors. In contrast to Sillanpaa and Arjas (1998), missing data was not considered.

Here we provide more details of one of the Bayesian methods proposed by Sillanpaa and Arjas (1998). The method can be applied to inbred line crosses and normally distributed phenotypes, and can handle incomplete data, similarly to our method we outline later. It falls into a framework of variable dimensional Bayesian models and uses MCMC algorithm to obtain the posterior distribution of QTLs and their locations.

The approach of Sillanpaa and Arjas (1998) bears some similarity with the CIM method of Zeng (1993) in that it incorporates information of QTLs on other chromosomes. The differences of this Bayesian method from the frequentist methods presented above lie in how the results of inference are quantified: they are summarised by two measures, the posterior QTL-intensity, considered as a function of its location, and the posterior distribution of the phenotypic effect of the corresponding putative QTL. This probabilistic summary seem to correspond to objectives of QTL mapping, localising QTLs and estimating their effect on the phenotype.

The parametrisation of the method is quite complicated. To convey the idea of the method we assume for now that there are no missing marker genotypes. The method uses an idea similar to composite interval mapping to account for the effects of QTLs on other chromosomes. Figure (1.5) shows the chromosome of interest, Chromosome 1, which contains a QTL Q_1, and Chromosome 2 that has markers affecting the phenotype. Vector $(g_1, ..., g_{N_{Gen}})$ includes all possible genotypes at any locus, QTL or marker, N_{Gen} being the number of possible genotypes. Parameters β's and c's are regression coefficients for QTLs and *background controls*, which are a selected set of markers that are hoped to be close to influential QTLs outside the interval. Sillanpaa and Arjas (1998) discusses various options for choosing the

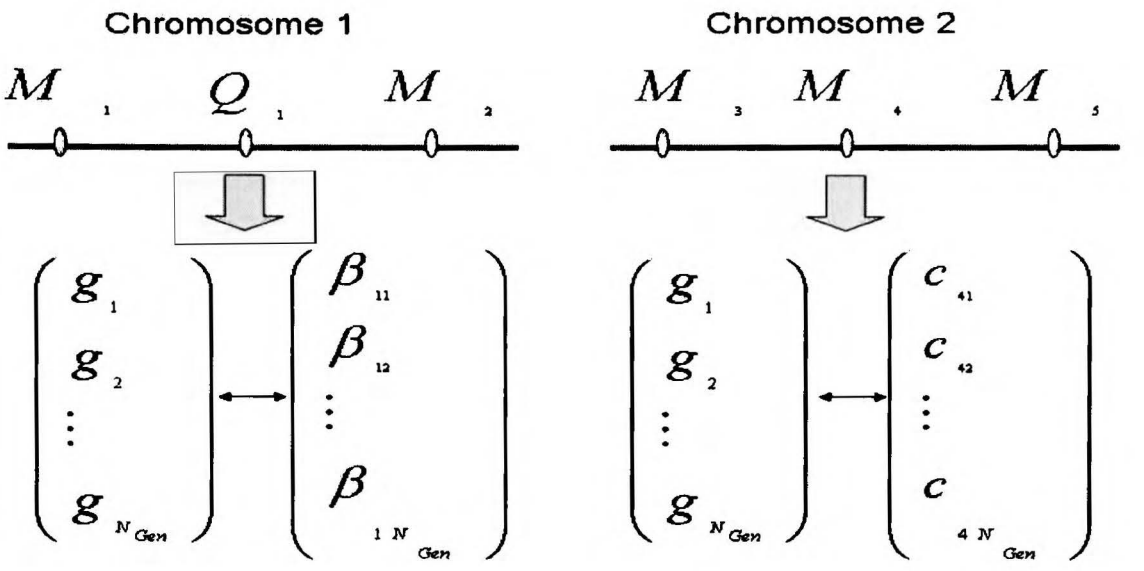


Figure 1.5: Illustration to the statistical model (1.25). We analyse QTL Q_1 on Chromosome 1 and the background controls are the markers on Chromosome 2. The graph shows all possible genotypes g_j and their corresponding effects on the phenotype. All elements of vectors β and c except one are zeros.

markers: preliminary stepwise regression analysis, use Bayesian QTL analysis of other chromosomes, choose from a set of candidate genes if they are available, or consider the entire genome in a single variable dimensional QTL analysis using the "current" QTLs as controls.

The method's statistical model is defined as:

$$\begin{aligned} y_i = \mu \quad &+ \sum_{q=1}^{N_{qtl}} \sum_{j=1}^{N_{Gen}} \beta_{qj} \mathbf{1}(x_{iq}^{Q} = g_j) \\ &+ \sum_{k=1}^{N_{BC}} \sum_{j=1}^{N_{Gen}} c_{kj} \mathbf{1}(x_{ik}^{BC} = g_j) + \epsilon_i, \end{aligned} \tag{1.25}$$

where μ is the intercept, $\epsilon_i \sim N(0, \sigma^2)$, x_{iq}^{Q} are elements from the $N_{ind} \times N_{qtl}$ matrix of QTL genotypes, X^Q, at QTL locations $l = (l_1, ..., l_{Nqtl})$. Elements from the second summation, x_{ik}^{BC}, are from the $N_{ind} \times N_{BC}$ matrix of background controls, X^{BC}. Note that N_{qtl} is an unobserved random variable, N_{ind}, N_{BC} and N_{Gen} are known. Also, $\mathbf{1}(x_{iq}^{Q} = g_j)$ is an indicator variable, taking value one if the genotype of the i-th individual at QTL q is g_j.

We can group the model's parameters as follows: the regression parameters are

denoted $\delta = (\mu, \beta, \sigma^2, c)$. Then all the unknown quantities are $\theta = (\delta, X^Q, l, N_{qtl})$. We further assume that we have a fixed marker map m with known recombination fractions between markers. Let us further define G as marker data on the chromosome of interest. Omitting some of the algebra which can be found in the original article, the posterior probability of the unknown quantities can be shown to equal:

$$P(\theta|y, X^{BC}, G, m) \propto P(y|\theta, m)P(\theta|m), \tag{1.26}$$

where

$$\begin{aligned} P(y|\theta, m) &= \prod_{i=1}^{N_{ind}} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big[-\frac{1}{2\sigma^2}\Big(y_i - \mu - \sum_{q=1}^{N_{qtl}} \sum_{j=1}^{N_{Gen}} \beta_{qj} \mathbf{1}(x_{iq}^Q = g_j) \\ &- \sum_{k=1}^{N_{BC}} \sum_{j=1}^{N_{Gen}} c_{kj} \mathbf{1}(x_{ik}^{BC} = g_j)\Big)^2\Big]. \end{aligned} \tag{1.27}$$

The joint prior distribution $P(\theta|m)$ conditioned on the marker map, can be factorised if we make the following assumptions: given N_{qtl}, parameters of the model (1.25), δ, are independent of the other parameters of θ, i.e. X^Q, l;

$$\begin{aligned} P(\theta|m) &= P(\delta, X^Q, l|N_{qtl}, m)P(N_{qtl}|m) \\ &= P(\delta|N_{qtl}, m)P(X^Q, l|N_{qtl}, m)P(N_{qtl}|m) \\ &= P(\delta|N_{qtl}, m)P(X^Q|G, l, N_{qtl}, m)P(l|N_{qtl}, m)P(N_{qtl}|m). \end{aligned} \tag{1.28}$$

Prior distributions

Depending on the goals of the study, priors can be specified to be "uninformative" if the goal is to present the information present in the data, or can represent experts' prior knowledge if it is available. In the latter case the posterior will be a synthesis of the experts' knowledge and evidence in the data.

It is important to note that the number of possible alleles and genotypes (N_{Gen}) and therefore the prior density $P(X^Q|G, l, N_{qtl}, m)$ depend on the experimental design. For example, in an F_2 design there are only three possible genotypes at each locus.

For computational reasons prior distribution for N_{qtl} is assumed to be truncated Poisson with $0 \leq \lambda \leq N_{QTLmax}$, where the upper bound N_{QTLmax} is introduced for computational reasons, and λ is a fixed control parameter. If there is now prior knowledge available, $P(l|N_{qtl}, m)$ and $P(\delta|N_{qtl}, m)$ are set to be uniformly distributed.

Prior probabilities of QTL genotypes can be further factorised

$$\begin{aligned} P(X^Q|G, l, N_{qtl}, m) &= \prod_{q=1}^{N_{qtl}} p(x_q^Q|x_i, ..., x_{q-1} l, m) \\ &= \prod_{q=1}^{N_{qtl}} \prod_{i=1}^{N_{ind}} p(x_{iq}^Q|g_{iq}^L, g_{iq}^R, r_q), \end{aligned} \tag{1.29}$$

where g_{iq}^L is genotype of the marker on the left and r_q is recombination fraction between flanking markers and the QTL location, which can be calculated using Haldane's formula. Appendices $B1$ and $B2$ of Sillanpaa and Arjas (1998) list two algorithms (for backcross and F_2 intercross) for constructing the probabilities of different QTL genotypes in (1.29).

Conclusion on QTL mapping methods

In this chapter we gave brief descriptions of methods commonly used in QTL mapping, starting from the basic and moving on to more complicated methods. Most of them were developed for and applied to F_2 and backcross populations. As mentioned in Section 1.1.1, many QTL regions have been identified but not as many genes. One of the reasons for the small number of discovered genes is poor resolution in F_2 and backcross studies. HS populations with corresponding methods can help resolve the issue of resolution. The next chapter introduce models used for QTL mapping in

HS. This is not to say that the previous methods are of no use for the new mouse stocks. Jansen's MQM method is very similar to the method we propose in Chapter 2 and the Bayesian model just described can probably be applied to heterogeneous stocks analysis.

1.3 Genome-wide association studies

In this section we provide background for the research presented in Chapter 3 and Chapter 4, including basic concepts of genetic epidemiology and population association studies.

1.3.1 Study designs

Let us begin with some brief definitions of the basic concepts. In epidemiology *incidence* means the rate of becoming diseased of an initially disease-free population. In contrast, *prevalence* is defined as the proportion of diseased at a given point in time and combines incidence, the average duration of the disease and survival rate. Thus prevalence should be used with care as there might be risk factors affecting fatality of cases but without influence on incidence. *Population at risk* is defined as a group who would be counted as cases if they had the disease under consideration.

Let us consider two very popular study designs used in statistical epidemiology. *Cohort* design is of fundamental importance and is very important conceptually. In this design, researchers identify people who are at risk but currently are free of disease. The subjects are followed over time to establish which of them develop the disease of interest. Then the relationship between the risk factors measured before the experiment and the risk of developing the disease are estimated. Cohort design is considered to be less prone to biases such as selection and information bias, which the case-control design is more subject to. For this reason cohort design is often preferred to the case-control design. However, cohort studies generally require very large groups of people and many years of observation for rare late onset diseases.

Breslow (1996) in his review article points out that that the contributions made

by statisticians to the development of *case-control* methodology in the second half of the 20-th century were the most important among their other contributions made to public health and biomedicine. In the early days of applying this design it was believed that it does not provide accurate information about the rate of diseases but then Cornfield (1951) showed that, under the rare disease assumption, the odds ratio of exposure is equal to the disease odds ratio and that in turn approximates relative risk.

The method randomly selects a representative sample of diseased individuals from the general population. Then, depending on the variation of the method (a few are described below), a sample of disease-free people is drawn. This is followed by an inquiry into the past history of cases and controls in order to establish possible reasons for different outcomes. The popular variants of the case-control design are *nested case-control design* and *case-base design.*

In the nested case-control design controls are individually matched to each case. Matching is done via random sampling from a group of selected individuals who are at risk at the time when the case occurred. This design allows to sample individuals as controls for more than one case. Also, a case can serve as a control for earlier cases (Thomas (2004)). Let us define *relative risk* as the ratio of *the risk in exposed subjects* (the probability of developing the disease if the individual is exposed to the environmental factors studied) to the risk in unexposed individuals. Lubin and Gail (1984) demonstrated that the alternative of excluding cases from the possibility of serving as controls would lead to biased estimates of relative risk.

The case-base design is sometimes referred to as the case-cohort design (Thomas (2004)) and is a scheme where controls are a random sample from the entire cohort at the time of enrollment to the study, regardless of whether they become later cases or not. The advantages of the case-base approach are that the same controls can be used for comparison with multiple case groups; also, the baseline data on controls can be collected early in the study, at the same time as cases are appearing. However, analysing case-base dataset requires more complex analysis.

The case-base (or case-cohort) design was used recently in the very prominent Wellcome Trust Case Control Consortium (2007) study recently, where a shared group of 3,000 controls taken from the whole population, and which potentially could be cases, was used for the analysis of all 7 common diseases.

In epidemiology, one of the main concerns in making inference is freedom from bias. *Selection bias* refers to aspects of the design that would make the sample not representative of the population it was selected from (for example, sampling controls from a hospital population to represent the population of unaffected subjects). *Information bias* is seen to be present when the quality of information about the individuals is not comparable between the groups (e.g. *recall bias* in case-control design when cases recall past exposures differently from healthy subjects). Since the information about the subjects is collected prior to the study, cohort design is believed to be less susceptible to the information bias (Thomas (2004)). Also, in the cohort design individuals are followed in the same way after enrollment. Information bias can occur in cohort studies though. For instance, sometimes subjects are lost for follow-up, and the probability of being lost may depend on the exposure of disease status.

In the context of exploring the genetic basis of diseases, concerns about recall bias are generally dismissed because establishing the genetic variant, which is the key exposure, does not rely on recall. However, there is still a possibility of bias in selection of cases and controls. Manolio *et al.* (2006) provides a thorough comparison of potential for various biases in different study designs. Among other conclusions, they state that the case-control design has particularly important advantages for studying rare diseases, although they still consider prospective cohort studies indispensable.

The purpose of population association studies is to find patterns of variation that change between individuals with different disease phenotypes and could represent the effects of deleterious or protective alleles. Since the genome is very large spurious associations can arise randomly. Therefore there is a need for reliable significance testing procedures which would help to identify true effects.

Linkage studies have been one of the most popular techniques preceding GWAS and are still widely used. Linkage studies follow the transmission of markers across generations within a pedigree, trying to establish correlations which imply linkage of the marker with a causal locus. Bingham and Riboli (2004) describe parametric linkage analysis where a likelihood-based model of marker transmission is used. In contrast, nonparametric linkage analysis seeks to identify differences in allele sharing between affected and healthy relatives and does not specify a model (Holmans (2003)). Information across families is combined in linkage methods so that it can accumulate even if different polymorphisms that segregate in different families are not the same. Thus linkage analysis can identify rare variants that contribute little to disease risk. The disadvantage of these methods is the need for combining many families in a study.

Ewens and Spielman (2003) considered an alternative method to test for linkage with a genetic marker when population association has been found, and termed it transmission/disequilibrium test (TDT). This method implements matched-pair design which compares alleles transmitted to a diseased individuals with untransmitted parents' alleles.

As linkage analysis, admixture mapping (Smith and O'Brien (2005)) is a nonparametric approach. For a group formed by recent admixture of two or more populations with very different disease prevalences, the case-only design is the most powerful for low prevalences (Hoggart *et al.* (2004)). A higher frequency of an allele that is more common in the founder population with a higher disease prevalence could mean that the allele is disease-predisposing. The model presented in Chapter 3 is applicable to admixture mapping, the null value of the model's parameter being the proportion of admixture.

One of the most popular types of studies is a "candidate gene" study, when a gene is identified first from a previous linkage analysis or based on homology with another gene with known function in model animals (e.g. mice). Then from 5 to 50 SNPs within the gene are typed.

Even though the statistical methods for candidate gene and GWA studies are similar there are a number of issues specific to GWAS. For example, typing all known variants on the whole genome is not feasible at present despite the recent rapid reduction in cost. Only about 300,000 or so SNPs are required to represent most of the most of the common genetic variation in Caucasians but this is still financially prohibitive for many researchers (Balding (2006)). For this reason a two stage approach is often adopted when a relatively small number of individuals are typed genome-wide and then many more for promising regions on the genome. Computational problems with analysing large datasets arising in such studies is another issue and the reason why single SNP analyses are still the main tool in GWAS. All the models presented in this work are for a single SNP analysis.

There are more than 10 million common human polymorphisms and only few of them are usually typed in a study. When a causal variant is not typed it can still be detected if there is *linkage disequilibrium* (LD) between the causal variant and a genotyped polymorphism. Therefore LD remains essential for the design of association studies while whole-genome sequencing is still not available.

Careful selection of SNPs for analysis is very important and "tagging" is often used to choose SNPs. There are several methods to select tag SNPs and they all attempt to exploit redundancy among SNPs, to minimise loss of information, while maximising efficiency in the laboratory (International HapMap Consortium (2005)). Some methods use a single SNP to be a proxy for others, which are untyped, whereas in other methods untyped SNPs are represented by combinations of alleles (haplotypes). Also, methods differ depending on whether or not they use LD blocks. One of the more sophisticated methods for selecting tag SNPs in described in Zeggini *et al.* (2005), a simpler method being leaving only one SNP from a pair with LD higher than a threshold value.

We assume throughout this chapter that the population from which samples are drawn is in Hardy-Weinberg equilibrium (HWE). However, as part of preliminary analysis a test for the equilibrium should be performed because it has been

pointed out that deviations from HWE can be due to processes important in disease causation (see Conrad *et al.* (2006)).

Allele frequencies

Minor allele frequency (MAF) means the lowest allele frequency observed at a locus, in a particular population. For single nucleotide polymorphisms (SNPs) it simply means the lesser of the two allele frequencies. We cover a broad range of minor allele frequencies in our investigation in the chapters that follow, and frequencies within this range have different implications for GWAS. There are two widely used hypotheses explaining the relationship between diseases and allele variants, the common-disease/common-variant hypothesis (CDCV) and the rare variant model (Zondervan and Cardon (2004)). According to CDCV complex traits are associated with alleles that have relatively high frequencies, larger than 0.1. Those variants have not been subject to high selection pressure and probably originated more than 100,000 years ago (Padmanabhan *et al.* (2008)). The rationale behind this hypothesis is that the human race expanded rapidly from a small group with low allelic diversity and low selective pressure, and that it takes a long time to dilute the founder population alleles with the new alleles produced by the increased population (Reich and Lander (2001)). Currently there is insufficient empirical evidence either to prove or to disprove this hypothesis. There are, however, examples of common diseases such as Alzheimer's, deep venous thrombosis (DVT), and type-II diabetes, where the common variants are known.

The rare variant hypothesis holds that environment has a great influence on complex traits and therefore allelic variants have low "attributable risk", that is the probability of carrying the risk allele given the disease phenotype. The detectable range of allele frequencies is often considered to be > 0.001 (Zondervan and Cardon (2004)).

The heated debate between the two camps is not crucial for the research presented here since our model in Chapter 4 exhibits an advantage in power for both

rare and common allele variants. However, the improvement is larger for rare variants.

Common statistical methods

For case-control phenotypes Pearson's 2df test (defined later in Section 3.3) is generally recommended but when genetic effect is additive Pearson's test is not as powerful as the specially tailored Armitage test. Figure 1.6 illustrates the principle behind the Armitage test. A formal definition is given in formula (4.5) of Chapter 4.

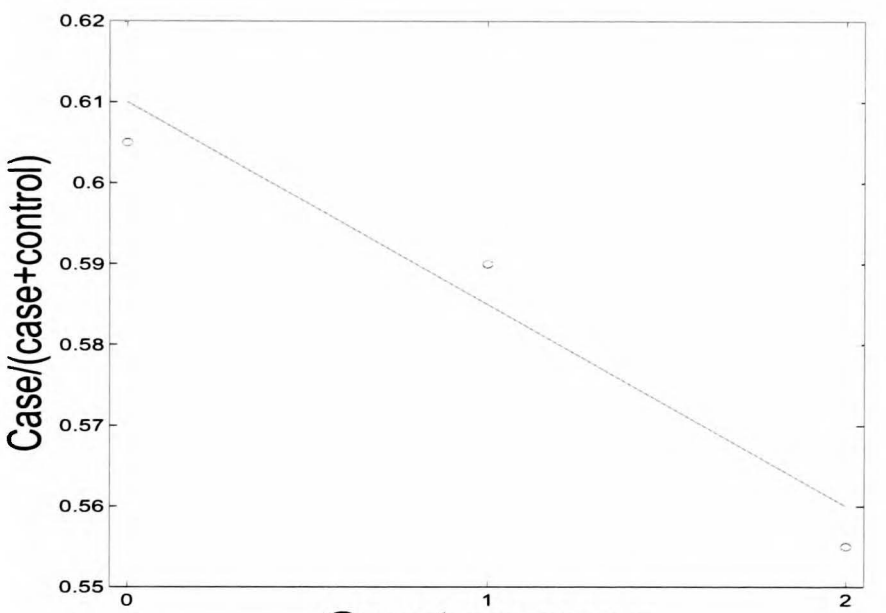


Figure 1.6: Illustration to the Armitage test. The circles represent the proportion of cases in a case-control sample, for each SNP genotype (coded as 0,1,2). The solid line is the least squares fit to the dataset. The Armitage test corresponds to testing the hypothesis that the line's slope is zero. The illustration is adopted from Balding (2006).

However, researchers normally do not know what type of genetic effect the allele has. Thus by choosing the most powerful test for additive effects we risk losing power if the effect is dominant or recessive. Suggested solutions include using maximum test statistics from special additive, dominant and recessive tests; designing a combined test statistic where additive model is given more weight because it is more likely; using Armitage test when MAF is low and Fisher test when counts of all genotypes are high.

Logistic regression is a more advanced option for this phenotype and by restricting its parameters the model can be modified for a specific disease risk. Also it can be extended to multiple SNP analysis and can accommodate interactions and covariates. For continuous outcomes ANOVA and linear regression models can be used.

In this section we provided necessary background for presenting the results of the findings in latter parts of the thesis. In the following chapters we offer improved models for both QTL mapping and GWAS.

We begin with QTL mapping in Chapter 2. Then in Chapter 3 we introduce an innovative model with optimal Bayesian prior which has been accepted for publication in a peer-reviewed journal (Antonyuk and Holmes (2009)). Chapter 4 deals with a wider class of problems in GWAS.

Chapter 2

Suggested methodology for QTL mapping in HS stock

2.1 Current approach

2.1.1 Background

Analysis of F_2 and backcross populations is relatively simple due to the fact that linkage and association are equivalent in these experiments. However, resolution is relatively poor, The 95% confidence interval often encompasses half a chromosome (Yalcin *et al.* (2004)). Such regions can contain many genes. It is easy to detect QTLs in such studies but more difficult to fine-map them. One of the reasons for this is that QTLs of large effect often turn out to be caused by a number of linked loci of much smaller effect. Lack of recombination in F_2 and backcross populations is the major obstacle when we need to fine-map QTLs, the reason being that resolution derives from the presence of many recombinants close to a QTL. The HS is especially suited to fine-scale mapping because of the large number of generations it has passed (Mott and Flint (2002)). Theoretically, the HS stock's resolution is 30 times that of an F_2 intercross (Darvasi and Soller (1995), Mott *et al.* (2000)). For QTLs of small to moderate effect, mapping to under 0.5 cM is possible with fewer than 2,000 animals. The larger number of founders than in two line crosses increase the chance

that a QTL will segregate in the HS population.

The data we used for the analysis is the result of applying the probabilistic ancestral haplotype reconstruction to the HS stock using a software package called HAPPY (Mott *et al.* (2000)). We give an example due to Flint *et al.* (2005), which compares multipoint model to single marker association (SMA), another commonly used technique that we described in Section 1.2.1.

Example in Figure 2.1 considers 4 inbred lines, A, B, C and D with a QTL that has alleles distributed among the strains in such a way that it increases the phenotype in strains A and C and decreases it in strains B and D. We assume that all markers are physically linked to the QTL. If we now consider an ancestral population and try to establish an association between a single marker and the phenotype, there would be no evidence of such association for separately considered markers p and q. Marker p, for example, has alleles 1 and 0 for both decreaser and increaser QTL alleles. Thus, when frequencies of a polymorphic marker do not coincide with linked QTL single marker association will fail to find the association.

Single marker versus multiple marker haplotypes

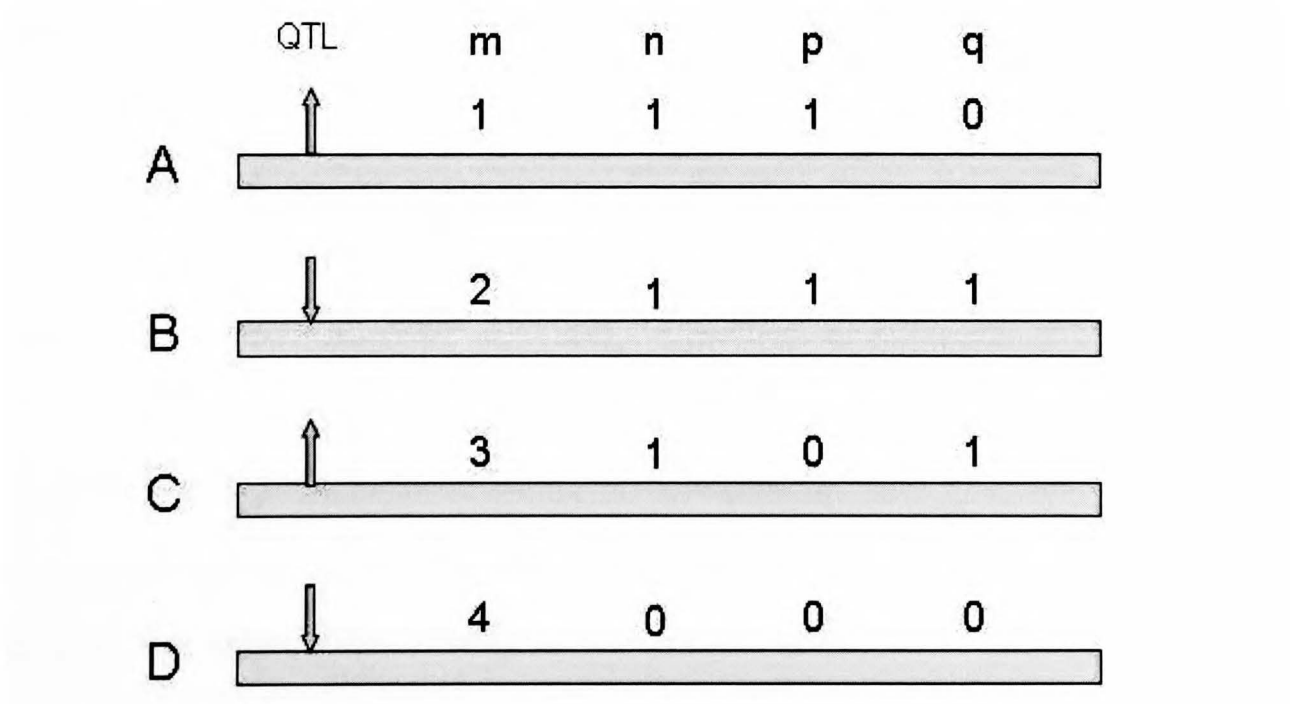


Figure 2.1: Illustration to single versus multiple marker comparison. The graphs shows QTL and marker alleles in the four inbred lines. Adopted from Flint *et al.* (2005).

Marker m though has 4 distinct alleles and will show evidence of association. Now, if we construct a haplotype by two markers, p and q, it can serve as a substitute for the efficient marker m, because it can distinguish all four strains. The example illustrates that haplotypes based on multiple markers are generally more efficient than single-point mapping.

There have been reports of failures of the SM model to detect association for known QTLs (Risch (2000)). For HS stock it failed in the open field behaviour study of HS mice's susceptibility to anxiety carried out by Talbot *et al.* (1999). In that study 67 markers were typed for 751 mice at an approximate distance of 1 centiMorgan over five regions where QTLs were previously detected. For each marker animals were grouped according to their genotype and the group means were tested for significant differences. Only two out of five regions were confirmed as QTLs by performing the SM testing.

Mott *et al.* (2000) sought explanation for the three failures, excluded the possibility of allele fixation by carrying out computer simulations, and arrived at the conclusion that it was caused by the identical marker alleles descending from different founder strains. Single marker analysis does not take into account neighboring markers and founder haplotypes, and therefore cannot distinguish between strains that have different QTL effects but the same alleles at a marker close by (consider again our example of marker p in Figure 2.1). At most markers there are only two or three alleles so it might be impossible to determine from which of the eight founder stains they descended. Mott *et al.* (2000) provides an example of two very close markers, for which the possibility of recombination is excluded, one yielding a significant association and the other showing no association, with the significant marker having different alleles in the founder strains whereas for the failed marker all the strains have the same allele. Later in the article Mott *et al.* (2000) demonstrate how, by incorporating information from the neighboring marker and progenitor haplotypes, the failed marker (or rather the region that includes it) is identified as a QTL.

2.1.2 Details of the HAPPY model

The model incorporates information from the flanking markers and the progenitor strains. It calculates the probabilities of each of the founder stains being the ancestor of a particular allele of an individual from the HS stock. Then, using these estimates, testing for QTL is not for a genetic effect at a single locus, as in SM, but for genetic effect of progenitor haplotypes. Note that reconstructing haplotypes for HS mice would not help since it can not determine ancestry of the alleles which are identical in the founder strains. The problem of reconstructing the progenitor haplotypes can be thought of as a Hidden Markov model, hidden states being the progenitor haplotypes and the observed data the genotypes.

Using the notation in the original paper, we define $P_{m,i}(s,t)$ to be the probability that, for individual i, the progenitor haplotypes at the marker m are (s,t), given:

1. The genotypes for the ordered markers labeled $1,...,m-1$

2. The founder strain haplotypes, expressed as the probability $\pi_m(s|a)$ that the ancestral state at marker m is s conditional on the allele observed at that locus being a

3. The genetic distances d_l between markers $l, l+1$ for all l.

It is assumed that the gap between recombinants has an exponential distribution with mean $1/Gd_m$, where G is the number of generations since the HS was founded (note that pedigree information and interference effects are ignored). It follows that the number of recombinations between two adjacent markers has a Poisson distribution with mean Gd_m. Thus, following Mott *et al.* (2000), if on a particular chromosome locus m is in state σ then the probability that marker $m+1$ is in state s is

$$r_m(s|\sigma) = \begin{cases} e^{-Gd_m} + (1-e^{-Gd_m})/S & s = \sigma \\ (1-e^{-Gd_m})/S & s \neq \sigma \end{cases}$$

Where S is the number of strains. The justification for the first probability is that if the alleles at the two loci are from the same progenitor, it can be either due to absence of recombination (the probability of this is e^{-Gd_m}); or from the joint occurrence of two events: at least one recombination between the loci (with probability $1-e^{-Gd_m}$) and inheriting the allele from strain s at marker $m+1$, which has the prior probability of $1/S$.

This allows us to find the transition probabilities for the Markov chain: the probability that the haplotype at marker $m+1$ is (s,t) given that the haplotype at marker m is (σ,τ) is simply $g_m(s,t|\sigma,\tau) = r(s|\sigma)r(t|\tau)$. Now, conditional on the genotype (a,b) for individual i at the marker $m+1$ and the ancestral haplotypes at m being (σ,τ) , the probability that the haplotypes at $m+1$ are (s,t) is:

$$f_{m,i}(\sigma,\tau|s,t,a,b) = \frac{g(s,t|\sigma,\tau)[\pi(s|a)\pi(t|b)+\pi(t|a)\pi(s|b)]}{\sum_{s',t'} g(s',t'|\sigma,\tau)[\pi(s'|a)\pi(t'|b)+\pi(t'|a)\pi(s'|b)]} \tag{2.1}$$

(the subscript m having been dropped for clarity). Note that the phase of the genotypes is unknown, and so the genotype (a,b) is not distinguished from (b,a). From this result we can find the $P_{l,i}(s,t)$ using the dynamic programming recurrence relation

$$P_{m+1,i}(s,t) = \sum_{\sigma,\tau} f_{m,i}(s,t|\sigma,\tau,a,b)P_{m,i}(\sigma,\tau) \tag{2.2}$$

The summation is over all possible haplotypes at m. It is now simple to calculate $P_{m,i}$ across the whole chromosome, starting with marker 1. Similarly, in the backwards direction we can find $Q_{m,i}(s,t)$, whose definition differs from that of $P_{m,i}(s,t)$, only in that we condition on the genotypes at markers $m+1,...,M$ instead of $1,...,m-1$.

Then one needs to consider the possible states of a QTL at locus L between markers $m, m+1$. It can be linked to both markers (signified by subscript B), linked only to the left marker (L), just to the right (R), or unlinked (U). If we define

the position of the QTL as an unknown distance cd_m from marker m, $0 \leq c \leq 1$, then the probabilities of recombination states of the locus (for one chromosome)

$$\begin{aligned} p_B(c) &= e^{-Gd_m} \\ p_L(c) &= e^{-Gcd_m} - e^{-Gd_m} \\ p_R(c) &= e^{-G(1-c)d_m} - e^{-Gd_m} \\ p_U(c) &= 1 - p_B(c) - p_L(c) - p_R(c) \end{aligned}$$

Two chromosomes can be linked in different ways so the linkage state for a diploid organism is equal to the product of the probabilities of the states of each chromosome. Also, we can average over the unknown distance variable c, assuming a uniform distribution for it, and calculate the marginal probabilities (called "prior probability" in Mott *et al.* (2000)): $p_{XY} = \int_0^1 p_X(c)p_Y(c)dc$, $X, Y \in \{B, L, R, U\}$.

Finally we can calculate the probability that the haplotypes are (s, t) at L, $F_{Li}(s, t)$ by summing over all possible linkage states (dropping subscript m for clarity again):

$$\begin{aligned} &F_{Li}(s,t) = \\ &P(s,t)Q(s,t)p_{BB} + P(s,t)Q(\cdot,t)p_{LB} + \\ &P(\cdot,t)Q(\cdot,t)p_{UB}/S + P(\cdot,t)Q(s,t)p_{RB} + \\ &P(s,t)Q(s,\cdot)p_{BL} + P(s,t)p_{LL} + \\ &P(\cdot,t)p_{UL}/S + P(\cdot,t)Q(s,\cdot)p_{RL} + \\ &P(s,\cdot)Q(s,\cdot)p_{BU}/S + P(s,\cdot)p_{LU}/S + \\ &p_{UU}/S^2 + Q(s,\cdot)p_{RU}/S + \\ &P(s,\cdot)Q(s,t)p_{BR} + P(s,\cdot)Q(s,\cdot)p_{LR} + \\ &Q(\cdot,t)p_{UR}/S + Q(s,t)p_{RR}, \end{aligned}$$

where $P(\cdot, t) = \sum_s P_{m,i}(s, t)$, and the probability that an unlinked locus is in a given state is $1/S$. To clarify the long expression above, $P(\cdot, t)Q(s, t)p_{RB}$, for example, is the probability that the first chromosome of the QTL is in state s and is linked only to the right marker, and the second chromosome is in state t and is

linked to both markers.

Single marker analysis tests for association between the marker and the phenotype whereas QTL can lie in the regions between the markers. As pointed out in Terwilliger (1998) if the correlation between a disease-predisposing gene and a marker is small testing for association will not reject the null hypothesis of no association at the marker location. A means to increase correlations between QTLs and markers is to increase the number of marker loci studied, though this will also increase the false positive rate. In contrast, the HAPPY method tests for a QTL in the intervals between the markers. The algorithm does not rely on the correlations, which might be negligible, but marginalises over possible states of linkage between markers and QTLs.

2.1.3 Current statistical model for HS

The reconstruction method is applied to the HS with 8 progenitor lines, so at the locus there are 36 non-redundant combinations of inherited alleles (inheriting alleles A_kA_m from strains k and m is equivalent to inheriting A_mA_k):

$$\begin{aligned} \mathbf{x}^1 &= (2,0,0,0,0,0,0,0) \\ \mathbf{x}^2 &= (0,2,0,0,0,0,0,0) \\ &\ldots \\ \mathbf{x}^{35} &= (0,0,0,0,0,1,0,1) \\ \mathbf{x}^{36} &= (0,0,0,0,0,0,1,1), \end{aligned} \tag{2.3}$$

where each of the 8 elements of the combinations represents the number of alleles inherited from a particular strain. For instance, $(2,0,0,0,0,0,0,0)$ means that both alleles come from the first strain.

The HAPPY software produces estimated probabilities for each of the $\mathbf{x}^j$. The approach taken in Mott *et al.* (2000) is then to use a linear regression with a design

matrix $X_E = E(X)$, where X has a distribution given by the HMM output:

$$Y = X_E\beta + \epsilon. \tag{2.4}$$

The estimator of β from this model is unbiased

$$\begin{aligned} E(\hat{\beta}) &= E((X'_E X_E)^{-1} X'_E Y) = (X'_E X_E)^{-1} X'_E E(Y) \\ &= (X'_E X_E)^{-1} X'_E X_E \beta = \beta. \end{aligned} \tag{2.5}$$

However, it is important to note that because the true progenitors are unknown, the standard calculation of the variance of the linear regression estimator (2.5) will be over-optimistic. When the design matrix X is known, we have

$$\begin{aligned} Var(\hat{\beta}) &= ((X'_E X_E)^{-1} X'_E) var(Y) ((X'_E X_E)^{-1} X'_E)' \\ &= ((X'_E X_E)^{-1} X'_E) \sigma^2 I ((X'_E X_E)^{-1} X'_E)' \\ &= \sigma^2 (X'_E X_E)^{-1}. \end{aligned} \tag{2.6}$$

This can not be applied to our model, however, since we have

$$var(Y) = var(X_E\beta) + \sigma^2 I = \beta' var(X_E)\beta + \sigma^2 I.$$

Note that this means the variance of the estimators is greater than that indicated by the regression (i.e. one in which the design matrix is deterministic rather than stochastic), unless β is a zero vector. Also note that the variance will be greater the more uncertainty there is over the true progenitor strains, since this will lead to greater values of $var(X_E)$.

Strain	Effect	Standard Error
A	0.32	0.13
AKR	441.24	164.38
BALB/c	-0.22	0.07
C3H/2	0.44	0.31
C57BL	0.09	0.08
DBA/2	-443.04	164.39
I	1.00	0.34
RIII	-0.01	0.06

Table 2.1: Parameter estimates from HAPPY programme. AKR and DBA have excessively high estimates due to collinearity.

It is also worth noting that the variance of the estimator depends on the true values of the parameters, something which was not the case above. In the null case, (that is, where β is a zero vector) we do still have the expression $var(\hat{\beta}) = \sigma^2(X'X)^{-1}$. Since we cannot know for certain whether we are in the null case or not, however, we must not simply assume this expression is correct.

A difficulty with this approach relates to the nature of the design matrix produced by the HMM. Some loci actually convey little information about the progenitor strain because several different strains can have regions with identical markers. Where two strains have identical markers over a significant length of the chromosome, it becomes very difficult to tell them apart. As a result the HMM method will reasonably assign approximately equal probabilities to those two possible ancestral types. This means that the corresponding two columns of the design matrix will be nearly identical (giving an almost rank-deficient matrix $(X'X)$). As a result, $(X'X)^{-1}$ will contain some very large entries, and so the variance of the associated estimators will also be very large, as shown in Table 2.1.

Single marker versus multiple marker haplotypes

It has been shown, however, that this "expectation-substitution approach is valid and efficient for testing the null hypothesis of $\beta = 0$ (Rebbeck *et al.* (2008)). Under the alternative hypothesis conditions it is less efficient than methods which use the full information about the distribution of unknown X ("marginal methods).

(Rebbeck *et al.* (2008)) also point out that even though the expectation-substitution approach may underestimate variability in estimation of β, for modest genetic effects the difference between performance of marginal and expectation-substitution methods is small. Later in the chapter we investigate their difference in statistical power and find that for larger genetic effects it is significant.

Heterogeneous stocks coupled with PHAR method opened a possibility for further progress in QTL mapping by allowing to find QTLs with finer resolution. This chapter showed that to make full use of the technique we need statistical estimators of the effect of genotype on phenotype with better properties. The next chapter suggests a method, which has a sound statistical foundation and is practical.

2.2 Alternative method

Since the problem described above falls into the class of problems with missing data, we now introduce the well-known statistical method for dealing with this type of datasets. In this chapter we define the model, show some estimation results, investigate important properties such as time of convergence, statistical power, sensitivity to starting points, ease of implementation.

First, assuming a linear regression model, we can write the marginal likelihood function of the $p \times 1$ vector of regression coefficients as follows

$$L(\beta) = P(y|\beta) = \sum_X P(y|X\beta)\pi(X) \tag{2.7}$$

where y is a $n \times 1$ vector of measurements, and the sum is taken over all possible realisations of the $n \times p$ design matrix (here $p = 8$) with corresponding probabilities $P(X)$, and where

$$P(y|X\beta) \sim N(X\beta, \sigma^2 I) \tag{2.8}$$

where I is the $n \times n$ identity matrix and σ is assumed known. Now we demonstrate that (2.7) can be rewritten as

$$L(\beta) = P(y|\beta) = \prod_{i=1}^{n} \sum_{j}^{k} P(y_i|x_i^j\beta)\pi(x_i^j) \tag{2.9}$$

with

$$P(y_i|x_i^j\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{1}{2\sigma^2}(y_i - x_i^j\beta)^2 \tag{2.10}$$

$$\pi(x_i^j) = p_i^j \tag{2.11}$$

where $p_i^1, ..., p_i^k$ are initial probabilities of the k possible values vectors x_i can take. To prove that expressions (2.7) and (2.9) are equivalent in full generality requires complicated notation so below we prove it for the case of two observations and two possible values for each row of the design matrix $x_i \in \{a_1, a_2\}$, with corresponding probabilities:

$$\pi(x_1 = a_1) = p_1^1, \quad \pi(x_1 = a_2) = p_1^2, \quad \pi(x_2 = a_1) = p_2^1, \quad \pi(x_2 = a_2) = p_2^2. \tag{2.12}$$

Then, given that probabilities of rows are independent, the four possible design matrices for the model have the following initial probabilities:

$$\pi \begin{pmatrix} x_1 = a_1 \\ x_2 = a_1 \end{pmatrix} = p_1^1 p_2^1 \ , \pi \begin{pmatrix} x_1 = a_1 \\ x_2 = a_2 \end{pmatrix} = p_1^1 p_2^2,$$

$$\pi \begin{pmatrix} x_1 = a_2 \\ x_2 = a_1 \end{pmatrix} = p_1^2 p_2^1 \ , \pi \begin{pmatrix} x_1 = a_2 \\ x_2 = a_2 \end{pmatrix} = p_1^2 p_2^2.$$

The marginal (2.7) simplifies to

$$\begin{aligned}
\sum_X P(y|X\beta)\pi(X) &= P(y_1|a_1\beta)P(y_2|a_1\beta)p_1^1p_2^1 + P(y_1|a_1\beta)P(y_2|a_2\beta)p_1^1p_2^2 + \\
&\quad P(y_1|a_2\beta)P(y_2|a_1\beta)p_1^2p_2^1 + P(y_1|a_2\beta)P(y_2|a_2\beta)p_1^2p_2^2 \\
&= P(y_1|a_1\beta)p_1^1\left(P(y_2|a_1\beta)p_2^1 + P(y_2|a_2\beta)p_2^2\right) + \\
&\quad P(y_1|a_2\beta)p_1^2\left(P(y_2|a_1\beta)p_2^1 + P(y_2|a_2\beta)p_2^2\right) \\
&= \left(P(y_1|a_1\beta)p_1^1 + P(y_1|a_2\beta)p_1^2\right) \times \left(P(y_2|a_1\beta)p_2^1 + P(y_2|a_2\beta)p_2^2\right)
\end{aligned}$$

This prove can be generalised to any number of observations and possible values of design vectors. The simplified representation of the likelihood (2.9) can be useful in many situations. However, since no full analytical expressions of (2.9) or (2.7) are available, maximising them is usually inefficient, since one can only numerically approximate the derivative of this quantity.

Here however, we show that this maximisation can be carried out very efficiently from a computational point of view, using an approach described in Ibrahim *et al.* (2005). That is, we maximise (2.7) using an EM stochastic optimiser, which we describe next.

2.3 The EM algorithm

The EM algorithm, first introduced by Dempster *et al.* (1977) is an optimiser, and is particularly useful in the context of missing data. In settings similar to ours, that is when the missing covariates are categorical and the missing mechanism is MAR (missing at random), it has been shown by Ibrahim *et al.* (2005) that the EM algorithm takes a very simple and computationally efficient form.

First, recall that an EM optimiser is a two-step iterative procedure described as follows:

The first step is an *Expectation* step where one takes conditional expectation of the log-likelihood over the missing covariates given the current coefficient estimate β^{curr} and the observed data.

The second step is a *Maximisation* step in which one maximises the quantity obtained from the previous *Expectation* step over β. The resulting $\hat{\beta}$ is then taken as the new current β^{curr} and both steps are repeated.

In our situation the algorithm can be described as follows:

- E-step for the i^{th} mouse. Let us consider

$$Q_i(\beta|\beta^{(s)}) = E\left\{l(\beta|x_i, y_i)|y_i, \beta^{(s)}\right\} \tag{2.13}$$

  where $l(\beta|y_i, x_i)$ is the log-likelihood of β when the covariates x_i are known and where the expectation is taken over the missing covariates x_i given the current regression coefficient estimate $\beta^{(s)}$ and the observed datum y_i. This leads to

$$Q_i(\beta|\beta^{(s)}) = \sum_{x_i} P(x_i|y_i, \beta^{(s)}) l(\beta|x_i, y_i) \tag{2.14}$$

- E-step for all responses. In our model we have the assumption that given a regression coefficient estimate β^s, the responses are independent. Then it is straightforward to see that

$$\begin{aligned} Q(\beta|\beta^{(s)}) &= E\left\{l(\beta|x, y)|y, \beta^{(s)}\right\} \\ &= \sum_{i=1}^{n} Q_i(\beta|\beta^{(s)}) \\ &= \sum_{i=1}^{n} \sum_{x_i(j)} w_{i,j,s} l(\beta|y_i) \\ &= \sum_{i=1}^{n} \sum_{x_i(j)} w_{i,j,s} l_{y_i|x_i}(\beta) \end{aligned} \tag{2.15}$$

  where $w_{i,j,s} = P(x_i|y_i, \beta^{(s)})$ are weights corresponding to incomplete observations.

- The M-step requires maximisation of the quantity $Q(\beta|\beta^{(s)})$ in (2.15) with respect to β. In the linear regression case this is straightforward to carry out

once we note that (2.15) is simply the log-likelihood of an augmented weighted complete-data regression model where the weights are given by $w_{i,j,s} = P(x_i|y_i, \beta^{(s)})$. The well known estimator of $\hat{\beta}$ is given then by

$$\hat{\beta} = (\tilde{X}'\Omega^{-1}\tilde{X})^{-1}\tilde{X}'\Omega^{-1}\tilde{y} \tag{2.16}$$

where $\tilde{X}$ and $\tilde{y}$ are the augmented complete-data design matrix (i.e. each row contains a possible combination for the realisation of the covariates) and augmented response vector respectively. Ω is a diagonal matrix with diagonal elements $w_{i,j,s}^{-1}$.More detail are provided in Ibrahim *et al.* (2005). For our setting we have the dimensions of the augmented dataset: $n * 36 \times 8$ and $n * 36 \times 1$ respectively, because the number of strains is 8, and $n = 1649$, the number of uncensored observations.

The weights are easily calculated using Bayes Theorem as described, for instance in Ibrahim *et al.* (2005). Note that in our situation, this calculation is greatly simplified because we are given the probability distribution over all the possible x_i's.

It is a useful property of the method that if we consider a generalised linear regression model rather than a linear regression model, as long as the responses are independent given the coefficients' estimates $\beta^{(s)}$, the methodology is easily extended. Implementation is straightforward in the GLM case as well because the calculation of the M step is provided by major statistical computer packages. This is important because it can be applied to many kinds of phenotypes that come from distributions covered by GLMs.

The EM algorithm has become a popular method for estimating maximum likelihood but one of its drawbacks is that it does not produce standard errors as a by-product (Jamshidian and Jennrich (2000)). One solution was proposed in Louis (1982) whereby one calculates the estimated observed information matrix of $\hat{\beta}$ estimated via the EM algorithm, which gives an approximation of the variance-covariance matrix of the estimator. We attempted to use this approach so that

variance of the estimator could be used for constructing test statistics but encountered difficulties similar to those reported in Meilijson (1989). Louis (1982) method requires calculation of the conditional expectation of the square of the complete-data score function which is specific for each problem. The algebra involved was overly complex and intractable so we did not proceed further with this approach.

2.4 EM algorithm for finding maximum-a-posteriori

EM algorithm can be used in situations when we need to maximise a joint distribution that is marginalised over some of the random variables. In our setting we marginalise over the unknown design matrix X. There is no restriction as to what the joint distribution is so we can apply the algorithm to the posterior distribution of the regression parameters β. Let us demonstrate first that the E-step is the same for finding maximum-a-posteriori (MAP) as for finding the MLE. We will prove it using the interpretation of the EM as *lower bound optimisation* described in Minka (1998).

Log of the posterior distribution is linked to the log-likelihood from (2.13) as follows:

$$log(Pr(\beta|y_i, x_i)) = l(\beta|x_i, y_i) + \pi(\beta) + const,$$

where $\pi(\beta)$ is *log* of the prior. According to the interpretation of the EM mentioned above the E-step constructs the lower bound to the objective function (posterior distribution in our case) by introducing an arbitrary PDF $f(X_j)$, which is optimised later, in the following way:

$$\begin{aligned} log(Pr(\beta|Y)) &= log(\Sigma_j Pr(\beta, X_j|Y)) \\ &= log\left(\Sigma_j \frac{Pr(\beta, X_j|Y)}{f(X_j)} f(X_j)\right) \\ &= log\left(E_{f(X)}\left(\frac{Pr(\beta, X|Y)}{f(X)}\right)\right). \end{aligned}$$

Then we can apply Jensen's inequality (Cover and Thomas (1991)) and define the

lower bound that needs to be optimised,

$$
\begin{aligned}
log\left(E_{f(X)}\left(\frac{Pr(\beta,X|Y)}{f(X)}\right)\right) &\geq E_{f(X)}\left(log\left(\frac{Pr(\beta,X|Y)}{f(X)}\right)\right) \\
&= E_{f(X)}\left(log\left(\frac{Pr(Y|X,\beta)Pr(\beta)}{f(X)}\right)\right) + const \\
&= B(\beta, f(X)),
\end{aligned}
\tag{2.17}
$$

which is followed by maximisation of $B(\beta, f(X))$ with respect to $f(X)$. Since $Pr(\beta)$ in 2.17 does not depend on X, it can be omitted and then the maximisation is equivalent to the E-step when the objective function is the likelihood rather than the posterior.

Thus, to apply the EM algorithm to finding MAP we only need to modify the M-step, which maximises the weighted log of the objective function. For this maximisation we use the iterative reweighted least squares (IRLS) method that is commonly used for finding ML. We demonstrate below how the Bayesian version of the algorithm is derived. Our derivation starts with a reminder of how IRLS maximisation algorithm is applied to the frequentist logistic regression, then we apply it to Bayesian regression models with Normal priors, and then further extend it to the case of the weighted log-posterior.

2.4.1 IRLS for logistic regression

Our starting point is the very clear derivation of the IRLS method for GLMs given in Green (1985), which we apply to logistic regression. Observations are distributed as follows: $y_i \sim Bin(1, \eta_i)$, link function $\eta_i = \{1 + exp(-x^i\beta)\}^{-1}$ and the log-likelihood is $L = \Sigma_{i=1}^{n}\{y_i log\eta_i + (1 - y_i)log(1 - \eta_i)\}$. We define the following variables that are applicable to any GLM and will be used later for the linear model:

$$
\frac{\partial L}{\partial \beta} = D^T u, \quad u = \frac{\partial L}{\partial \eta}, \quad D = \frac{\partial \eta}{\partial \beta}, \quad A = E\left(-\frac{\partial^2 L}{\partial \eta \partial \eta^T}\right) \tag{2.18}
$$

For logistic regression the i-th row of those variables are (when observations are

independent):

$$u_i = \frac{dL_i}{d\eta_i} = \frac{y_i - \eta_i}{\eta_i(1-\eta_i)},$$

$$D_i = x_i\eta_i(1-\eta_i).$$

Matrix A is diagonal with elements $1/(\eta_i(1-\eta_i))$.

The IRLS algorithm is effectively an implementation of the Newton-Raphson method with Fisher scoring and provides an iterative solution to the likelihood equation (derivative of the likelihood equals zero). The key feature of the method is that the equation for each iteration of the Newton-Raphson method has the form of the common weighted linear regression:

$$(X^TWX)\beta = X^TWY.$$

The Newton-Raphson (N-R) equation, after applying Fisher scoring method, is

$$(D^TAD)(\beta^* - \beta) = D^Tu,$$

where β^*, β are the parameter to optimise and the current value respectively. It is not immediately obvious that the two equations are of the same form but we can re-write the N-R equation as:

$$(D^TAD)\beta^* = D^TA(D\beta + A^{-1}u),$$

and now the analogy with the weighted linear regression equation is obvious if we consider $Y^* = D\beta + A^{-1}u$ as the response variable and A as the weight matrix.

2.4.2 Logistic-Normal Bayesian model

The Logistic-Normal Bayesian model has the following log-posterior

$$log(Pr(\beta|Y)) \propto L + \pi(\beta),$$

where L is logistic log-likelihood and $\pi(\beta)$ is again log of the prior for p independent parameters of interest, the prior itself having the distribution $N(0, \Sigma)$ with the covariance matrix $\Sigma = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_p^2)$. To apply IRLS algorithm to the Bayesian model we only need to replace everywhere log-likelihood with log-posterior (not standardised), and it is easy to see that the N-R equation becomes (for any prior):

$$\left(D^T A D - \frac{\partial^2 \pi(\beta)}{\partial \beta \beta^T}\right)(\beta^* - \beta) = D^T u + \frac{\partial \pi(\beta)}{\partial \beta}, \tag{2.19}$$

Conveniently, some of the elements cancel out for the case of a Normal zero-mean prior. We illustrate it for the $p = 2$ case, when $\beta_1 \sim N(0, \sigma_1^2), \beta_2 \sim N(0, \sigma_2^2)$:

$$\frac{\partial^2 \pi}{\partial \beta \beta^Y} = \begin{pmatrix} -\frac{1}{\sigma_1^2} & 0 \\ 0 & -\frac{1}{\sigma_2^2} \end{pmatrix}, \quad \frac{\partial \pi}{\partial \beta} = \begin{pmatrix} -\frac{\beta_1}{\sigma_1^2} & 0 \\ 0 & -\frac{\beta_2}{\sigma_2^2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma_1^2} & 0 \\ 0 & -\frac{1}{\sigma_2^2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Now we see that the second derivative times β in the left hand side of 2.19 cancels out with the first derivative on the right. Thus, the N-R equation reduces to (after moving all the elements with β to the right)

$$\left(D^T A D + \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}\right) \beta^* = D^T A (D\beta + A^{-1} u) = D^T A Y^*. \tag{2.20}$$

The last equation can still be seen as weighted least squares if we augment the matrices D, A and Y^*. Namely, we can construct

$$D_a = \left(\begin{array}{c} D \\ \hline 1 \quad 0 \\ 0 \quad 1 \end{array}\right), \quad Y_a^* = \left(\begin{array}{c} Y^* \\ \hline 0 \\ 0 \end{array}\right), \quad A_a = diag(a_1, a_2, ..., a_n, 1/\sigma_1^2, 1/\sigma_2^2). \quad (2.21)$$

Using simple matrix algebra it can be verified that the resulting weighted linear regression $D_a^T A_a D_a = D_a^T Y_a^*$ is equivalent to (2.20).

2.4.3 Weighted log-posterior via IRLS

For the M-step we need to maximise the weighted log-posterior of the form:

$$\Sigma_i^n w_{i,j,s} L(y_i, |x_i, \beta) + \pi(\beta),$$

where $w_{i,j,s}$ were defined in the section for the frequentist EM. IRLS algorithm is then easily modified. Out of all variables in 2.18 only the first and second derivatives of the objective function wrt to link η are affected,

$$u_i = w_{i,j,s} \frac{y_i - \eta_i}{\eta_i(1-\eta_i)}, \qquad A_{ii} = \frac{w_{i,j,s}}{\eta_i(1-\eta_i)}.$$

The new factors $w_{i,j,s}$ cancel out in $Y^* = D\beta + A^{-1}u$ so the only change is in the weight matrix (in IRLS terms) A.

2.4.4 IRLS for the Bayesian Normal-Normal model

All of the above simplifies significantly for the Bayesian model of Normal likelihood and Normal priors: $y_i \sim N(\eta_i, \sigma), \eta_i = x^i\beta, L = (-n/2)log(2\pi\sigma^2) - \Sigma_{i=1}^n (y_i - \eta_i)^2/(2\sigma^2)$, $\beta_1 \sim N(0, \sigma_1^2), \beta_2 \sim N(0, \sigma_2^2)$, where σ is assumed known. The variables defined earlier become:

$$u_i = \frac{dL_i}{d\eta_i} = \frac{y_i - \eta_i}{\sigma^2},$$

$$D = X,$$

$$A = 1/\sigma^2 I,$$

where I is the identity matrix. Substituting all the variables into equation (2.20), we note that the current value of the parameter, β, cancels out:

$$\left(X^T A X + \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix} \right) \beta^* = X^T A (X\beta + Y - X\beta) = X^T A Y, \qquad (2.22)$$

which means that the algorithm has only one iteration. This is hardly surprising if we take into account that the N-R algorithm for finding roots via linear approximation was applied here to the derivative of the log-posterior, which is a linear function in the case of the Normal-Normal model.

To solve the equation (2.22) its matrices are augmented in the same way as for the logistic case in (2.21). To maximise the weighted log-posterior we need only, as for the logistic likelihood, multiply the elements of matrix A by the weights: $A_{ii} = w_{i,j,s}/\sigma^2$.

2.5 Estimation Results

Let us examine the behaviour of the EM algorithm when applied to our data set. For this purpose, consider two simulated examples. In the first one, the $N = 1000$ responses (e.g. phenotypes) are taken to be normally distributed while they are taken to be binomially distributed in the second simulation. This allows us to apply our methodology using a linear and generalised linear regression framework. Also, a version of the algorithm for survival analysis has been implemented using an R package that allows for survival models with weights.

For both distributions we take the missing data for each response to have a multinomial distribution with weights taken to be identical to that of the real data (provided by HMM). The vector of regression coefficients β is fixed to some value. Thus, each response y_i is generated by drawing its genotype from the given multi-

nomial distribution and multiplying the latter by the fixed effects of each progenitor strain, β.

2.5.1 Normal distribution

Let the parameters of the model be defined as follows,

$$\begin{aligned} y_i|x_i,\beta &\sim N(x_i\beta', \sigma^2) && (2.23)\\ \beta &= (1,0,0,0,0,0,0,0) && (2.24)\\ \sigma &= 0.3 && (2.25)\\ \pi(x_i) &\sim MN(1,\alpha_1^i,...,\alpha_p^i), && (2.26) \end{aligned}$$

where $MN(1,\alpha_1^i,...,\alpha_p^i)$ is multinomial distribution with 1 trial, $\alpha_1^i,...,\alpha_p^i$ are given multinomial distribution's cell probabilities which are taken to be equal to the initial probabilities of the real data set. Simulated phenotypes are shown in Figure 2.2. The EM algorithm was applied with starting value $\beta = (1.5, 1.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5)$. Convergence to the true value of β was remarkably quick (15 iterations) as shown in Figure 2.3. Note however, that one needs to consider various initial starting values in order to ensure that the algorithm converges to the global maximum rather than a local one. This feature is shared however by all optimising algorithms and is discussed later.

2.5.2 Binomial distribution

Let us now consider a logistic regression defined as follows,

$$\begin{aligned} y_i|\beta,x_i &\sim B(1,p),\ p=\frac{e^{x_i\beta}}{1+e^{x_i\beta}} && (2.27)\\ \beta &= (0.8,0.1,0.1,0.1,0.1,0.1,0.1,0.1) && (2.28)\\ \pi(x_i) &\sim MN(1,\alpha_1^i,...,\alpha_p^i), && (2.29) \end{aligned}$$

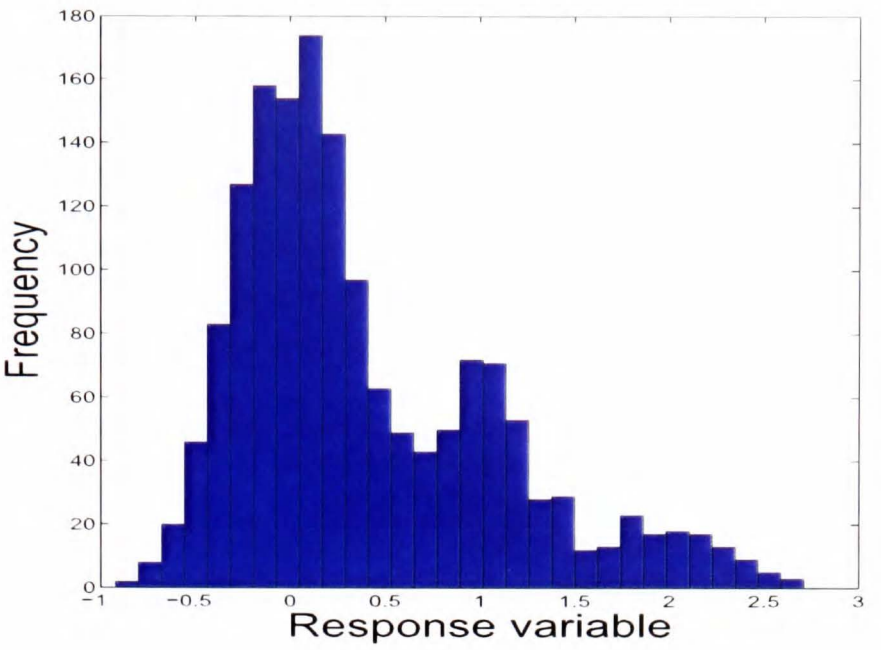


Figure 2.2: The distribution of y_i's for fixed β's, σ and X matrix drawn using the initial probabilities. Because of the particular β we assume, note that the generated phenotypes behave as a mixture of 3 normal distributions.

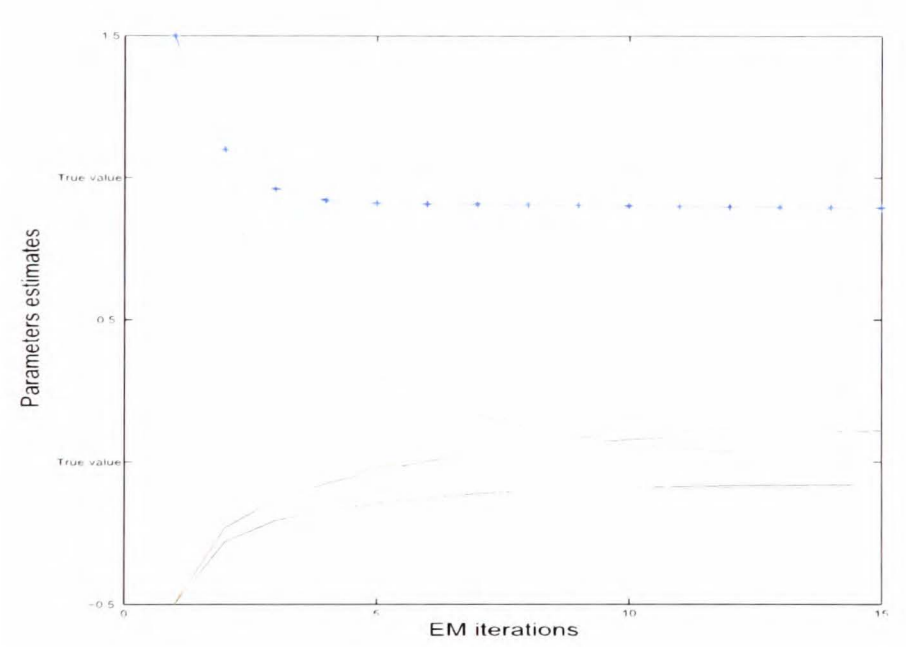


Figure 2.3: Linear model's convergence of estimates of β parameters, $N = 1000$, $\sigma = 0.3$. The true values for generating the dataset were $\beta_1 = 1$ and the other seven β's equal zero. The starred line represents estimates of β_1, all the solid lines are estimates of the other β's. Initial values for the algorithm: $\beta^0 = (1.5, 1.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5)$.

As already mentioned in Section 3.2, it is straightforward to implement our methodology within this setting. One only needs to replace in the M-step a weighted linear regression optimiser with a standard weighted GLM optimiser. Similarly, in

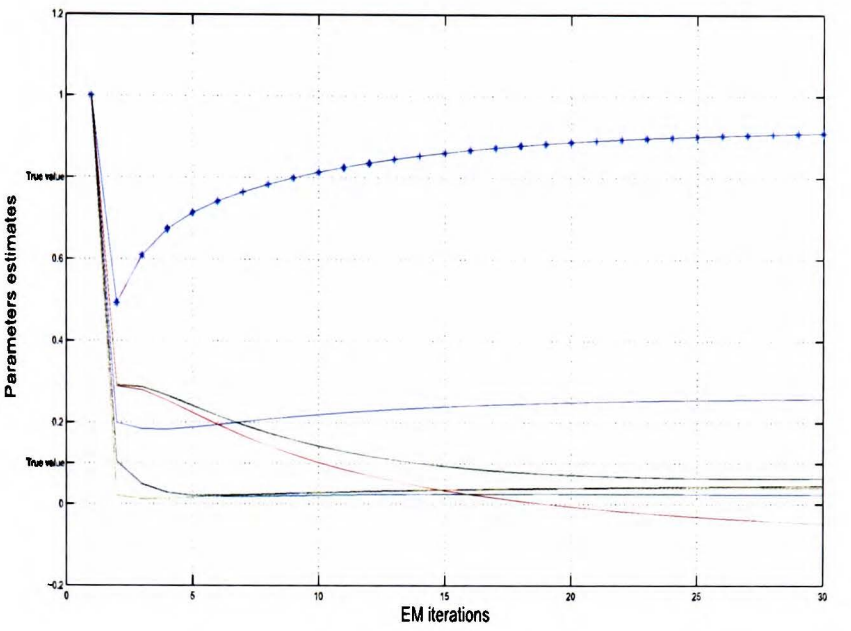


Figure 2.4: Binomial model's convergence of estimates of β parameters, $N = 1649$. The true values for generating the dataset were $\beta_1 = 0.8$ and the other seven β's equal 0.1. The starred line represents estimates of β_1, all the solid lines are estimates of the other β's. Initial values for all β's were set to 1.

the linear regression case it is straightforward to adapt the E-step to this setting.

We have plotted the corresponding EM output in Figure 2.4. Again, the algorithm converged to the true estimate, and the time of convergence was very short.

2.5.3 Computational efficiency

All simulations as well as the run on the real data set using an exponential GLM framework were carried out using Matlab software. When running the EM algorithm on the real data set ($N = 1649$ mice), convergence was *typically* obtained after around 30 iterations, which took *5 seconds* to run. The next section contains more details on convergence.

2.5.4 EM convergence

The problems with parameter estimation caused by collinearity of the output from the HAPPY algorithm can also arise in our suggested method. We noticed that the algorithm can be sensitive to starting values when the non-zero genetic effect parameter is collinear with another parameter. In search of a strategy for choosing

Initial values								
Set N	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
1	-3.2	-5.6	-3.0	-7.4	0.1	5.3	2.7	9.4
2	10.0	10.0	10.0	10.0	10.0	10.0	10.0	10.0
3	-15.0	-15.0	-15.0	-15.0	-15.0	-15.0	-15.0	-15.0
4	0	0	0	0	0	0	0	0
5	0	2.0	0	10.0	0.10	0	5.0	0
6	0.10	0	0	0	0	0	0	0.3
7	0	0	0	0.3	0.3	0	0	0
8	0.5	0.2	0	0.4	0.5	0.5	0.0	0.2
EM estimates at 200th iteration								
Set N	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
1	0.0052	-0.0205	0.0428	0.2895	-0.0106	0.0060	-0.0042	0.0033
2	0.0052	-0.0205	0.0428	0.2896	-0.0106	0.0009	0.0009	0.0033
3	0.0052	-0.0205	0.0428	0.2896	-0.0106	0.0009	0.0009	0.0033
4	0.0052	-0.0205	0.0428	0.2896	-0.0106	0.0009	0.0009	0.0033
5	0.0035	-0.0219	0.2906	0.0547	-0.0114	0.0017	-0.0024	0.0021
6	0.0052	-0.0205	0.0428	0.2896	-0.0106	0.0009	0.0009	0.0033
7	0.0052	-0.0205	0.0428	0.2896	-0.0106	0.0009	0.0009	0.0033
8	0.0052	-0.0205	0.0428	0.2895	-0.0106	0.0057	-0.0039	0.0033

Table 2.2: Initial values and corresponding EM algorithm estimates. The true betas are all zero except $\beta_3 = 0.3$. Only set 5 provides the right convergence.

starting values that would guarantee correct convergence (at least on average) we performed simulations with many different starting points and true values of the parameters.

We used true values where one β is positive and the rest are zero, and also starting sets of the same form, e.g. $\boldsymbol{\beta} = (0.0, 0.0, 0.0, 0.0, 0.0, 1, 0.0, 0.0)$. Also, we investigated randomly generated true values, and a few combinations of true values and starting points selected "by hand". Unfortunately, for some β's collinearity turned out to be a serious issue. For instance, for the true values $\boldsymbol{\beta} = (0, 0, 0.3, 0, 0, 0, 0, 0)$ the algorithm converted to the right numbers only for two starting points, set number 5 in Table 2.2 and when the starting set was equal to the true values (not shown in the table). Set number 5 does not seem to be closer to the true value than any other sets which led us to conjecture that random starting points might converge to the right figures, at least on average.

We carried out simulations for a data set generated by the effect $\beta_3 = 0.3$ with 100 randomly simulated starting values. The algorithm converged to estimates of β_3 close to 0.3 only 34 times and 41 times estimates of β_4 were close to 0.3 (the rest were zero). Thus, for some combinations of true β's we could not find a strategy for choosing starting points that would make the EM converge to the true values. It is necessary to point out that the HMM probabilities we used had *several* columns with identical probabilities for all individuals, which posed a great difficulty for both the old (plugging in $E(X)$) and the proposed approaches.

2.5.5 Convergence under no collinearity

We can reasonably expect that in many situations distribution of the design matrix is not "collinear", i.e. there are no possible values of the design matrix that have the same probability for all individuals in the sample. An investigation into behaviour of the EM algorithm under such a condition was carried out. To remove collinearity we permuted probabilities produced by HMM, independently for each individual. We found that for the normally distributed response, $N = 1600, \sigma = 0.3$, and a range of true betas (13 sets) and 16 different sets of starting values the EM always converges to values very close to true β's. However, for some starting points the algorithm first appears to converge to wrong values. Namely, a few randomly generated starting points produced the estimates shown in Figure 2.5. Estimates initially appear to converge to wrong values but then suddenly change direction and converge to the true β's. The graph is is designed to illustrate the sudden change in the estimates, so all the lines are of the same type.

2.5.6 Empirical power estimation

We performed simulations for comparison of our method's power to the currently used regression model with the expected design matrix. We ran 20000 simulations with all β's set to zero except for the first one, which was given values from 0.05 to 0.5, with fixed $\sigma = 0.3$, $N = 300$ (results are shown in Figure 2.6). The simulations

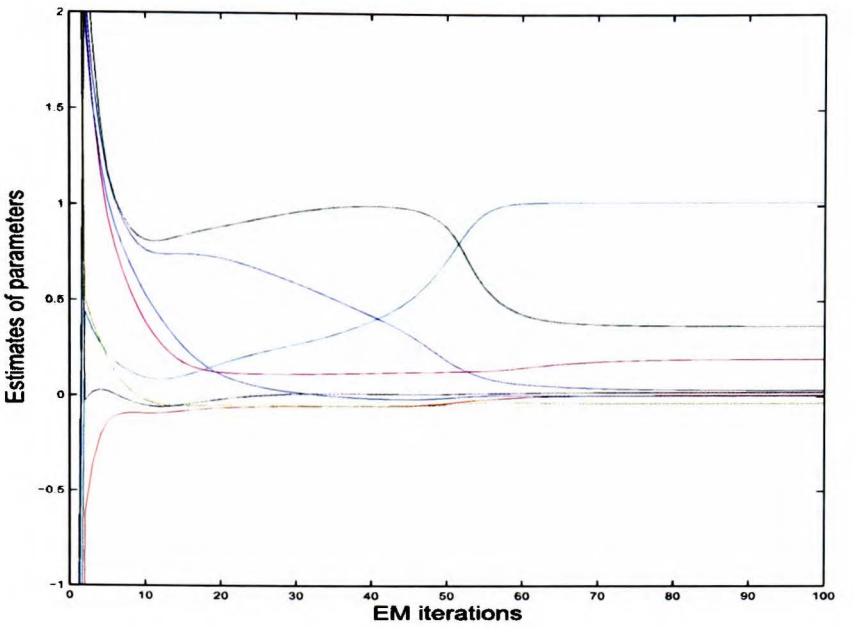


Figure 2.5: The EM estimates for randomly generated starting values: $(-2.5, -1.9, -6.6, -4.3, 8.5, 5.8, 5.0, 8.5)$. The estimates converge eventually to numbers very close to the truth. The true values are: $(0.0, 0.4, 0.0, 1.0, 0.2, 0.0, 0.05, 0.0)$. Each line represents iterative estimates of one of the eight parameters.

were done as follows: the design vector (i-th row of the design matrix) for each Y_i was drawn from the initial probability distribution $\pi(x_i)$, which were taken from the HMM output. Before running the power simulations we checked if the null distribution is χ^2 with seven degrees of freedom as it should be from the standard theory. We simulated datasets under H_0 and the distribution turned out to be χ^2 with five degrees of freedom. The whole of the next chapter addresses the question of why this is so, and how to overcome the issue, but for the purpose of estimating power we simply use the χ^2_5 null distribution, which we obtained empirically.

Results of empirical power calculations are shown in Figure 2.6. The Monte Carlo standard error was calculated as in Boos and Brownie (1986), using the following approximation: $\alpha(1-\alpha)/N_{sim}$, where α is a nominal level of the test and N_{sim} is the number of Monte Carlo simulations. The power of the test using our estimates was significantly higher. For example, at $\beta_1 = 0.3$ the estimate of our power was 64.1% compared to about 56.7% of the linear estimators from the model with the expected matrix (2.4). Seeing that power is one of the most important feature of any method for QTL mapping and that there are other desirable properties of the

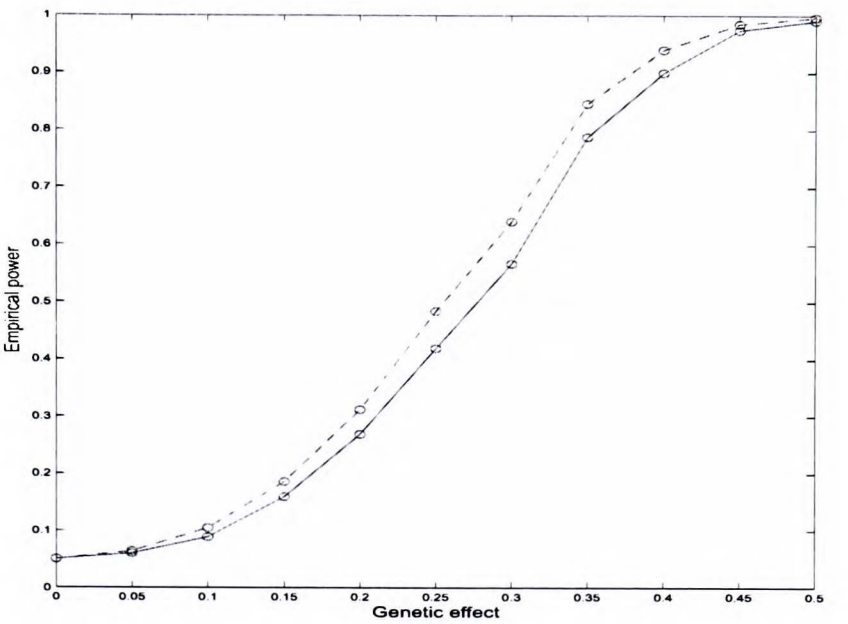


Figure 2.6: Power curves for the two models: the dashed line - estimates via the EM, solid line - linear model with the expected design matrix plugged in. Sample size $N = 300$, variance $\sigma = 0.3$. Lines link power estimates for the values $\beta_1 \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$. Power estimates at those values are displayed as circles. The null hypothesis was tested at nominal level 5%, power was estimated from 10,000 simulations, Monte Carlo standard error ≈ 0.002

technique just described, we strongly recommend using the presented method.

2.6 Sensitivity to errors in HAPPY estimates

So far, in our investigation of the properties of the suggested method of testing for association, we have provided the EM algorithm with the true probabilities of possible design matrices (progenitor haplotypes). However, the output probabilities of the HAPPY method are only estimates and can contain errors. Thus, it is of interest to explore how the EM method performs when the probabilities deviate from those used to generate the design matrices in the analysis above, as well as to see how the EM method performs relative to the plug-in of expected phenotype method under such conditions.

The HAPPY algorithm's effectiveness depends on its ability to distinguish ancestral haplotypes across the interval. The power will be lower where all markers have the same noninformative allele distribution, but markers share information where

there is a mixture. Noninformative allele disttribution can result in "collinearity" in estimated probabilities. The effect of such collinearity on the performance of the EM method is investigated in Section 2.5.4. Here we look at the effects of other possible errors in estimation of the probabilities of allele ancestry.

Our approach is to generate design matrix X using one set of probabilities, then to distort the probabilities in various ways and pass them to the two methods of estimation of the genetic effect (with subsequent testing for association). We look at both, deterioration of estimation of genetic effects β and reduction in statistical power. A few ways of introducing errors to the probabilities were considered:

- Random deviations in all 36 probabilities
- Swapping the largest probability with the fourth largest

We experimented with swapping the first two largest probabilities. The resulting changes in estimation were not very large, and therefore we chose swapping with the fourth. This can probably be explained by observing that sometimes the top few probabilities are of similar magnitude, for example in the second set in Figure 2.7.

Random deviations were implemented via drawing from Beta distribution with mean at the probability estimate provided by HAPPY and variance set by experimentation so that the distortions look substantial but do not completely change the ranking of the probabilities. An illustration of what these errors produce is given in Figure 2.7. Somewhat surprisingly these errors did not cause much deterioration in the estimates of genetic effects, neither by the EM method nor by the expected genotype method, as can be seen in Table 2.3.

We conjecture that random deviations do not affect the two models very much because the highest probability can stay the highest after the errors are introduced. It is possible that the methods are more dependent on allocation of the highest probability, that is on which progenitor haplotype is the most likely. Intuitively, when the design matrix is not given, having the highest probability allocated to the right combination would help significantly with estimation. Thus we exposed the

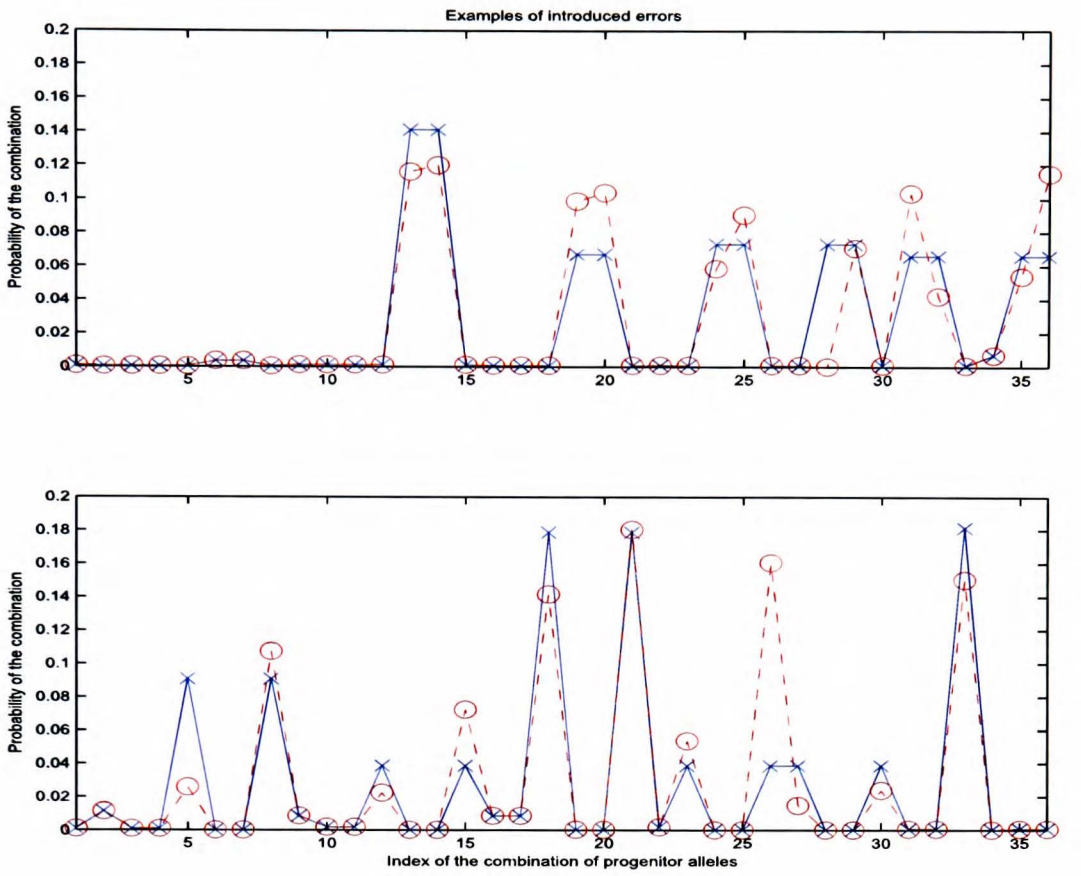


Figure 2.7: Original and distorted probabilities of combinations of ancestral alleles. Blue Xs are original probabilities, red circles are results of distortion. The lines join Xs and circles for better visualisation.

methods to a different kind of error, *swapping probabilities*, ie assigning the largest probability to the combination with the fourth largest probability and visa versa.

As we conjectured the deterioration in precision of estimates becomes more profound (Table 2.4), although for some parameters such as β_2 estimation improved. Figure 2.8 illustrates the deterioration for a few parameters, and also demonstrates that the introduced errors caused slower convergence of the EM algorithm. However, the combination of true values and starting points used in the graph is more difficult for the EM method than other settings as it was especially chosen for demonstration of possible slow convergence in Section 2.5.5.

Ultimately we are interested in how the power of the considered methods compares. We performed simulations to assess power empirically, the results for the "swapping" distortion are shown in Figure 2.9. Both methods' power deteriorates, more so for the EM method, even though it stays higher than the expected design

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
True value	0.000	0.400	0.000	1.000	0.200	0.000	0.050	0.000
EM (true)	0.045	0.454	-0.009	1.005	0.117	0.036	-0.004	-0.003
EM (errors)	0.044	0.464	0.001	1.005	0.129	0.029	-0.013	0.004
$E(X)$ (true)	-0.034	0.568	0.054	1.149	0.056	-0.003	-0.082	-0.054
$E(X)$ (errors)	0.059	0.512	0.094	0.906	0.038	0.056	-0.059	0.031

Table 2.3: Changes in accuracy of estimation of the genetic effect after introduction of errors in the probabilities of ancestral combinations. Results are for a dataset of $N = 1600$, $\sigma = 0.3$, starting values for the EM $\{-2.5, -1.98, -6.63, -4.35, 8.55, 5.89, 5.08, 8.53\}$

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
True value	0.000	0.400	0.000	1.000	0.200	0.000	0.050	0.000
EM (true)	0.045	0.454	-0.009	1.005	0.117	0.036	-0.004	-0.003
EM (errors)	0.077	0.425	0.039	0.999	0.027	0.037	-0.023	0.043
$E(X)$ (true)	-0.034	0.568	0.054	1.149	0.056	-0.003	-0.082	-0.054
$E(X)$ (errors)	0.147	0.333	0.116	0.720	0.005	0.152	0.011	0.158

Table 2.4: Effect of swapping the probabilities of possible combinations of ancestral alleles, $N = 1600$, $\sigma = 0.3$, starting values for the EM algorithm $\{-2.5, -1.98, -6.63, -4.35, 8.55, 5.89, 5.08, 8.53\}$

matrix method's. The more severe decrease in the EM method's power can probably be explained by the fact that in the E-step it reassesses the probabilities of haplotypes in the E-step, using both phenotypes and the initial probabilities, and so passing the wrong initial weights does not get corrected sufficiently. Perhaps it may require more iterations to converge. Also, it is possible that this particular distortion of probabilities and allocating the genetic effect to the first parameter β_1 (in our power simulations) is especially deleterious to its performance.

2.7 Testing procedures

In the previous sections it has been shown how the EM algorithm can provide estimates that quickly converge and have higher power to detect the effect. However, the ultimate goal of QTL analysis is to perform tests of whether or not the locus in consideration has an association with the phenotype.

In this chapter we consider a few possible tests for the presence of a genetic

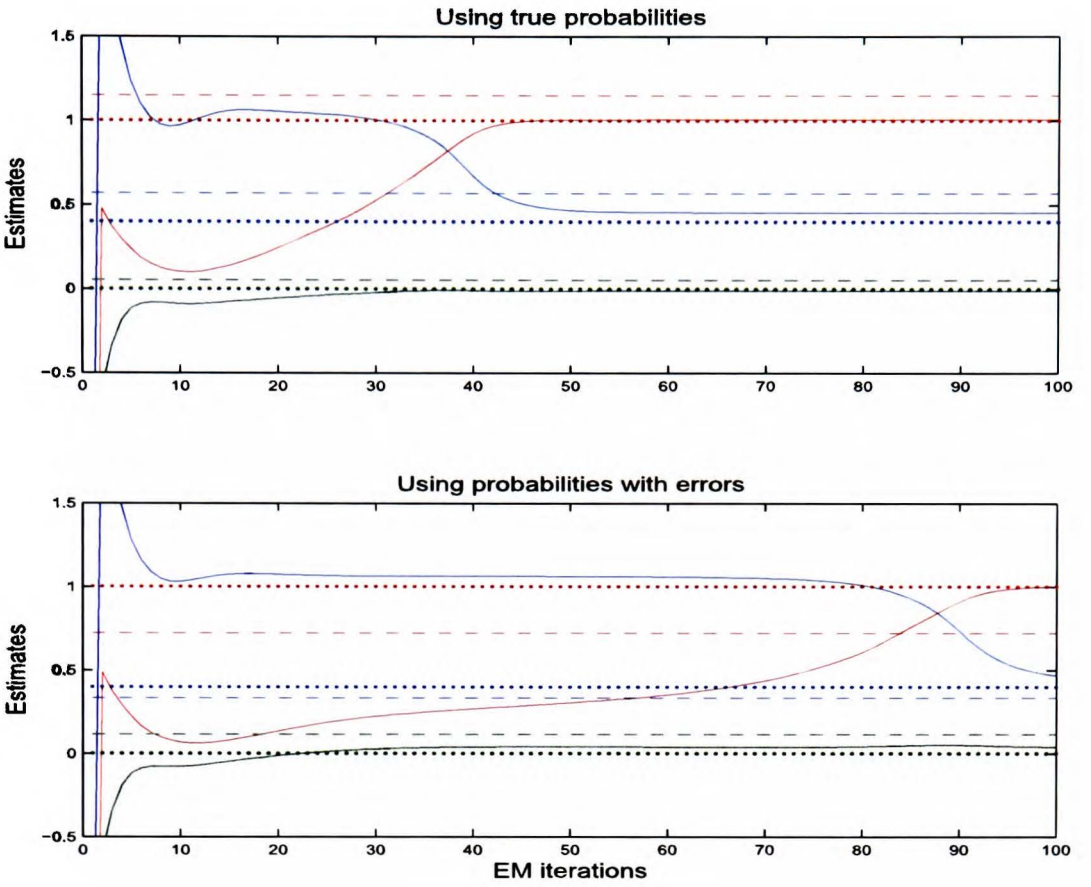


Figure 2.8: Estimation of β_2 (blue), β_3 (green), and β_4 (red) before and after swapping the highest probability with the fourth. Dots represent the true values of the parameters, dashed lines - pluging-in expected design matrix, solid lines - EM estimates.

effect. The reason for considering a few tests rather than using just one with desirable properties (for instance, the uniformly most powerful LR test) was that in our setting there is a deviation from the null distribution suggested by the standard theory. The null distribution used for the power simulations in Section 3.3.4 is an example of such deviation. Given that QTL mapping is done genome-wide, on thousands of loci, it is not feasible to perform simulations at each locus to determine the null distribution as we did in our power estimation section. Even though estimation using the EM algorithm takes a few seconds at each locus, performing it thousands of times at hundreds of loci is not feasible.

The implications of and possible solutions to this common situation in QTL mapping have not been fully addressed in the literature, though the property has been stated, so we intend to fill the gap here. This feature of the distribution of the

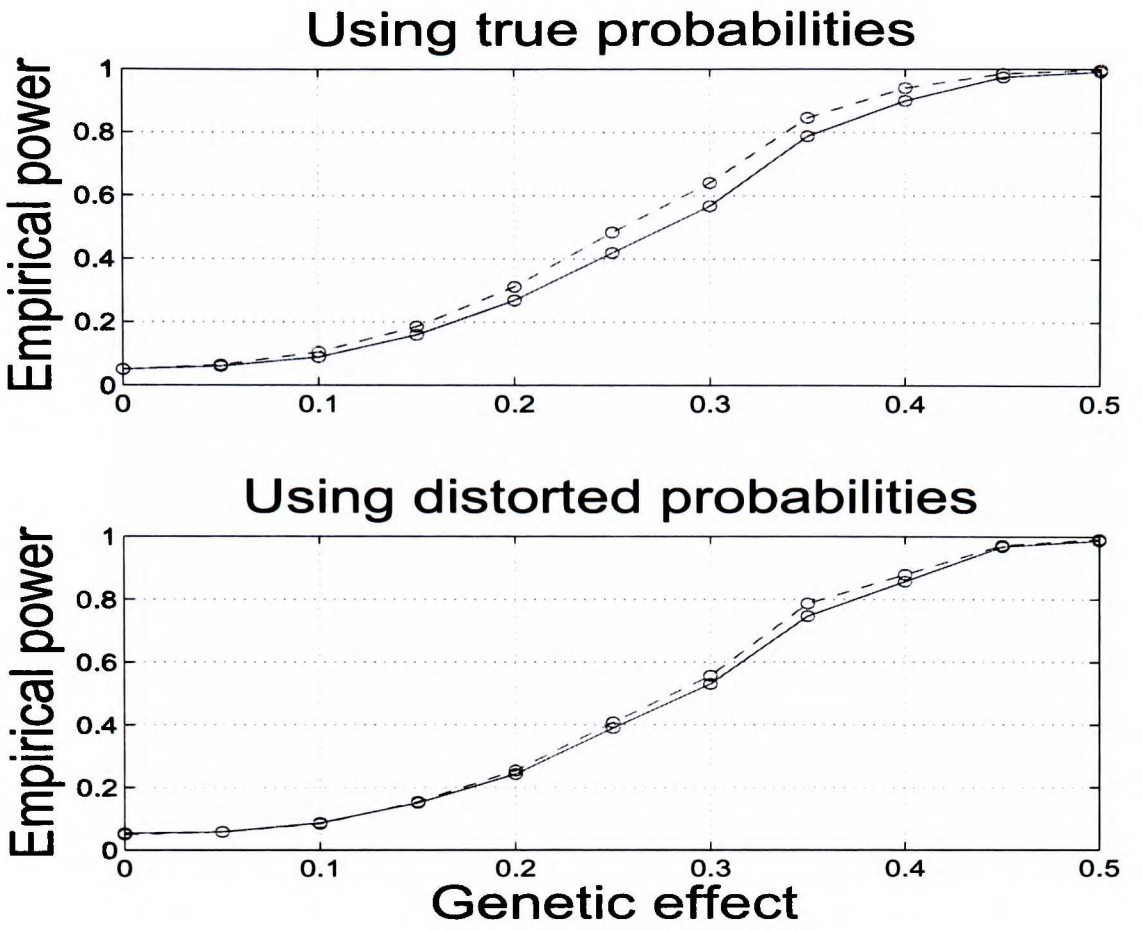


Figure 2.9: Power estimation for both methods, solid lines represent the expected design matrix method, dashed lines - the EM method. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

test statistics in models like ours is of considerable importance because association studies, where the genotypes are inferred from adjacent markers, are commonplace.

In our testing procedures the null hypothesis is that there is no QTL effect at a given locus. This is equivalent to the statement that all the alleles, inherited at this locus from different progenitors' strains, have the same effect on the phenotype. Taking into account that the sum of each row in our design matrix is 2, we have

$$\begin{aligned} E(Y_i) &= \beta_0 + x_1^i\beta_1 + x_2^i\beta_2 + ... + x_8^i\beta_8 \\ &= \beta_0 + \beta(x_1^i + x_2^i + ... + x_8^i) \\ &= \beta_0 + 2\beta, \end{aligned} \tag{2.30}$$

which shows that having no QTL effect is equivalent to a model with an intercept

and all β's equal zero.

We identified two possible areas in our search for a testing approach suitable for our situation:

- Test statistics with null distributions that change when initial weights change but we know how the distribution changes (theoretically or by a quick estimation). We consider LR test and contemplate about using the weights matrix from weighted OLS.

- Instead of calculating p-values for all loci we could try to exclude loci that are very unlikely to be a QTL. Then we can run simulations on a subset of loci to find the correct null distribution. We used cross-validation techniques for implementation of this idea.

2.7.1 Likelihood ratio test

We use the likelihood ratio (LR) test following the "EM-LR method" from Niu *et al.* (2005). The statement of the problem in Niu's article is the same as ours except that the initial probabilities p_i^j of (2.9) are not given and instead are treated as parameters to be estimated. The idea of the method is that we use the EM algorithm to find ML estimates of the parameters and LR test to establish the presence of QTL.

Before performing tests using the EM-LR method we ran simulations to check the null distribution of the LR statistics. We used the real initial probabilities from the HMM described in Section 1.2.2. Surprisingly, the closest match for the null distribution was χ^2 with 5 degrees of freedom (df) instead of 7 degrees of freedom suggested by the standard theory (for example Theorem 6.3.1 from Bickel and Doksum (2001)). Figure 2.10 shows the empirical distribution and the fit.

This finding prompted us to start an investigation into why the null distributions of various test statistics for the EM estimators depend on the initial probabilities. We ran simulations for different sets of initial probabilities. Figure 2.11 demonstrates how the null distribution can change dramatically depending on what set of initial probabilities we use.

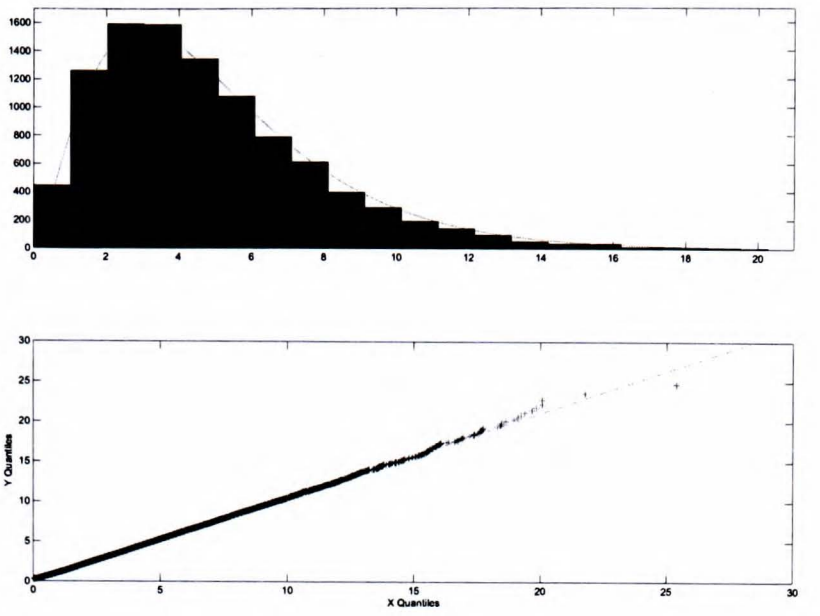


Figure 2.10: Empirical null distribution of the LR using real initial probabilites from 10,000 simulations of response under H_0. The fitted line and the QQ plot are for χ^2 with 5 df.

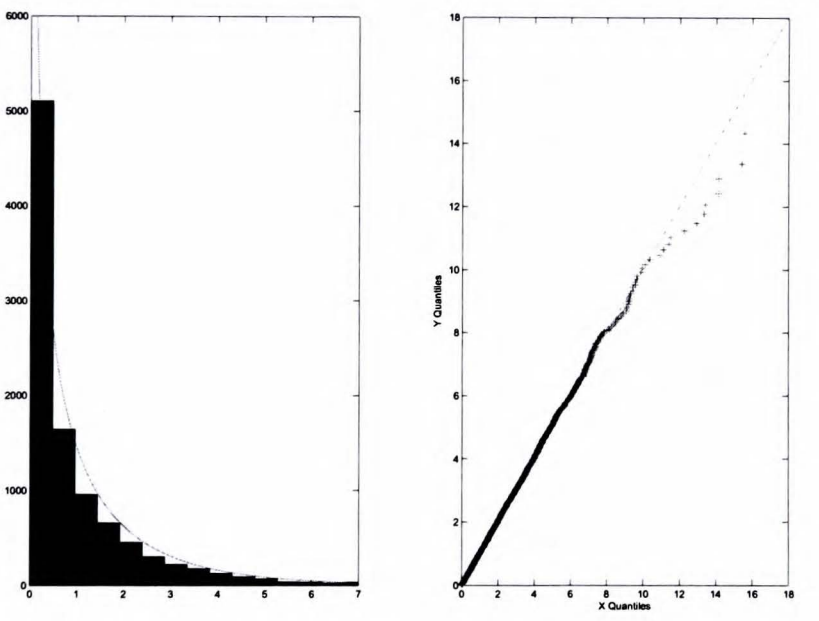


Figure 2.11: Empirical null distribution of the LR using simulated "flat" initial probabilities (equal to 1/36 plus a small random deviation). The fitted line is χ^2 with 1 df.

This property is applicable to many situations because our statement of the problem is quite broad. A lot of phenotypes are normally distributed and the EM algorithm is often used for parameter estimation. Thus, whenever genotypes at loci are inferred and their possible values are assigned probabilities, the situation of

deviation from the standard null distribution can arise. Even the famous article by Lander and Botstein (1989) that gave rise to QTL mapping falls precisely into the category.

2.7.2 Overview of mixture models

Broadly speaking, the data-generating mechanism considered here is of the class of *finite mixture models.* In this section we provide a brief overview of this large field, focusing on the literature related to the issues with the likelihood ratio test. Then we discuss application of previous findings to our situation and conjecture on the reasons for the departure from the standard null distribution.

Finite mixtures of distributions provide a modelling approach to a wide variety of random phenomena (McLachlan and Peel (2000)). As any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities with common variance, mixture models provide a convenient framework to model unknown distributional shapes. One of the first analyses involving the use of mixture models was undertaken by Karl Pearson to analyse the ratio of forehead to body length of crabs sampled from the Bay of Naples (Pearson (1894)), which suggested that there were two subspecies present. Pearson used the method of moments and the computational effort involved was daunting. Various attempts were made over the ensuing years to simplify fitting of mixture models. During the first half of the 20th century research methods based on the method of moments focused on the use of graphical techniques (Fowlkes (1979)).

Considerable advances have been made in the last 30 years in fitting of finite mixture models by the method of maximum likelihood. Even though applications of the method to finite mixture models was studied in the 1960s (Day (1969), Wolfe (1970)), interest in the use of finite mixture distributions was greatly stimulated by the seminal paper of Dempster *et al.* (1977) on the EM algorithm. Also, Tan and Chang (1972) and Fryer and Robertson (1972) demonstrated that the method of maximum likelihood has advantages over the method of moments used in Pearson

(1894). For deeper reviews of the research area readers are referred to Titterington *et al.* (1985), McLachlan and Basford (1988), McLachlan and Peel (2000).

Now we outline basic definitions of finite mixture models, closely following Picard (2007), which will help with establishing how our problem fits into the mixture models framework. In the mixture model based approach, data is presumed to arise from a mixture of an initially specified number of "populations" in different proportions. A random sample of size n is defined, $Y = \{Y_1, ..., Y_n\}$, where Y_i is a random variable with probability density function $f(y_i)$, and y_i is its realisation. It is also assumed that the density of Y_i can be written in the form

$$f(y_i) = \sum_{p=1}^{P} \pi_p f_p(y_i),$$

where $f(y_i)$ is called a component density of the mixture, and π_p are "weights" or "mixing proportions" which are constrained to be $0 \leq \pi_p \leq 1$ and $\sum_p \pi_p = 1$. Note that in mixture models the weights are the same for each random variable Y_i, whereas in our HS analysis we have different probabilities for each mouse.

The component densities are often assumed to belong to some parametric family, and are specified as $f(y_i; \theta_p)$, where θ_p is unknown vector of parameters for the p-th component of the mixture. All the unknown parameters are then denoted by vector $\psi = (\pi_1, ..., \pi_{P-1}, \theta_1, ..., \theta_P)$. Thus, in its general form, the mixture model problem has another difference from our formulation: the weights are unknown. We consider the weights known, and they are provided by the HAPPY algorithm described in Section 2.1.2. Mixture models with known weights are studied in Goffinet *et al.* (1992) which we discuss later in this section.

As mentioned above, the EM algorithm is often used for fitting mixture models. To apply the EM method the problem is formulated as incomplete-data: assignment of observations to populations is considered as missing data. A new random variable Z_{ip} is introduced which equals 1 when y_i belongs to population p and 0 otherwise. The labeling variables Z_{ip} are considered independent, and the conditional density of Y_i given that $Z_{ip} = 1$ is $f(y_i; \theta_p)$. If we define $X_i = \{Y_i, Z_i\}$ as the complete data

vector, with only Y_i observed, the density is given by

$$g(x_i;\psi) = \prod_{p=1}^{P} [\pi_p f(y_i;\theta_p)]^{z_{ip}}, \tag{2.31}$$

where z_{ip} are realisations of the random variables Z_{ip}. The complete data log-likelihood is written in the form:

$$logL^c(x;\psi) = \sum_{i=1}^{n}\sum_{p=1}^{P} z_{ip} log\{\pi_p f(y_i;\theta_p)\}, \tag{2.32}$$

and the parameters ψ are estimated by repeating the E and M steps of the algorithm. The details of implementation can be found in Picard (2007). McLachlan and Peel (2000) draw attention to a nice feature of the EM algorithm, that the solution of the equation for finding θ_p that maximise log-likelihood often exist in closed form, and demonstrate it on a normal mixture model.

Now we move on to discussion of applying the likelihood ratio test (LRT) to mixture models and issues with its null distribution, which are similar to the problem we faced while performing hypothesis testing for our model for the HS data. It has been noted by several authors that the null distribution of the LRT for mixture models can be different from the standard chi-squared distribution. McLachlan (1987) point out that when testing for the null hypothesis $H_0 : P = p_0$ versus $H_1 : P = p_1$, $p_1 > p_0$, regularity conditions do not hold for $-2log\lambda = 2\{logL(\hat{\psi}_1) - logL(\hat{\psi}_0)\}$ to have its usual asymptotic null distribution of chi-squared; bootstrapping is suggested for the assessment of the null distribution of $-2log\lambda$. Feng and McCulloch (1994) discovered that LRT is unbounded above and does not satisfy standard regularity conditions for normal mixture models with unequal variances, when the goal is to determine the number of components. Thode *et al.* (1988) considered normal mixture models with unknown mean and variance under the null, versus the alternative of a mixture of two normal distributions each with unknown mean and unknown (but equal) variances. They examined many sample sizes n and found that for $n < 1000$ the LRT distribution does not follow χ^2_2 as expected. Goffinet *et al.* (1992) studied

a particular case of a mixture of two normals, namely when the weights are known. They proved, among other results, that for the hypotheses (π_1 is fixed)

$$\begin{aligned} H_0: \quad f(y_i;\psi) &= f(y_i;\mu_1,\sigma) \\ H_1: \quad f(y_i;\psi) &= \pi_1 f(y_i;\mu_1,\sigma) + (1-\pi_1) f(y_i;\mu_2,\sigma), \end{aligned} \tag{2.33}$$

(2.34)

the null distribution of the LRT follows $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ when $\pi_1 = 0.5$ (McLachlan and Peel (2000)), and χ_1^2 otherwise.

Can the previous developments in mixture models be used for association studies of the HS stock? In particular, do they provide any insight into the reasons for the deviation from the standard null distribution? To answer these questions let us first reiterate the differences already mentioned. In mixture models weights are postulated to be equal for all Y_i (phenotypes of HS mice in our case), while they are different in our model since the HAPPY algorithm calculates individual probabilities for each mouse. Also, in mixture models weights are usually considered unknown. So most results are not directly applicable to our situation.

One could reformulate our problem of testing the HS stock for genetic association in terms of mixture models. For example, consider the setting presented in Section 2.5.1, where $\beta = (1,0,0,0,0,0,0,0)$. Then the phenotype is distributed as a mixture of three normal distributions with means $\mu_1 = 0$, $\mu_2 = 1$, $\mu_2 = 2$ (from multiplication of β by only three possible values of $x_i^1 \in \{0,1,2\}$, representing the number of inherited alleles from the first strain). Figure 2.2 plots a simulated sample from the distribution.

Thus, we could specify a mixture model:

$$
\begin{aligned}
H_0: \quad f(y_i;\psi) &= \phi(0,\sigma) \\
H_1: \quad f(y_i;\psi) &= \sum_{p=1}^{P} \pi_p \phi(\mu_p,\sigma), \qquad (2.35)
\end{aligned}
$$

where $\phi(\mu,\sigma)$ is normal density. The unknown parameters would be $\psi = \{P, \pi_1, ..., \pi_{P-1}, \mu_1, ..., \mu_P\}$. We could then estimate the unknown parameters and test the null hypothesis. Imagine that there is a genetic effect at the locus under investigation and our test rejects the null. In this case we would have estimates of the number of components in the mixture, the means of the populations (which are a product $x_i * \beta$), and the weights. Thus, we would not know which strain's allele influences the phenotype as the model is not regression-based. We would establish, however, that the locus is a QTL.

Going back to the EM method which we introduced in this chapter, let us re-visit the issue with the null distribution shown in the previous section. McLachlan and Peel (2000) point out that for mixture models such as (2.35), under H_0, parameters lie both in a nonidentifiable subset and on the boundary of the parameter space. To explain why this is so, imagine that the component densities are completely specified. Then the null hypothesis corresponds to all π_p except one being zero, which is on the boundary of the parameter space. H_0 will also hold when all μ_p are equal, then π_p are nondidentifiable.

When parameters under H_0 lie both on the boundary of the parameter space and in a nonidentifiable subset, the classic regularity conditions about the asymptotic properties of the MLE are not valid under H_0 (Cramer (1946)). This results in deviation of the null distribution of the LRT from chi-squared (McLachlan and Peel (2000)). We conjecture that this is the reason for the observed deviation of the LRT for our model presented in Section 2.2. If we used the model (2.35) for our dataset this explanation would be unconditionally valid, but our regression model with missing covariates has a different likelihood and parameters. However, our

parameters are probably also nonidentifiable, which may lead to the ML estimators not being normally distributed.

We performed simulations to check if the MLEs are normally distributed, as they should be when the regularity conditions hold. Figure 2.12 demonstrates that some parameters do not follow a normal distribution.

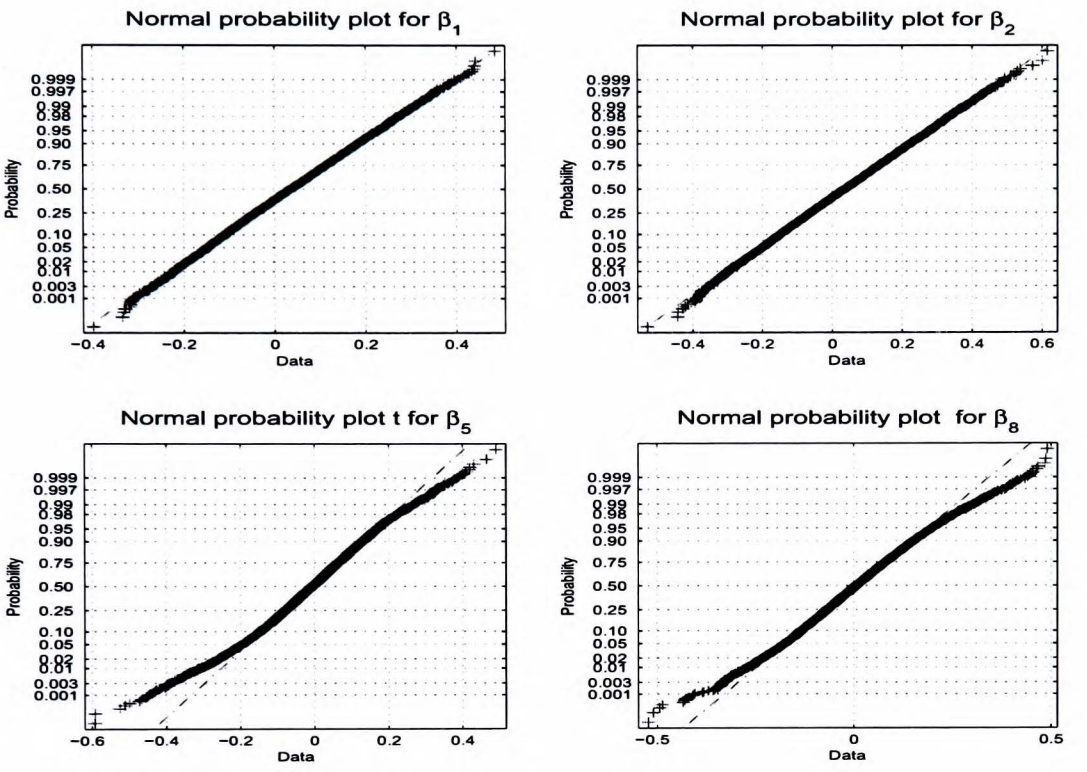


Figure 2.12: "Normal probability plot" for estimates of β_1, β_2, β_5 and β_8, implemented using Matlab function normplot(). The purpose of a normal probability plot is to graphically assess whether the data could come from a normal distribution. If the data are normal the plot will be linear. Estimates of β_5 and β_8 do not appear to be normally distributed.

In further support of our conjecture, we mentioned in Section 2.5.4 that initial probabilities are "collinear" in the sense that there is no information in them to be able to distinguish between some strains (parameters β). In other words, some of the eight parameters are nonidentifiable, a situation which is known to lead to deviations from the standard null distribution.

2.7.3 F-like test

The test statistic we use here is not standard and does not theoretically follow F distribution. The purpose is to show how, depending on the initial weights, "F-like statistic" with the EM estimators can have a distribution either very close to F-distribution or deviate considerably from it.

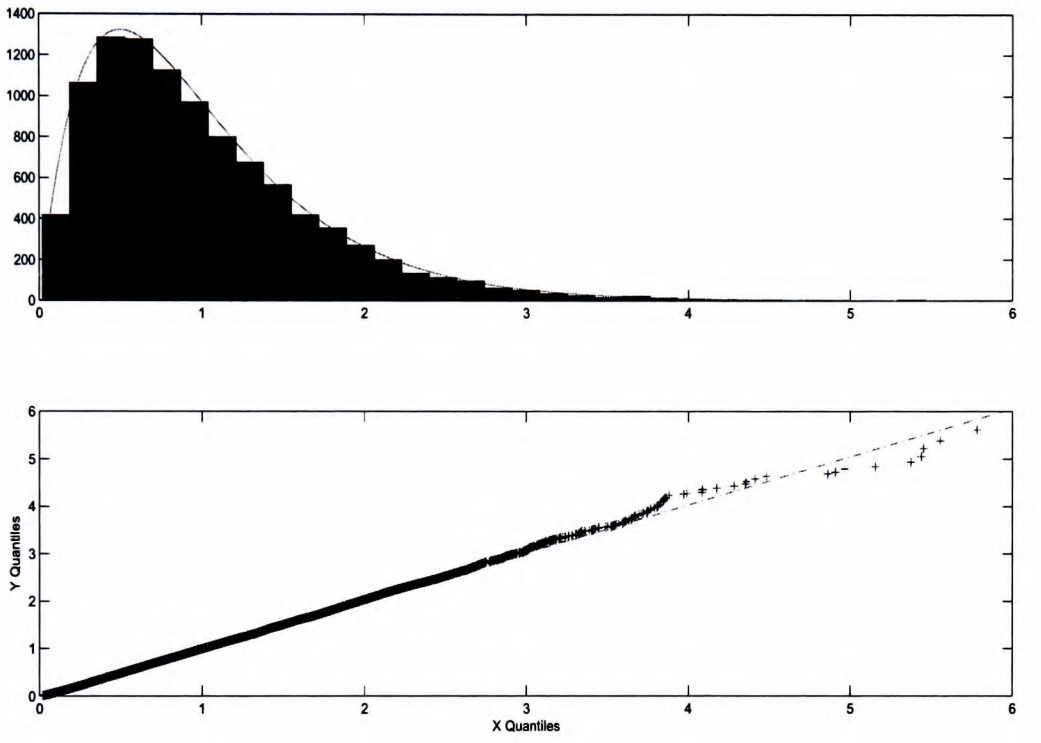


Figure 2.13: Simulation results for the null distribution of F-like test. 10,000 sets of Y were generated under H_0, β's estimated via the EM with "real" initial weights. The fitted line and the QQ plot are for $F_{4,995}$.

Under the null conditions, our model's likelihood (2.7), which is a mixture of normals, reduces to a single normal distribution. As mentioned before, the estimators from the EM maximise the likelihood. Thus, under the null hypothesis, we can expect our estimates to be very similar to the usual regression estimates because they maximise the same likelihood. Therefore we can substitute, as an experiment, our EM estimates for the linear estimates in the common F test of all parameters being zero.

Figure 2.13 shows the null distribution for a set of initial weights from an HMM output. The estimates of β from the EM were plugged into

$$\frac{(TSS - RSS)/df_1}{RSS/df_2} = \frac{\left[\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]/df_1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/df_2} \tag{2.36}$$

The null distribution is remarkably close to F distribution but a different set of weights reveals that it is not always the case. When we use "flat" initial weights that are equal 1/36 plus a small random deviation, the parts in the equation (2.36) do not have the required distributions to make it F-distributed. Namely, for the previous weights $TSS \sim \chi^2_{999}$, $RSS \sim \chi^2_{995}$ and thus their difference follows χ^2_4. For the flat weights we get a distribution of RSS shown in Figure 2.7.3. It appears to have degrees of freedom close to $1,000$ and as a result $TSS - RSS$ is very tight and does not follow χ^2 distribution.

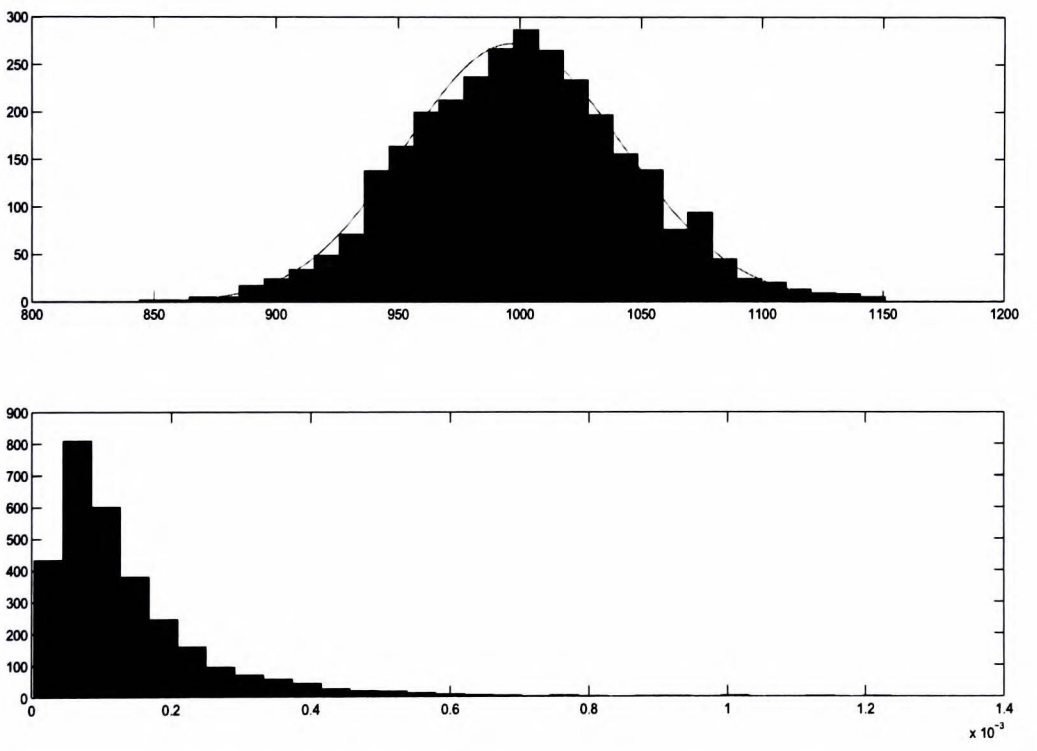


Figure 2.14: Null distributions for "flat" weights. The first graph is the distribution of RSS and the fitted line is χ^2_{999}. The second figure is histogram of $TSS - RSS$.

2.7.4 Conjecture

Testing for our method remains an open research question at present. However, we are strongly confident though that a solution exists for reasons shown in simulation

results in Figure 2.15. The graph suggests that the null distribution of the LR is always χ^2 and the number of degrees of freedom vary depending on the initial weights. If this hypothesis is correct, then to determine the null distribution all that needs to be done is to determine degrees of freedom of χ^2. That can be done by running a manageable number of simulations at each locus. We note, however, that until the conjecture is proved theoretically we can not assume that the the null distribution is always χ^2. In Section 2.7.2 we saw that in a similar setting of mixture models the null distribution sometimes is not χ^2 distributed.

Now we provide a few more details of how close the distributions in Figure 2.15 are to chi-squared with various degrees of freedom. One way of assessing the fit is to take theoretical values of the 95-th percentile of the chi-squared distribution with corresponding degrees of freedom, and calculating the percentage of the simulated points lying to the right of this value. If the percentage is close to 5% then the null distribution is well approximated by the chi-squared in the tail area, and Type I error rate is close to the desired 5%. For the first set of initial weights in Figure 2.15 this percentage is 4.95%, for the second - 4.91%, the third - 4.92%, for the fourth - 5.27%.

2.8 Future work

The work started here has potentially very interesting extensions.

- Applying the EM method to imputation

  The situation that arises when the genotypes are not known but are assigned a probability distribution is not unique to the HS stock. In fact, imputation techniques currently receive much attention in the literature because SNP data is often uncertain in population association studies (Marchini *et al.* (2007), Servin and Stephens (2007)). There are certainly a few improvements which can be made to the currently used methods utilising the results above: Monte Carlo integration for calculating Bayes factors in Marchini *et al.* (2007) can be speeded up by using a deterministic formula (2.9) rather than simulating

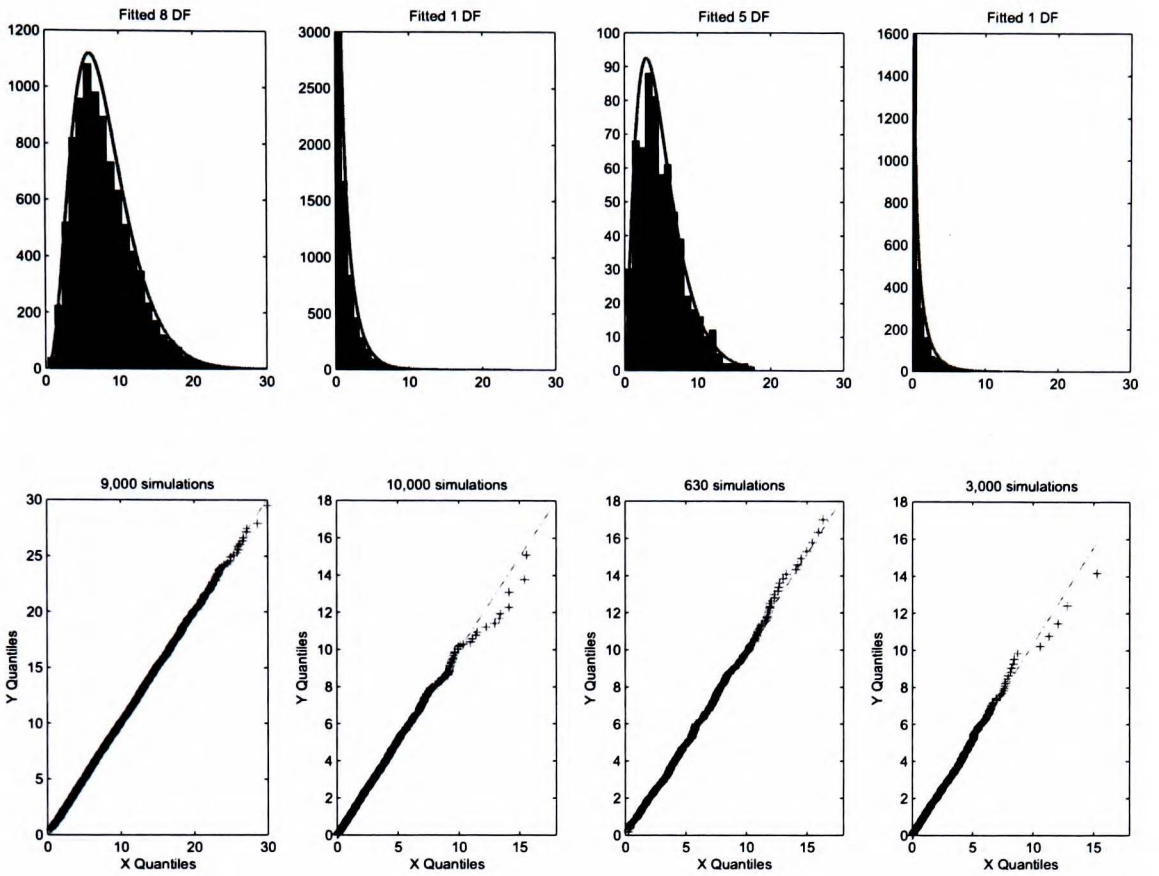


Figure 2.15: Null distribution of the LR statistic for 4 different sets of initial weights. Each column is for one set. Degrees of freedom were based on calculated gamma distribution fit. QQ plots are versus χ^2 distribution with the same DF as in the graphs above them. The first set is for most of the weight randomly allocated to two combinations, second - 95% of weight on the first combination, third - real weights from HMM, fourth - weight evenly spread across first 8 combinations.

X's; drawing β's can be achieved more efficiently by exploiting knowledge of the mode which our modified EM algorithm can find.

- Theoretical null distribution of the LR

  Given that the method presented here has very appealing properties and the only remaining hurdle appears to be knowing the null distribution, developing a theoretical proof of the empirical result in Section 4.4 is probably worthwhile.

  Next steps can be: searching for results similar to Thode *et al.* (1988), who found that the distribution is χ^2 provided there are enough observations; going through the steps of the theorem that proves χ^2 distribution of the LR; analysing the weights matrix of the last iteration's of the EM.

- Extending the model to multiple QTL

  Our collaborators from the Wellcome Trust Centre of Human Genetics in Oxford have been developing tools to account for the background effects of the other loci. Therefore it might be beneficial to extend the model to multiple QTL analysis, perhaps similarly to the MQM method of Jansen (1993) from Section 1.2.5.

Chapter 3

Association studies when allele frequencies are known

In this chapter we consider testing for association when the population-wide allele frequency is known. We suggest a Bayesian model which has a number of advantages. Using extensive simulation, the following results are demonstrated:

- For rare diseases the most powerful conventional (frequentist) design is to actively discard the available control data even though it contains information for association.
- That the most powerful (frequentist) design changes with the underlying prevalence of the disease; from a case-only-discard-controls design for low prevalence to a case-control design for higher prevalence.
- That a single Bayesian test, using a Beta-Binomial model, is more powerful than all frequentist tests we considered.
- That the Bayesian test utilises all the available information (both cases and controls) irrespective of disease prevalence.
- That the retrospective ascertainment of the genotypes from case-control GWAS implies parameter constraints on the support of the Beta prior in a Bayesian association model.

3.1 Introduction

Genome-wide association studies (GWAS) are becoming increasingly important for localising genetic variants which might convey differential disease susceptibility. With the emergence of Biobanks and the publication of data from large-scale GWAS (WTCCC (2007)) we will soon be in the enviable situation of obtaining very precise estimates of the population minor allele frequencies for SNPs which make up the panels in standard genotyping arrays, such as those produced from Illumina and Affymetrix. Under these circumstances we have investigated the statistical implications for GWA study designs. We note that everything we discuss here is also applicable to problems of testing for addmixture and in animal FC crosses (Hoggart *et al.*, (2004)) which is discussed later in this chapter in Section 3.4.

When the population allele frequencies are available it is well known that a case-only design has the largest statistical power to detect a genetic effect, see for example Hoggart *et al.* (2004). That is, for a limited budget allowing for only n individuals to be genotyped, the optimal strategy is to use only affected individuals from the population. However, it is clear that there will be many situations when the genotypes of controls (known unaffecteds) will also be freely available; such as from genotyped-phenotyped individuals' data obtained from other GWAS. In this chapter we explore the statistical implications for testing for association with n cases where controls are available and the population MAF is known.

In the Results section we show the estimates obtained from frequentist (non-Bayesian) and Bayesian approaches to testing with known population MAF.

Bayesian approaches to testing for a single SNP association remain somewhat unexplored, possibly because of the additional computation that they can require, but this is probably going to change in the near future (Balding (2006)). Here we transform the problem from a model with categorical exposure to one with binary exposure. The early work on binary exposure was done by Zelen (1986), Nurminen and Mutanen (1987), and Marshall (1988). All three papers derive posterior distribution of the odds ratio or risk ratio parameters. Zelen (1986) suggested using

a ratio of posterior probabilities with arbitrary critical value. Nurminen and Mutanen (1987) based inference on the highest posterior density, and Marshall (1988) used posterior credible intervals. Müller and Roeder (1997) considered continuous exposure. A very important result of equivalence of prospective and retrospective likelihoods was established by Seaman and Richardson (2004): they demonstrated that for case-control studies the posterior distribution of the log-odds ratio obtained from the "prospective" likelihood, i.e. based on the probability of disease given exposure, is the same as the posterior obtained from the "retrospective" likelihood (based on the probability of exposure given disease status). Retrospective likelihood is the natural likelihood for case-control studies because subjects are recruited according to their disease status, but the result of Seaman and Richardson (2004) proves that the Bayesian analysis of case-control studies can be done using the relatively simple logistic regression model which is based on the prospective likelihood.

Other, more recent, papers on the use of Bayesian methods include Lunn *et al.* (2006), Servin and Stephens (2007) and Morris (2005). Also, the Bayesian version of SNPTEST software was used in WTCCC (2007). In Section 4.2, we discuss how the models presented here and in Chapter 4 relate to the Bayesian framework.

3.2 Statistical models

To start with, let us put the problem in a familiar framework. If we consider only one chromosome, Table 3.1 shows a 2×2 table arising from a case-control assignment where the allele frequencies are known under the null hypothesis of no association. Also, this design is applicable to admixture mapping under genome-wide levels of mixture and FC animal crosses. Table 3.1 has an interesting structure. The column sums are fixed (deterministic) and known. Assuming allele frequencies are known, under the null hypothesis of no association, the marginal conditional distributions within columns, and within rows, are also known; namely, $\text{Bin}(n, \theta)$.

In case-control studies either allele frequencies or genotype frequencies can be compared in order to evaluate association. In order to compare allele frequencies one

		Y		
		$y=0$	$y=1$	
X	$x=0$	n_{00}	n_{01}	$n_{0.} \sim Bin(2n, 1-\theta)$
	$x=1$	n_{10}	n_{11}	$2n - n_{0.}$
		n	n	

Table 3.1: Contingency table for a case-control study. Y is disease status and X is a genotype. The distribution of the row totals are for the null hypothesis of no association, θ is either a known population admixture proportion or proportion of alleles of interest among all alleles in the sample.

Genotype distribution table

	Number of alleles			
	0	1	2	Total
Case	r_0	r_1	r_2	N
Control	s_0	s_1	s_2	N

Table 3.2: The table shows counts of the genotypes.

needs to treat the two alleles of each individual as independent. This requires that the genotype frequencies to be in Hardy-Weinberg equilibrium (HWE) (Slager and Schaid (2001)). Sasieni (1997) showed that when HWE is violated the allele-based test is biased and the Armitage test (see Section 1.3.1) should be used instead. When HWE holds, the two tests are asymptotically equivalent (Sasieni (1997), Guedj *et al.* (2008)). When an additional condition of the allele effect being codominant holds, Sasieni (1997) demonstrated that the tests are locally most powerful.

Here we assume that the genotypes are in HWE since in practice it is often recommended (Guedj *et al.* (2008)) that a two-stage procedure is applied, in which the markers are first tested for HWE. After this first "quality check" step, they are either used for testing for association if the equilibrium holds, or discarded otherwise.

Now we demonstrate that the likelihoods for both tables are proportional. Let us define a known $\theta = MAF$ at a particular locus. Then, assuming Hardy-Weinberg equilibrium, frequencies of genotypes with zero, one and two alleles of interest are correspondingly $(1-\theta)^2, 2\theta(1-\theta), \theta^2$, see Table 3.3.

Table 3.3 has the same structure but contains frequencies rather than counts.

Genotype fequencies table

	Number of alleles		
	0	1	2
Case	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2
Control	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2

Table 3.3: Genotype distribution table with frequencies. Under the null hypothesis the rows are independent. The table shows the multinomial distribution probabilities for each genotype. θ is MAF of the allele of interest.

Allele distribution table

	Allele of interest	Other	Total
Case	$r_1 + 2r_2$	$r_1 + 2r_0$	$2N$
Control	$s_1 + 2s_2$	$s_1 + 2s_0$	$2N$

Table 3.4: For diploid organisms the sample size is double that of the genotype table.

Under the null hypothesis the rows of the table have the same distribution. The table shows the multinomial distribution probabilities for each genotype. θ is MAF of the allele of interest.

From the two tables the likelihood of the data is:

$$\begin{aligned} L = \quad & [((1-\theta)^2)^{r_0}(2\theta(1-\theta))^{r_1}(\theta^2)^{r_2}] \times [((1-\theta)^2)^{s_0}(2\theta(1-\theta))^{s_1}(\theta^2)^{s_2}] \\ & \propto \theta^{r_1+2r_2+s_1+2s_2}(1-\theta)^{2r_0+r_1+2s_0+s_1}, \end{aligned}$$

and the second line is the form of the likelihood for a 2×2 table with cell probabilities θ and $1-\theta$; so that for each heterozygous individual in Table 3.2 we add one and for each homozygous - two, as is illustrated in Table 3.4.

Genotype data was simulated from a retrospective (case-control) design using a prospective, generative, logistic regression model. First we generated populations using multinomial distribution with probabilities given by Hardy-Weinberg equilibrium for a known MAF. Then disease status of each individual was generated using the logistic regression model

$$Pr(Y|x) = 1/(1 + exp(-\beta_0 - x\beta_1)), \tag{3.1}$$

where Y is disease status, X genotype. This was followed by randomly sampling n cases and n controls.

For our Bayesian model, to test a null hypothesis H_0 versus an alternative H_1 we computed Bayes factors which are marginal likelihoods of the data under both models where the model specific parameters are integrated out. For example, for the alternative model:

$$Pr(H_1|X) \propto Pr(X|H_1)Pr(H_1)$$

where X is the observed random variable (generic, not necessarily disease status as above), and the second term on the right hand side is the marginal likelihood

$$Pr(X|H_1) = \int Pr(X|H_1, \theta)Pr(\theta|H_1)d\theta. \tag{3.2}$$

It has been noted that the natural likelihood that should be used in case-control studies is a "retrospective" likelihood, which is based on the probability of genotype given disease status (Seaman and Richardson (2004)). Hence our likelihood, based on $Pr(X|Y)$, Y being disease and X genotype, uses the right dependency, which is that the observed genotype depends on whether the individual was sampled from cases or controls. For the observed genotypes we considered the binomial likelihood with a conjugate Beta prior for both cases and controls, which were taken as independently sampled. Thus we had two parameters, θ_0, θ_1 with priors Beta(a_0, b_0) and Beta(a_1, b_1). Let us designate the null value for both of them by θ^*, which is a known MAF.

The constraints on the parameter space mentioned in the introduction affect the calculation of the marginal likelihood and can be formulated as: if $\theta_1 > \theta^*$ then $\theta_0 < \theta^*$; and vice versa if $\theta_1 < \theta^*$ then $\theta_0 > \theta^*$. These constraints are induced by the retrospective design and can be proved in a few ways but the simplest is

probably a proof by contradiction presented below.

Let us ignore sampling error and assume that cases' and controls' parameters are equal to the population's minor allele frequencies among diseased and disease-free groups, which should be very close to the reality if sampling is properly randomised. Consider the case when the allele of interest increases the risk of disease. Then the frequency of the susceptibility allele should be higher among the diseased individuals than in the whole population, ie $n_1/N_1 = \theta_1 > \theta^*$, where n_1 is the number of alleles of interest among the diseased part of the population, which is of size N_1. Imagine now that the frequency of the susceptibility allele is higher than that of the whole population among disease-free individuals as well, $n_0/N_0 = \theta_0 > \theta^*$. For a population split into two groups MAF is

$$\theta^* = \frac{n_0 + n_1}{N_0 + N_1} = \frac{\theta_0 N_0 + \theta_1 N_1}{N_0 + N_1} = \theta_0 \frac{N_0}{N_0 + N1} + \theta_1 \frac{N_1}{N_0 + N_1}$$

On the other hand it is equal to

$$\theta^* = \theta^* \frac{N_0 + N_1}{N_0 + N_1} = \theta^* \frac{N_0}{N_0 + N_1} + \theta^* \frac{N_1}{N_0 + N_1}$$

Comparing the two expressions we arrive at a contradiction $\theta^* > \theta^*$ because in the first expression each component is larger. Thus, if $\theta_1 > \theta^*$ then θ_0 must be less than θ^*. By the same reasoning it can be proved that if $\theta_1 < \theta^*$ then $\theta_0 > \theta^*$. A simulated demonstration of this property is shown in Figure 3.1.

The above argument leads to a restricted integration region in (3.2). This constrained integration, together with the Beta priors' conjugacy, means that the numerator of Bayes factor becomes a sum of products of two incomplete Beta integrals:

$$\begin{aligned} Pr(Y|H_1) \quad \propto \quad & \int_0^{\theta^*} \theta_0^{a_0^*-1}(1-\theta_0)^{b_0^*-1} d\theta_0 \times \int_{\theta^*}^1 \theta_1^{a_1^*-1}(1-\theta_1)^{b_1^*-1} d\theta_1 + \\ & \int_{\theta^*}^1 \theta_0^{a_0^*-1}(1-\theta_0)^{b_0^*-1} d\theta_0 \times \int_0^{\theta^*} \theta_1^{a_1^*-1}(1-\theta_1)^{b_1^*-1} d\theta_1, \end{aligned} \tag{3.3}$$

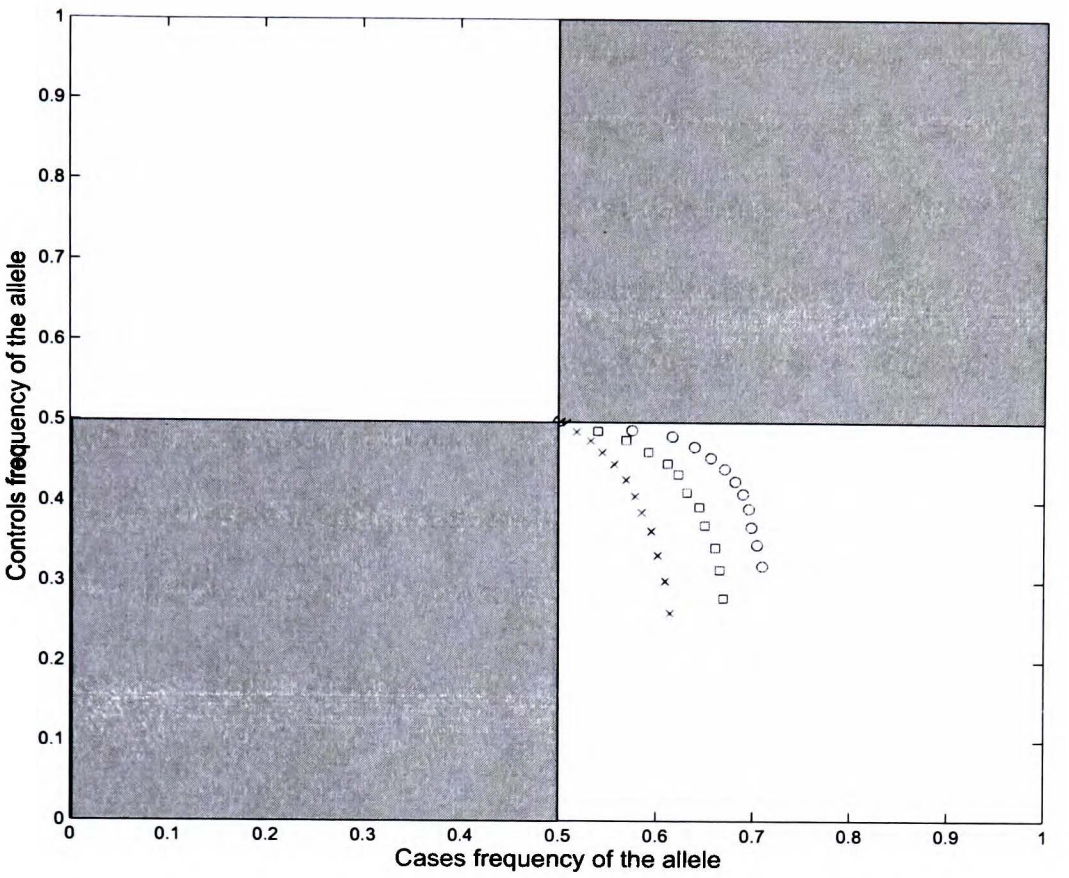


Figure 3.1: Empirical proportions of the allele of interest in the sample. Each series, circles, squares and crosses, represents different values of prevalence, 8%, 18% and 37% respectively, with increasing genetic effect, $\theta^* = 0.5$. The allele of interest was assumed to increase the probability of being diseased, which excludes the coloured part of the graph from the parameter space.

where the posterior parameters are $a_0^* = a_0 + s_1 + 2s_2, b_0^* = b_0 + s_1 + 2s_0, a_1^* = a_1 + r_1 + 2r_2, b_1^* = b_1 + r_1 + 2r_0$ (see Table 3.4), and the normalising factors were omitted for clarity. Incomplete Beta integrals can be efficiently computed, for example, via a standard function in the software package Matlab. This provides a great advantage in terms of computational efficiency and accuracy over MCMC methods.

We ran an automated procedure that went through a set of values of prior Beta parameters for controls and estimated power for a fixed genetic effect that roughly corresponded to 50% power of detection. In our search we kept cases' parameter values fixed at $a_1 = b_1 = 1$ which corresponds to the uniform distribution. For sample sizes $N = 500$ and $N = 250$ the procedure set $a_0 = b_0 = 3000$ and these provided an improvement in power over a number of standard frequentist procedures. As the sample size increases the values of the controls' parameter deviate from the

null very little, hence the procedure set tighter priors. For $N = 1000$ we found $a_0 = b_0 = 4700$ provided an improvement. The optimisation procedure was run for fixed $MAF = 0.5$, but for loci where $MAF \neq 0.5$ we propose to rescale the prior to have mean equal to MAF but with the same variance as the $a_0 = b_0 = 4700$ setting. This value may seem extremely large, however, observe Figure 3.2 where we see that priors with such parameters have variance that is not insignificant. We have not fully studied the optimal settings of these to maximise power though clearly one could do this through extensive large scale simulation over a fine grid of values.

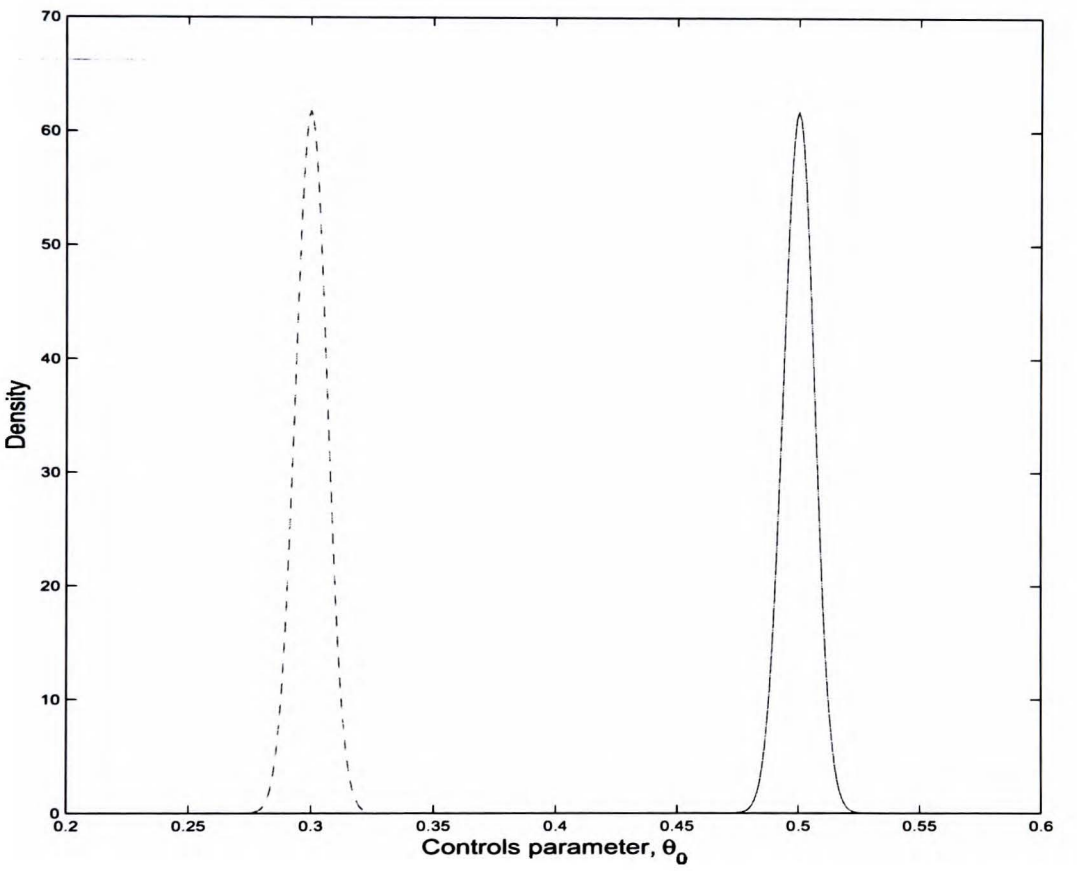


Figure 3.2: Beta prior on controls, $a_0 = b_0 = 3000$. Solid line represents prior with the optimised parameters when prior mean is 0.5, $a_0 = b_0 = 3000$. Dashed line shows the prior with parameters recalculated for the mean equal 0.3, $a_0 = 1512, b_0 = 3527$. The two priors have equal variance.

For smaller sample sizes, e.g. $N = 100$, one can use a method that has better precision than power simulations and is computationally efficient. Each possible dataset is uniquely defined by a combination of two numbers when the column totals are fixed, the number of cases with $x = 1$ and the number of controls with

$x = 1$, $((n_{11}, n_{10}))$. For each combination we calculated the Bayes factor. Then, for a given genetic effect β_1 and assuming the data-generating model (1), one can calculate the probability of each combination. This was carried out under the null conditions and determined the combinations that lay outside the 95% region. Then, because we know the theoretical probability of each combination (n_{11}, n_{10}) under the alternative, the power is easy to calculate: it is the sum of probabilities (under H_1) of all combinations from the rejection region.

An illustration to the method is given in Figure 3.3. The graph also demonstrates how shrinkage of the controls' prior gives an advantage in power: one can see in the graph that the dots which correspond to the shrunk prior on controls, have a rejection area of the same size as the uniform priors model but one that has a larger intersection with the contours of H_1 and therefore a higher power.

3.3 Results

We investigated power curves for a range of scenarios covering different study sample sizes n, disease prevalence λ and MAFs θ. In particular for n cases and n controls, $n \in \{250, 500, 1000\}$; with prevalence $\lambda \in \{0.12, 0.18, 0.26\}$; and MAF, $\theta^* \in \{0.1, 0.3\}$ we performed the following experiments.

The prevalences we have considered here do not represent the whole range of prevalences of common diseases and conditions. The main purpose of the presented results is to demonstrate the new approach. The performance of the approach should be tested on smaller prevalences since common diseases with a possible genetic basis often have much lower prevalences. For example, the most common cancer among women, breast cancer, had a prevalence of 3.6% among women aged 60-69 in the United States in January 2006 (Horner *et al.* (2006)). Also, sample sizes of case-controls studies are increasing and it would be of interest to investigate performance of the models when $N > 1000$.

We considered six common frequentist tests available for testing this data, namely,

1. χ^2 test of homogeneity, 1 degree of freedom (1df);

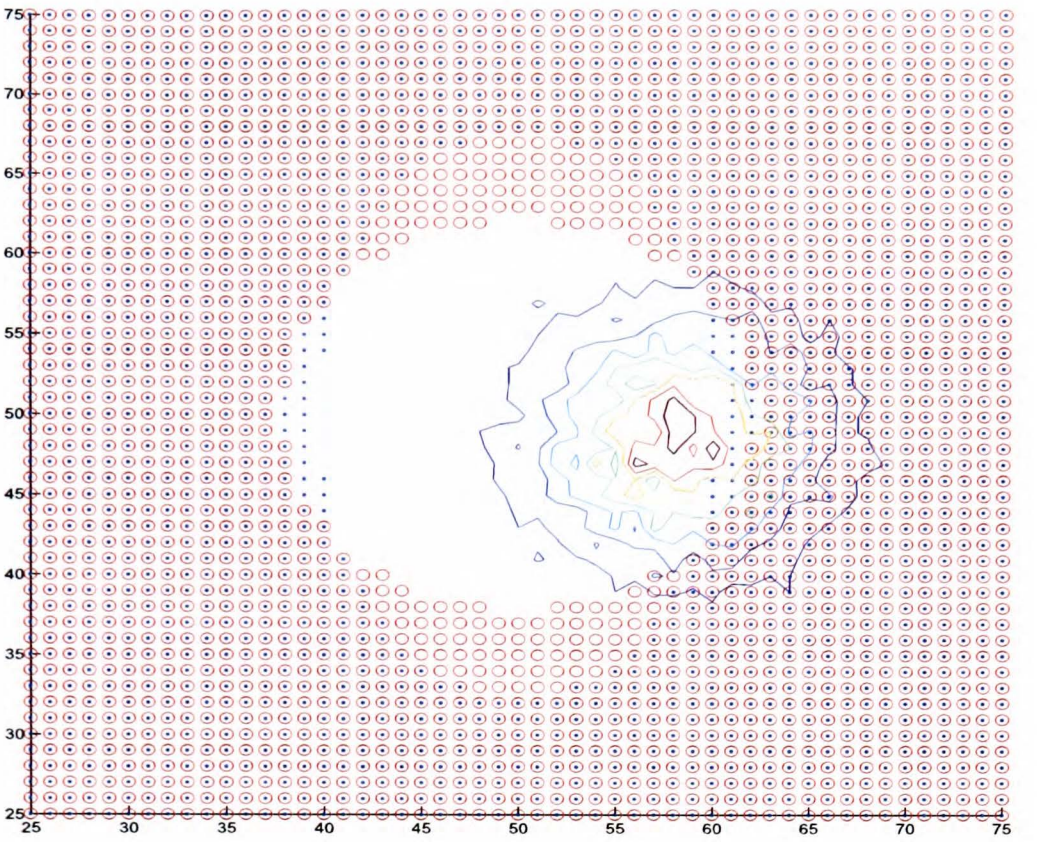


Figure 3.3: Rejection regions for two models with different priors. Circles represent the combinations of (n_{11}, n_{10}) that lead to rejection of H_0 for the model with uniform priors on both cases and controls parameters. Dots are for the model with uniform prior on cases and Beta prior with $a_0 = b_0 = 100$ on controls. The contour plot shows the distribution of combinations under H_1: logistic model parameters $\beta_0 = -2$ and $\beta_1 = 0.4$. The axis show the number of individuals with genotype $x = 1$. $N = 100, \theta^* = 0.5$.

The test estimates the probabilities of an allele, from a case or a control individual, falling into the first and second columns of Table 3.4. Namely, $\hat{\pi}_1 = n_{.1}/4N$ for allele of interest and $\hat{\pi}_2 = n_{.2}/4N$ for the other allele, where $n_{.j}$ is a column sum. Then those probabilities are used to calculate the expected values of the four cells, $E_{ij} = 2N\hat{\pi}_j = n_{.i}/2$, which results in the statistics being: $T_1 = \sum(O_{ij} - E_{ij})^2/E_{ij} = \sum(O_{ij} - n_{.j}/2)^2/(n_{.j}/2)$

2. Logistic regression;

   Generalised likelihood ratio test (LRT) using a prospective logistic regression. This test compares two models for the i's individual from the sample, M_1 : $logit(P(y_i|x_i)) = \beta_0 + x_i\beta_1$ and $M_0 : logit(P(y_i|x_i)) = \beta_0$. The test statistics

uses ML estimates via logistic regression likelihood for both models and is χ^2_1 under H_0:

$$T_2 = 2ln(\max_{\beta_0,\beta_1} L_1(\beta_0, \beta_1|Y) / \max_{\beta_0} L_0(\beta_0|Y))$$

3. Pearson's χ^2 with known cell probabilities, 2df;

   In contrast to the test of homogeneity, Pearson's test uses the known null frequencies and estimates the expected values of the cells directly: $E_{i1} = 2N\theta, E_{i2} = 2N(1-\theta)$, which are plugged into the chi-squared test as well: $T_3 = \sum(O_{ij} - E_{ij})^2/E_{ij}$

4. Cases and controls test;

   Generalised LRT for Binomial test on $Pr(X|Y)$ with two independent parameters for cases and controls. The Bernoulli test is again a generalised LRT, where for the composite H_1 the parameter is maximised over the whole parameter space and for the simple H_0 it is a fixed value θ_0:

$$LRT = 2ln(\max_{\Theta} L(\hat{\theta}|Y)/L(\theta_0|Y))$$

   Sets of cases and controls are treated as independent and thus the test statistics is a sum of LRT for cases and controls, and has χ^2_2 distribution under the null: $T_4 = LRT_{case} + LRT_{contr}$

5. Control-only test;

   The control-only test is equal to LRT_{contr} from the above test and hence have χ^2_1 distribution under H_0

6. Case-only test;

   Equals to LRT_{case} and follows χ^2_1 under H_0

Note that under the null hypothesis of no association, $\theta_{\text{cases}} = \theta_{\text{controls}} = \theta^*$. Two pairs of the tests described above produced equivalent results: the test of homogeneity (test 1) is equivalent to the LRT for logistic regression (test 2); Pearson's χ^2 (test

N	MAF	Prevalence	β_1	Bayesian	Case-only	Logistic	Case-Control	Control-only
250	0.1	0.12	0.35	0.614	0.582	0.417	0.484	0.067
250	0.1	0.18	0.35	0.542	0.501	0.411	0.424	0.077
250	0.1	0.27	0.42	0.605	0.524	0.525	0.479	0.140
250	0.3	0.12	0.24	0.594	0.564	0.416	0.473	0.064
250	0.3	0.18	0.28	0.665	0.625	0.545	0.556	0.101
250	0.3	0.27	0.28	0.563	0.509	0.537	0.486	0.152
500	0.1	0.12	0.25	0.605	0.564	0.416	0.490	0.056
500	0.1	0.18	0.25	0.514	0.478	0.410	0.431	0.070
500	0.1	0.27	0.30	0.619	0.526	0.539	0.506	0.129
500	0.3	0.12	0.15	0.497	0.482	0.344	0.395	0.061
500	0.3	0.18	0.18	0.600	0.554	0.458	0.479	0.079
500	0.3	0.27	0.18	0.505	0.456	0.456	0.416	0.115
1000	0.1	0.12	0.17	0.568	0.564	0.397	0.464	0.067
1000	0.1	0.18	0.18	0.538	0.518	0.412	0.436	0.081
1000	0.1	0.27	0.21	0.606	0.543	0.525	0.494	0.134
1000	0.3	0.12	0.12	0.581	0.575	0.419	0.480	0.060
1000	0.3	0.18	0.13	0.570	0.544	0.446	0.468	0.081
1000	0.3	0.27	0.13	0.502	0.451	0.446	0.414	0.117

Table 3.5: Empirical power of detecting an association at various simulation parameter values. The value of β_1 denotes the effect size, converted to log-odds, defined in equation 3.1; chosen here to produce power close to 50%. The null hypothesis was tested at nominal level 5%, power was estimated from 10,000 simulations, Monte Carlo standard error ≈ 0.002

3) is equivalent to the cases and controls test (test 4). Since the test of homogeneity is classified as Pearson's test as well, this corresponds to the theoretical result of asymptotic equivalence of Pearson's test and LRT test (Hosmer and Lemeshow (2000)). Thus we omit results for tests 1 and 3 in Table 3.5, and in estimated power graphs they always have the same values.

Results of the simulations are shown in Figs 3.4 - 3.6 and in Table 3.5; the individual graphs in Figs 3.4 - 3.6 are reproduced in the Appendix A for greater detail. The table lists estimates of power at a value of β_1 for which the case-only test's power is close to 50%. The curves in the graphs are constructed using 10,000 generated data sets at each of a number of points. For example, for $n = 500$, $MAF = 0.3$ and prevalence of 12% the points are:

$\beta_1 \in \{0.0, 0.03, 0.06, 0.09, 0.12, 0.15, 0.18, 0.21, 0.24, 0.27, 0.30\}$.

For the frequentist tests we observe some interesting behavior. For rarer disease simulations the case-only-discard-controls design is the most powerful amongst frequentist tests. That is, we are lead to actively discard the control data even though it is available and contains information, as can be seen by the shallowly increasing power curve for control-only (squares in Figs 3.5 - 3.6).

The reason that case-only-discard-controls is most powerful among frequentist tests is that for lower prevalence the cases data contains much more information than the controls as they are greatly oversampled relative to controls due to the retrospective nature of the design. The information imbalance is so great that the extra variability from the increased degree of freedom of the 2 degree of freedom case-control test relative to the one degree of freedom case-only test is not enough to generate an increase in power. As the prevalence increases so does the relative information in the controls until at some point the information gain outweighs the extra degree of freedom in the test at which point the logistic model based test proves most powerful.

This behaviour can be observed in Figures 3.4 - 3.6 (see also the Appendix). For the lower prevalence of 12% (graphs at the bottom) the case-only test (solid line) has an advantage over all other frequentist tests. Consider next the graphs above, with 18% prevalence, and we note that all frequentist test, with 1df and 2df, get closer to the case-only test in terms of power. In Figure 3.4, for $N = 250$, we see that at the highest investigated prevalence of 27% the 2df Pearson's and case-control tests (crosses, and dashes and dots) outperform the case-only test. The crossover point is determined by the MAF and the underlying prevalence.

3.4 Applications of the method

In this section we discuss applicability of the introduced model to real datasets. As explained in the introduction to this chapter, we envisage the model becoming more and more relevant in the future, as genotyping costs decrease and the number of genotyped people and subjects of animal models increase. Here we describe the

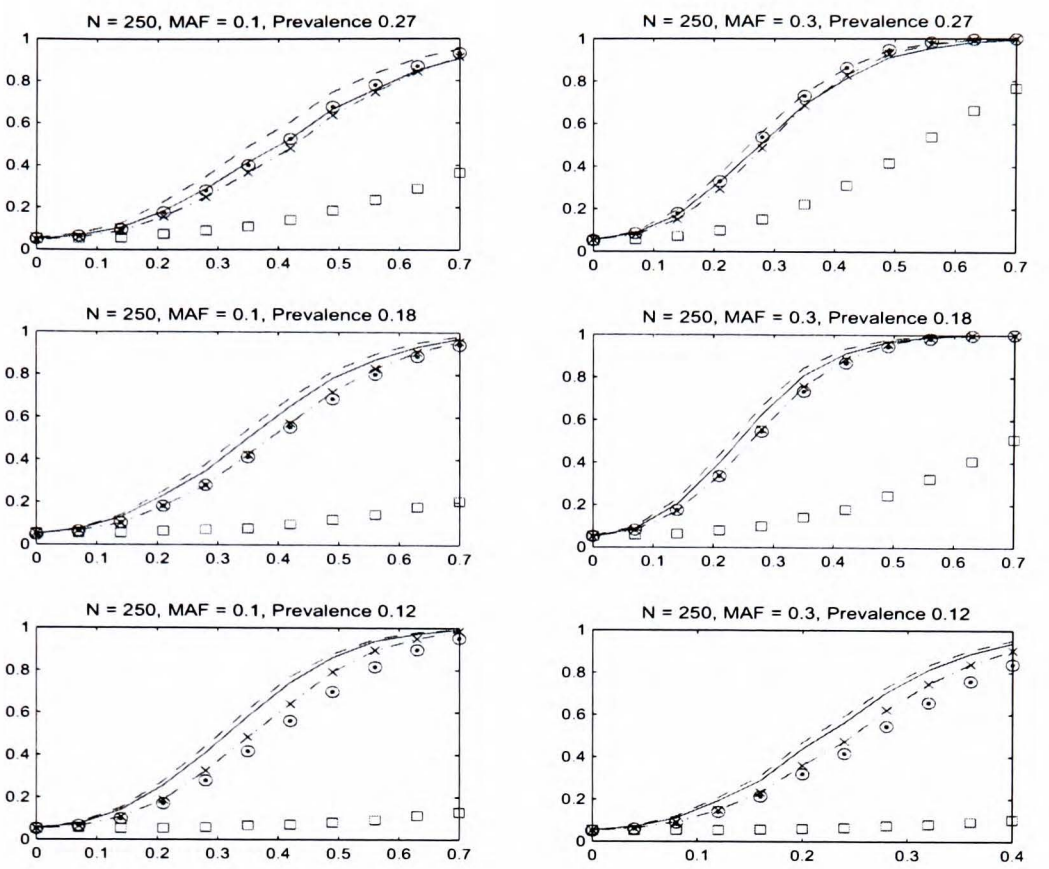


Figure 3.4: Estimated power of the six frequentist tests and Bayesian test for $N = 250$. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

current difficulties and prospects for two situations in which our model can be used: when estimates of allele frequencies of relatively small and isolated populations are available, and when we have estimates of admixture proportions and locus ancestry.

In the post human genome project era there is an increased interest in large-scale cohort studies with a focus on the effect of genetic variation on the common human diseases (Davis and Khoury (2007)). Population-based biobanks have distinct advantages over resource-intensive large-scale cohort studies, and many countries around the world are currently considering such projects. Austin *et al.* (2003) has already identified plans for eight biobanks: in Iceland, the United Kingdom, Estonia, Latvia, Sweden, Singapore, the Kingdom of Tonga, and Canada. The authors note that the characteristics of the projects varied but all intended to map genes for common diseases. There are also several major biobank networking initiatives

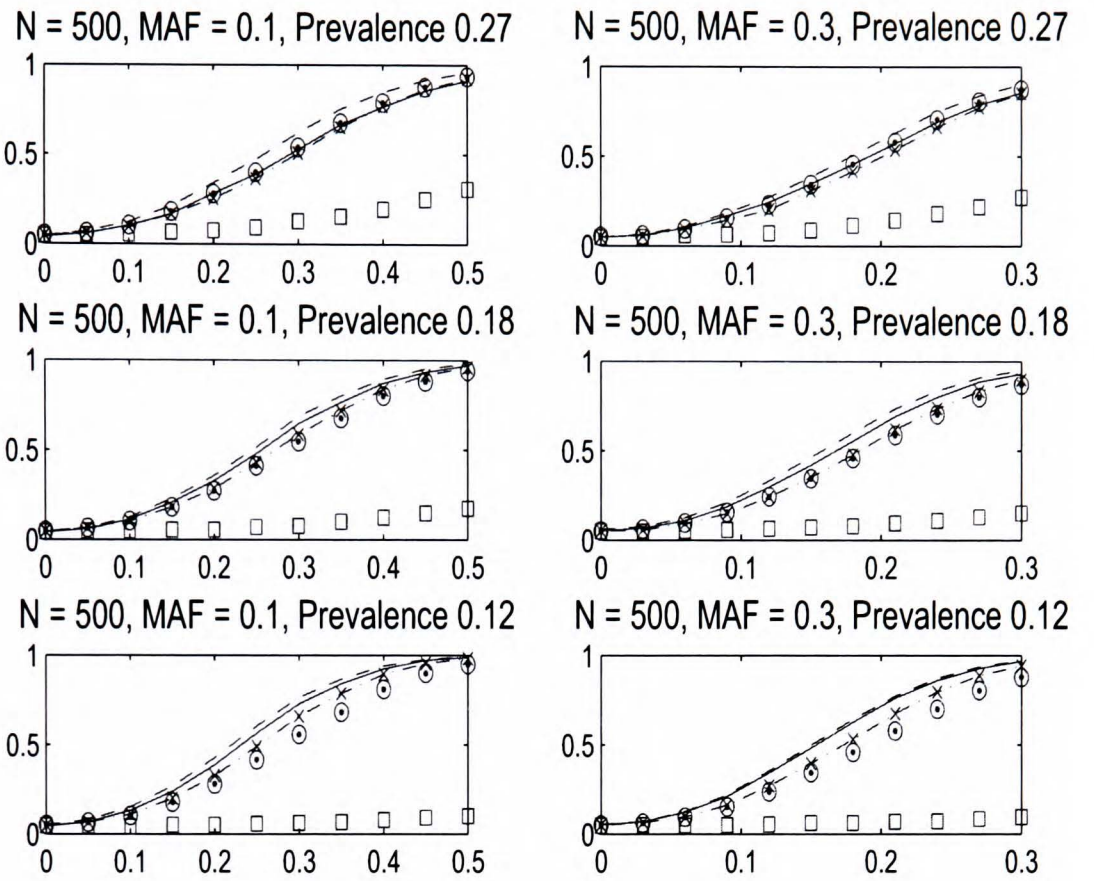


Figure 3.5: Estimated power of the six frequentist tests and Bayesian test for $N = 500$. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

worldwide (Asslaber and Zatloukal (2007)).

Austin *et al.* (2003) provide a summary of the planned biobanks. In smaller countries such as Iceland and Estonia implementation of the projects would result in acquiring genetic material from a large proportion of the population. The Estonian Genome Project is planning to genotype approximately one million Estonians in the country of 1.4 million people. The DeCode project in Iceland owns a sample of genotypes of about 100,000 volunteers, having also purchased the Icelandic Genealogic Database from its government (Hagen and Carlstedt-Duke (2004)). Helgason *et al.* (2005) report the population size of Iceland to be 290,000.

In addition to the Icelandic and Estonian projects, there are other large collections of genetic material (Hagen and Carlstedt-Duke (2004)): the Genomics Collaborative has data on 120,000 patients; the European Prospective Investigation

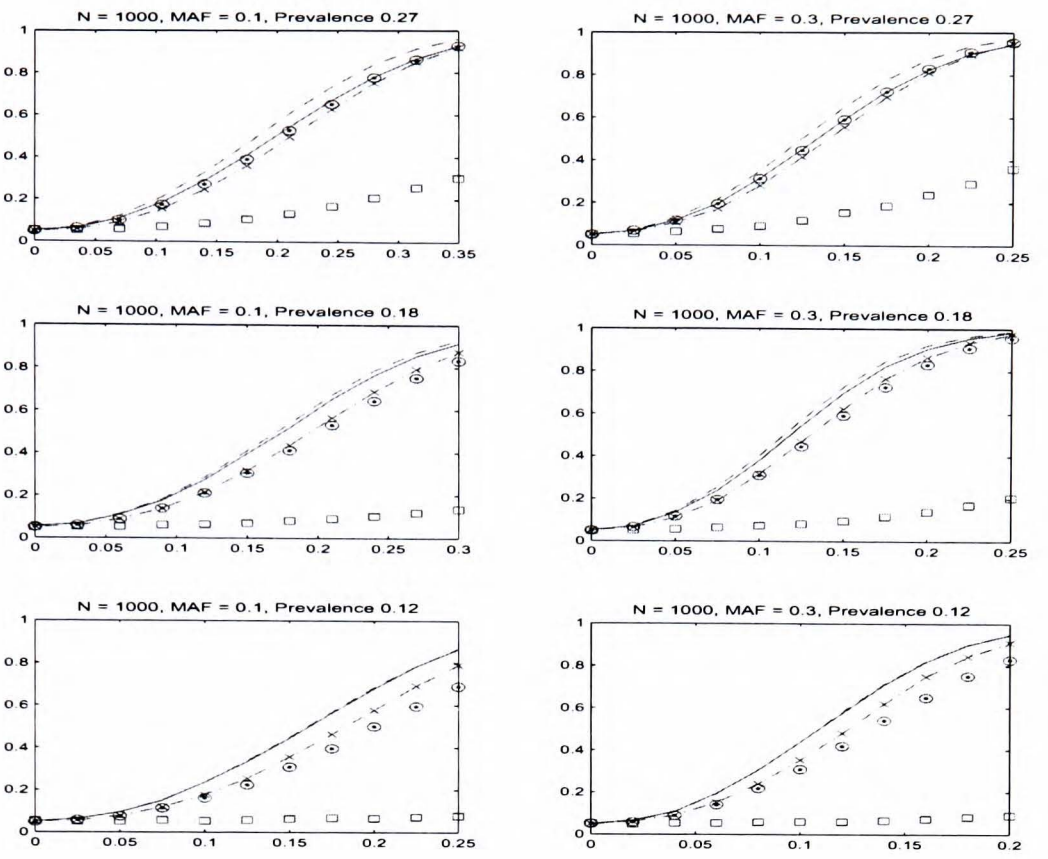


Figure 3.6: Estimated power of the six frequentist tests and Bayesian test for $N = 1000$. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

into Cancer and Nutrition has material from 520,000 individuals; the Clinical Trials Service Unit at the University of Oxford collected samples from 250,000 individuals.

Thus, there are large and growing deposits of genetic data which can help to establish populations' allele frequencies. Scandinavian populations are relatively small and live in relatively isolated regions, so the planned and implemented biobanks could provide very good estimates of the population-wide allele frequencies for some loci. However, even in these favourable cases there are some caveats. Lewis and Brunner (2004) carried out meta-analysis of association studies of longevity and apolipoprotein E gene and found that allele frequencies varied considerably not only between nations but within countries as well (e.g. Finland). Iceland had been considered as a relatively homogeneous (Helgason *et al.* (2003)) genetic isolate (Cavalli-Sforza *et al.* (1994)) but Helgason *et al.* (2005) demonstrated that there

is a substructure that can cause spurious association. On using populations' allele frequencies as the null value for our model, we conclude that however small the studied population is one would need to check for the presence of a substructure.

More generally, it can be challenging to define objectively what constitutes a population. While it may seem logical to define a population based on physical or geographical characteristics, it would be difficult to determine whether selection using these criteria reflects, in genetic terms, a naturally occurring assignment (Pritchard *et al.* (2000)). Ewens and Spielman (1995) noted that subjective classifications can miss "cryptic" population structure, and showed how this undetected structure can invalidate standard tests in association studies, although the extent to which it is a genuine cause of false positives has been the topic of much debate (Balding (2006)). Genetic information can be used to detect the presence of hidden population structure, which was considered in Pritchard and Rosenberg (1999) in association mapping context. Later Pritchard *et al.* (2000) developed a Bayesian clustering method that estimates allele frequencies of the constituent populations.

Another context in which our model is relevant is admixture mapping studies. Here we closely follow Hoggart *et al.* (2004), whose approach is based on comparing the observed and expected proportions of gene copies from an ancestral population. The authors combined data from admixed and unadmixed populations, and showed that locus ancestry and admixture proportions (defined as the proportion of the genome that have ancestry from each subpopulation) can be measured almost without uncertainty. Hoggart *et al.* (2004) note that this can be achieved if a genomewide panel of ancestry-informative markers is typed, and regions of putative linkage are then saturated with additional markers to extract information about locus ancestry.

For a binary disease the probability of observing locus ancestry from the high-risk population in an affected individual is (Hoggart *et al.* (2004)):

$$z = \frac{\gamma\sqrt{r}}{\gamma\sqrt{r} + 1 - \gamma}, \tag{3.4}$$

where γ is admixture proportion from the high risk population, r is the ancestry

risk ratio, or the risk ratio for disease in individual with 2 versus 0 gene copies and who have ancestry from the high-risk population at the locus. To apply our model in this context we only need to redefine the parameter θ from section 3.2 to signify the probability of locus ancestry instead of MAF. Our model only requires the null value for this parameter and under the null hypothesis $r = 1$, and so 3.4 simplifies to $z = \gamma|H_0$.

There is, however, uncertainty in measurements of admixture proportion, locus ancestry, and allele frequencies. Now we examine how this uncertainty affects the relative performance of the tests considered in this chapter. We use $Beta(a, b)$ distribution to model this uncertainty, similarly to Ambrosius *et al.* (2004) in their study of effects of allele frequencies uncertainty on power of association studies.

In order to investigate the tests' behaviour under uncertainty we added one step to the case-control simulation procedure: before simulating a population's genotypes the allele frequency was drawn from a Beta distribution rather than kept fixed as before. Parameters of the Beta(a,b) distribution for allele frequency were chosen so that it was centered at the previously fixed value ($maf = 0.1$), and we investigated two standard deviations, 0.005 and 0.01. Results of the power comparison are ploted in Figure 3.7, where we observe that for random MAFs:

- The null distributions of the frequentist tests which rely on the null value of MAF do not follow their standard χ^2 distribution (the solid line in Figure 3.7 does not cross the Y axis at 5%). The Bayesian test, Pearson's test, and LRT for logistic regression still follow their standard null distributions because they do not require a MAF under the null.

- As the variance of the distribution of MAF becomes larger, the frequentist tests which do not require the null value of MAF (Pearson's and logistic LRT) become more powerful that the Bayesian test (Figures 3.7 and 3.8)

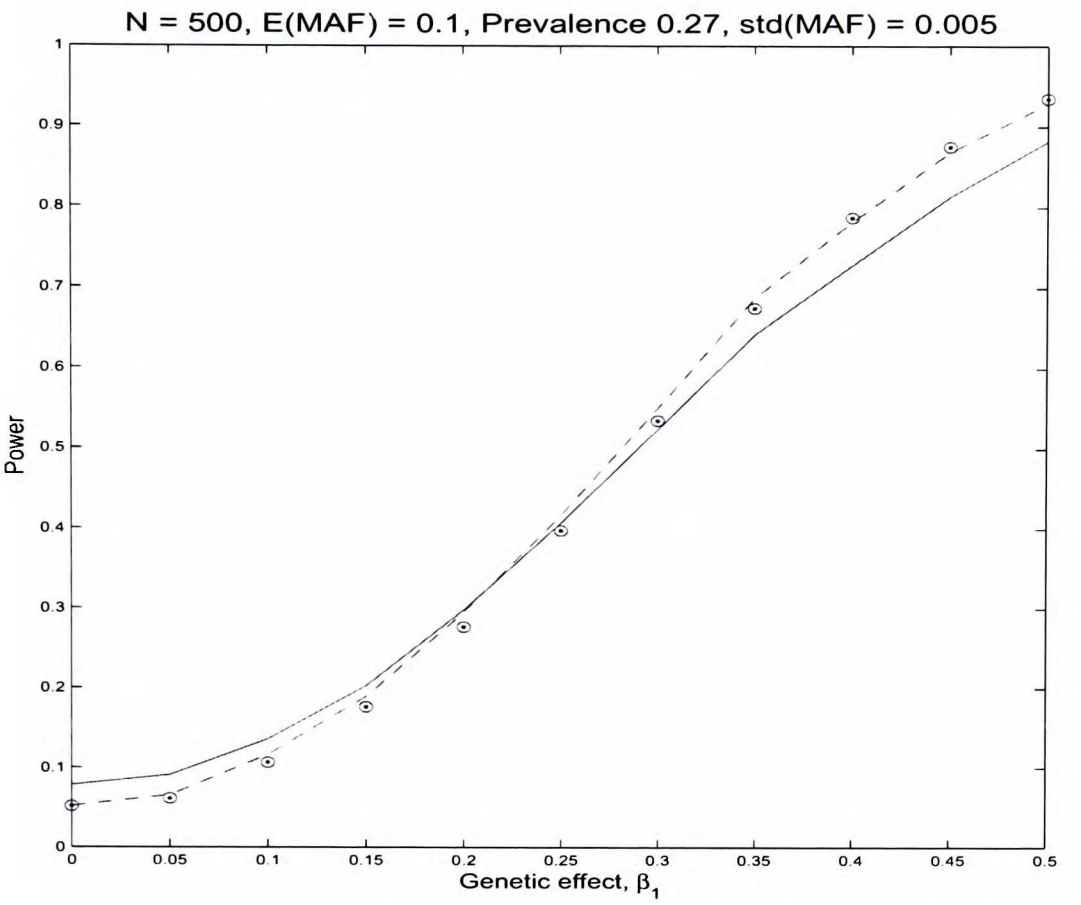


Figure 3.7: Statistical power for MAF drawn from Beta distribution with the expectation equal 0.1 and the standard deviation equal to 0.005. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity. Power estimates for case-only test are not valid and is displayed for illustration. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

3.5 Discussion

We have shown for the case of known population allele frequency that, based solely on frequentist measures, a Bayesian test outperforms conventional frequentist tests. This highlights the strengths of the Bayesian paradigm in being able to accomodate via prior distributions the relative information gain that cases have over controls in a retrospective ascertainment. The retrospective case-control ascertainment also induces a contraint on the parameter space in the Bayesian prior which leads to an increase in power.

The Bayesian test can accommodate the unequal relative information in the two categories by adopting a prior which has greater concentration around the null for controls than for cases. This shrinks the Bayes estimates for controls more heavily

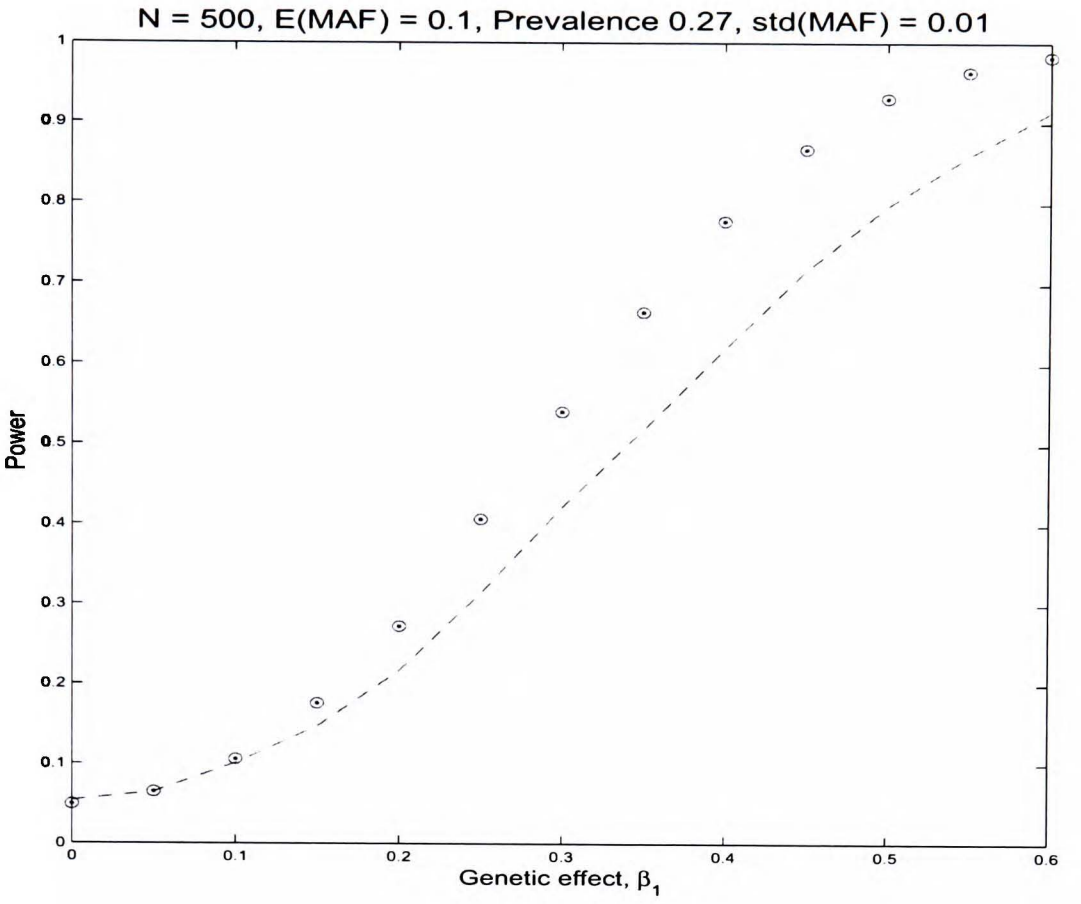


Figure 3.8: Standard deviation of MAF equals 0.01. Dashed line represents the Bayesian test, circles - logistic, dots - test of homogeneity. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

than cases and adjusts for their extra variability. We see that the resulting Bayes test appears uniformly more powerful than all frequentist tests we applied and that only a single test is needed for any prevalence.

It is interesting to investigate the relative contribution of optimal prior parameters and taking into account constraints on the parameters. We compared our models to that of using a uniform prior on both cases and control allele frequency as suggested by Balding (2006) without constraints on the parameter space. Note the uniform prior is a special case of the Beta prior we consider. From Figure 3.9 we can see that the shrinkage of the controls towards the population MAF (lines without circles) produces a much larger increase in power than the constrained integration over a smaller parameter space; though this last feature still provides some gains.

Further development of the results in this chapter could include investigation of how the linkage between allele of interest and marker affects power of the tests

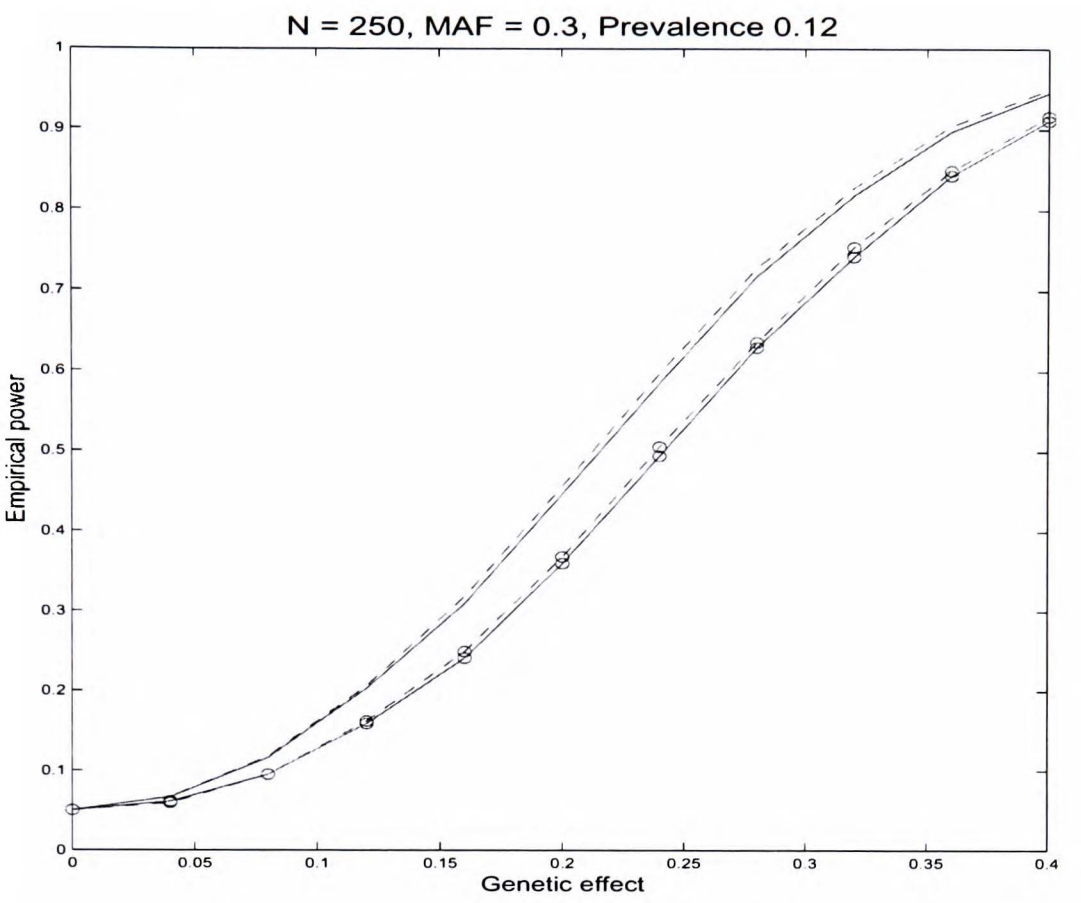


Figure 3.9: Comparison of statistical power for optimal versus uniform priors; and incomplete integration versus full integration. Dashed line represents the Bayesian test with optimal priors and incomplete integration, solid - optimal priors (on both cases and controls) and full integration, dashed with circles - uniform priors and incomplete integration, solid with circles - uniform priors and full integration. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

considered here. This study tested for association between the phenotype and a marker and we have not investigated the effect of this fact on power to detect the "true" effect. However, Zondervan and Cardon (2004) provide very clear *deterministic* expressions for the relationship between marker odds ratios and the true disease-causing allele odds ratio, which depends on the disease and marker allele frequencies and linkage disequilibrium between them. Also the article considers the effect of the joint influence of those parameters on the statistical power to detect the true genetic effect. It would be interesting to examine whether the results of this chapter would still hold true (in terms of comparative power analysis) if the formulae were applied.

Chapter 4

Hypothesis testing of association when MAF is not known

In this chapter we consider a single SNP analysis as part of genome-wide association studies again but for a more common situation, when minor allele frequency is unknown. Our findings include the following:

- Application of optimal Bayesian methodology to GWAS with optimised prior parameters provides a test for association with more power than the most powerful frequentist test.
- The difference in power is greater for rare alleles and thus our models can be a very useful tool when the "rare variant model" hypothesis is deemed more appropriate than the common-disease/common-variant hypothesis.
- We demonstrate the importance of using constraints on parameter arising from retrospective ascertainment.
- We show how optimisation of parameters provides more power by comparing our model to another Bayesian model recently used in the largest association study yet conducted.

4.1 Statistical models

Genome-wide association studies, and single SNP analyses on which they are based, are becoming increasingly important for localising genetic variants that convey differential disease risk. As such, the statistical methods used to test for association, being critical to the performance, are also of high importance.

Historically there has been a fierce debate between frequentist and Bayesian methods for conducting the statistical analysis. The structure of the frequentist framework was established between the two world wars (Krzanowski (2007)). Many statisticians were unhappy with the the fact that results have to be interpreted in terms of long-run frequencies. Some frequency-based results were misinterpreted as probability statements. Dissatisfaction with the frequentist framework led to adopting interpretation of probability as a degree of belief, and therefore to a Bayesian approach. The early years of Bayesian methods were marked by the fierce opposition of the frequentists whose cause was further supported by the mathematical difficulties involved in calculating some Bayesian results. However, advances in computational techniques towards the end of the last century revolutionised the application of Bayesian techniques, and led to their much wider acceptability. Krzanowski (2007) gives a very clear comparison of the two frameworks. Fundamentals of the Bayesian approach can be found in Bernardo and Smith (1994), Berger (1993), Robert (2001), to name only a few among many good sources.

In this chapter we introduce optimal Bayesian procedures as an attractive compromise between frequentist and Bayesian approaches when substantive prior knowledge is either unavailable or where non-subjective inference is desired. Optimal Bayesian model carries the benefits of full Bayesian inference in the use of marginal probability calculus, but without the need to specify subjective priors, which are "objectively" set using pre-defined protocols to deliver optimal performance characteristics. Using extensive Monte Carlo simulations and other methods we demonstrate that for the three most widely used GWAS genetic tests the optimal Bayesian modeling leads to procedures with uniformly greater power than the current fre-

quentist (non-Bayesian) methods. Hence, from a purely frequentist argument we are lead to discard frequentist tests in favour of Bayesian methods. It is also shown that optimisation protocols lead to proper priors with a structure that is intuitively understandable. The knowledge of uniformly (up to 13%) greater power and intuitive priors when combined with other reasons to be Bayesian leads us to believe that optimal Bayesian modeling procedures will provide an important framework for scientific inference in genetic epidemiology.

There are many reasons for choosing Bayesian methods for problems of scientific inference. Proponents of Bayesian statistics cite coherency, the proper treatment of nuisance parameters and accounting for all aspects of uncertainty in model specification as some of the main strengths of their doctrine. However, there is also understandable hesitation in using Bayesian methods due to the subjective, personalistic nature of conclusions drawn from data. By way of middle ground, optimal Bayesian modeling procedures provide a principled way to make inference.

Perhaps the most important single difference between Bayesian and non-Bayesian methods when testing for genetic association is in the use of marginal likelihood. Under a null and alternative hypothesis $\{H_0, H_1\}$ the Bayesian calculates the odds of H_1 given the observed data $\mathcal{D}$,

$$\frac{Pr(H_1|\mathcal{D})}{Pr(H_0|\mathcal{D})} = \frac{Pr(\mathcal{D}|H_1)}{Pr(\mathcal{D}|H_0)} \frac{Pr(H_1)}{Pr(H_0)}$$

whereas the non-Bayesian typically looks at some function of the likelihood ratio

$$\frac{Pr(\mathcal{D}|\hat{\theta}_1, H_1)}{Pr(\mathcal{D}|\hat{\theta}_0, H_0)}$$

where θ_a denotes parameters in the model under H_a, $a \in \{0, 1\}$, and the "hat" on $\hat{\theta}_a$ specifies that the parameters typically have to be estimated from the data.

There are a number of important features to note in the above. First, the Bayesian directly answers the question of interest, namely $Pr(H_1|\mathcal{D})$, the probability of H_1 given the observations; whereas the non-Bayesian cannot infer this. In

order to do this, however, a prior specification must be made, which is an inevitable consequence of conditional probability. Second, Bayesians do not condition on things they do not know, namely, $\hat{\theta}_1$ which must be estimated from the data, so in truth the data $\mathcal{D}$ enters on both sides of the conditioning bar in the non-Bayesian likelihood $Pr(\mathcal{D}|\hat{\theta}_1, H_1) \equiv Pr(\mathcal{D}|\{\hat{\theta}_1|\mathcal{D}\}, H_1)$. The Bayesian on the other hand uses the marginal likelihood to quantify the evidence from the data,

$$Pr(\mathcal{D}|H_a) = \int Pr(\mathcal{D}|\theta_a, H_a Pr(\theta_a|H_a) d\theta_a. \tag{4.1}$$

The marginal likelihood does not condition on unknown θ. However, it does require the specification of the prior $Pr(\theta_a|H_a)$ and it is here that the two camps disagree so vehemently. To the Bayesian the specification of the prior allows for the question of interest to be answered and for external knowledge to be incorporated. To the non-Bayesian the use of subjective inference is worrying due to sensitivity of conclusions to personalistic beliefs. Optimal Bayesian modeling offers a prescriptive method for specifying $Pr(\theta_a|H_a)$ thus removing subjective inference from Bayesian calculations.

To make things more concrete we consider the case of testing for an additive effect in a GWAS case-control design on n individuals. We will adopt a prospective logistic regression model as is done in the supplementary material on Bayesian modeling in Balding (2006). According to Seaman and Richardson (2004) and Balding (2006), this model, based on prospective likelihood, should provide power estimates equal to estimates from a more natural retrospective likelihood. The data set is derived from SNP genotyping arrays, $\mathcal{D} = \{y_i, x_i\}_{i=1}^n$ with y_i a case indicator and $x_i \in \{0, 1, 2\}$ denotes the number of copies of the minor allele at a particular marker. SNP markers are predominantly diallelic because of their low mutation rate and many studies include only such SNPs. However, triallelic analyses may become more popular in the near future as interest in deletions, which can be seen as a third allele, is growing.

For now we consider only two alleles, and the logistic regression model specifies that

$$\log\left(\frac{Pr(y_i=1|x_i)}{Pr(y_i=0|x_i)}\right)=\beta_0+x_i\beta_1,$$

and a typical Bayesian prior might then be,

$$\beta_0\sim N(0,v_0^{(a)}),\qquad \beta_1\sim N(0,v_1^{(a)}), \tag{4.2}$$

where $v_i^{(a)}$ is prior variance of β_i under hypothesis H_a. It is interesting to note that the structural part of the prior, namely normality $N(\cdot,\cdot)$, is not usually considered but rather the subjective part is in the specification of $\{v_0^{(a)},v_1^{(a)}\}$. It is perhaps under-appreciated that the independent prior specification in (4.2) is inappropriate for case-control designs. The constraint that the number of cases and controls remains fixed implies that $\beta_0\times\beta_1\leq 0$. That is, that the coefficients must have opposite signs. While this constraint is rarely a problem for non-Bayesians who maximise $\hat{\beta}_0,\hat{\beta}_1$, for Bayesians the evidence $Pr(\mathcal{D}|H_a)=\int\int Pr(\mathcal{D}|\beta_0,\beta_1,H_a)Pr(\beta_0,\beta_1|H_a)d\beta_0\beta_1$ with independent priors (4.2) would place half its measure on a space which has probability zero. To solve this we will consider multivariate priors,

$$H_a:\beta_0,\beta_1\sim N(0,\Sigma^{(a)})I(\beta_0\beta_1\leq 0) \tag{4.3}$$

where $\Sigma^{(a)}=\begin{pmatrix} v_0^{(a)} & 0\\ 0 & v_1^{(a)}\end{pmatrix}$ and $I()$ is the indicator function (see Figure 4.1 for illustration). We show later that incorporating this constraint leads to increased power.

The choice of the hyperparameters, $\{\widehat{v_0^{(a)}},\widehat{v_1^{(a)}}\}$, is driven by our optimal Bayesian modelling approach. Our method seek automated protocols for prior specification which lead to optimal performance (in some well defined sense explained later). In this chapter we consider frequentist statistical power, the probability of rejecting the null when the alternative is true, though other measures could easily be used. To the Bayesian the setting of the hyperparameters is a decision process with respect

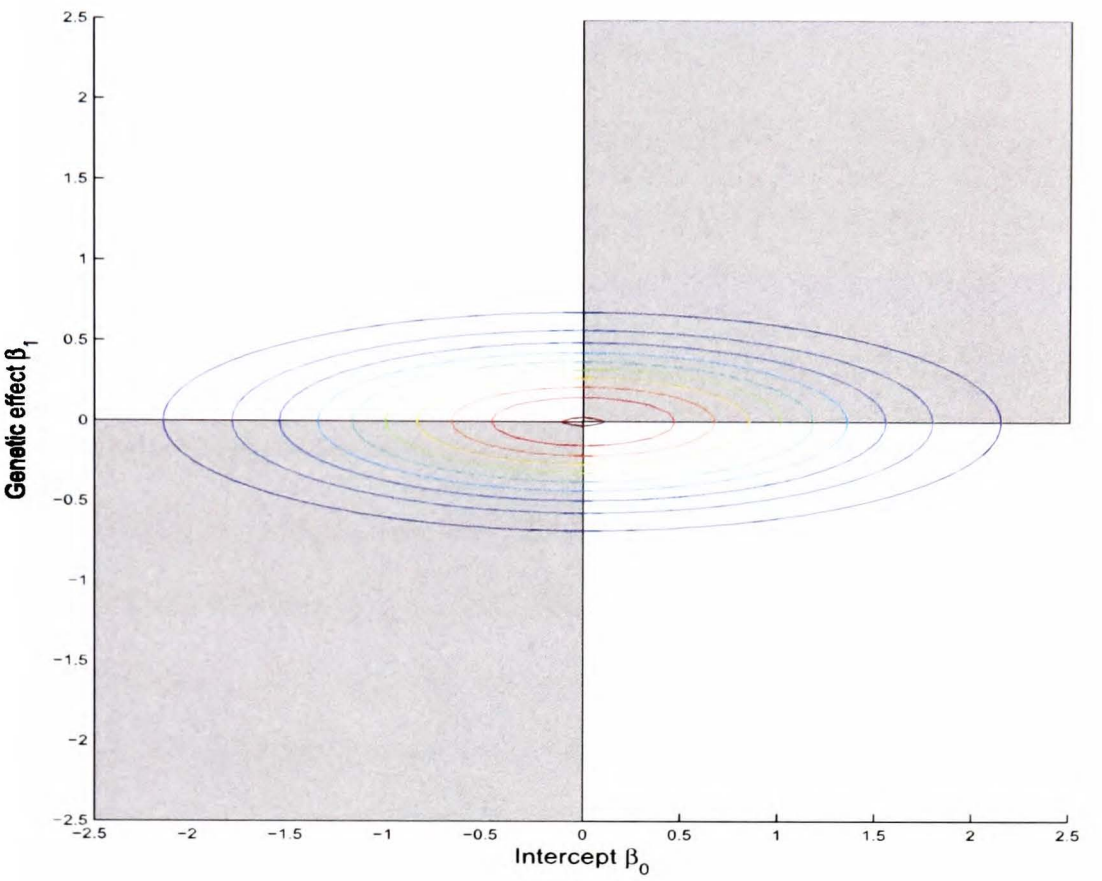


Figure 4.1: Contours of the multivariate normal prior with constraints on the parameter space, defined in equation (4.3). Contours signify parameter values with equal probability density of bivariate normal distribution $N(\mathbf{0}, \mathbf{\Sigma^{(a)}})$, $v_0^{(a)} = 1$, $v_1^{(a)} = 0.1$. The areas in grey have zero probability and are excluded from the parameter space.

to a loss function, in this case one minus frequentist power, for fixed Type I error. That is, we must call H_0 or H_1 based on a threshold that meets a frequentist Type I error rate and seeks to optimise $\{\hat{v}_0, \hat{v}_1\}$ to lower the Type II error. This is a well defined Bayesian decision process that proceeds in the following manner. Suppose one must provide a (Bayesian) model to the GWAS community with a predefined setting of $\{\widehat{v_0^{(a)}}, \widehat{v_1^{(a)}}\}$. The model must classify markers as H_0 or H_1 and a reward is received for calling correctly and a penalty paid for incorrect calls. The ratio of the loss-reward defines the fixed Type I error rate. For GWAS in the absence of prior information across markers we can arbitrarily fix the prior ratio $Pr(H_1)/Pr(H_0)$ to be 1. We then set $\{\widehat{v_0^{(a)}}, \widehat{v_1^{(a)}}\}$ in order to maximise our expected utility (minimise our expected loss) from the community.

We generated 100,000 data sets from the additive model $Pr(y_i = 1|x_i) = \beta_0 +$

$x_i\beta_1$ and then performed a grid-search over the space of realistically possible values of prior variance and identified the values that maximise power at the value of β_1 for which the corresponding Armitage test has approximately 50% power. As pointed out in Balding (2006), for complex traits it is often assumed that contributions to disease risk from individual SNPs are roughly additive and that the Armitage test is the most powerful for such genetic effects because it is specifically tailored for such conditions. Thus, we conducted our comparative investigation under very favourable conditions for the most powerful of the existing frequentist tests.

For calculation of the marginal probabilities in (4.1) we used Laplace's method since the integrals can not be solved analytically. The method uses a second order Taylor expansion of the integrand and provides a very computationally efficient alternative to Monte Carlo simulation for such integrals. As explained in Kass and Raftery (1995), the method is very accurate for large samples, such as 250 for two parameters, when the likelihood peaks near its maximum.

Laplace's method is easily derived by expanding the log of unstandardised posterior $l = log(P(Y|\beta)\pi(\beta|H_a)$ up to the quadratic term about its maximum-a-posteriori β^* and then exponentiating. The resulting approximation has the form of a normal density with mean β^* and covariance matrix $\Sigma = (-H)^{-1}$, where $H = D^2 l(\beta^*)$ is the Hessian matrix. Integration of this approximation yields the estimate for the marginal m:

$$\hat{m} = (2\pi)^{d/2}|\Sigma|^{1/2}P(Y|\beta^*)\pi(\beta^*|H_a)$$

The iterative re-weighted least squares algorithm described in Chapter 2 was used to find the required posterior mode. The algorithm always converged in less than 7 iterations which took less than a second to run. Figure 4.2 demonstrates how well Laplace's method works. For comparison, the Monte Carlo simulation for generating the data in the figure took about 10 minutes.

For our investigation of the dominant and recessive allele effects we generated data using these models. For testing we modified the logistic regression model

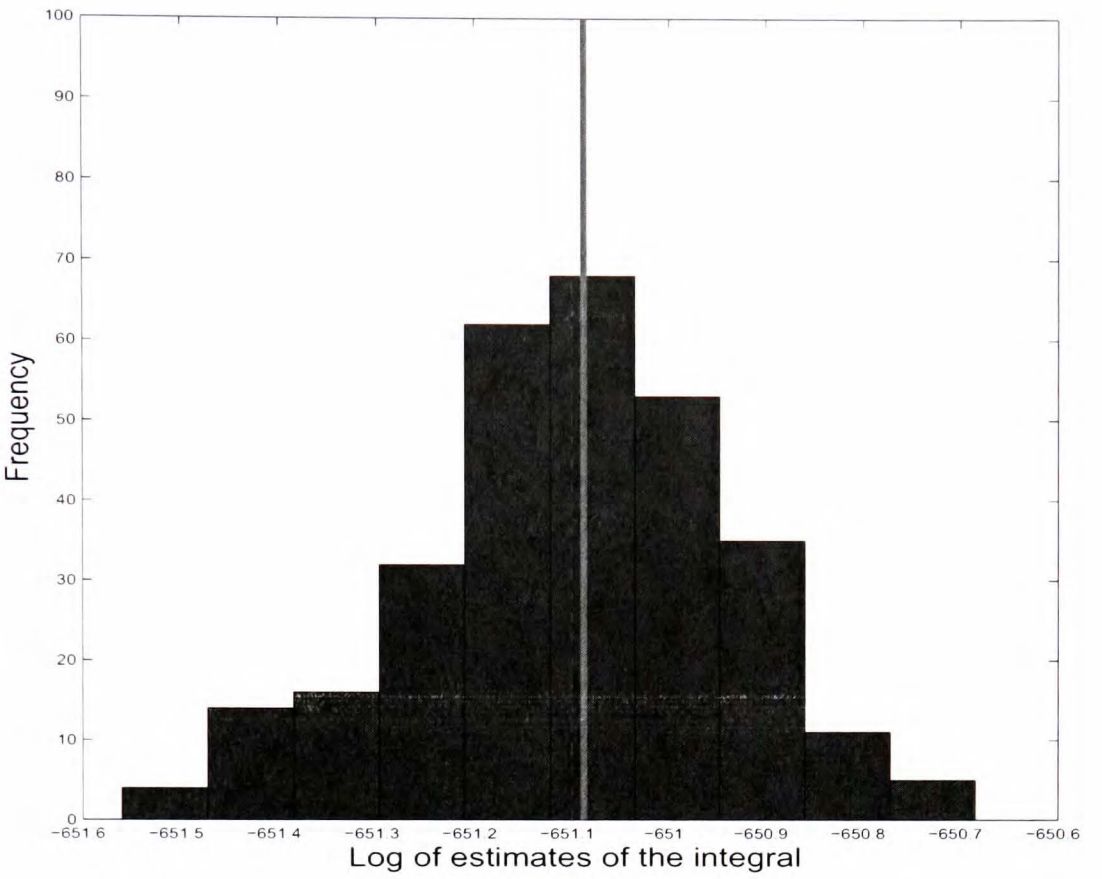


Figure 4.2: Comparison of Laplace's method estimate to Monte Carlo integration. The histogram is of 300 Monte Carlo estimates, with 5000 simulations in each. The thick line is Laplace's method estimate, the dotted line is the Monte Carlo sample's mean. The 500 cases and 500 controls dataset was generated from a population model with MAF=20%, prevalence 10% and $\beta_1 = 0.5$.

accordingly. For the general form of the logistic regression,

$$logit(Pr(Y_i = 1|x_i)) = \beta_0 + \beta_1 I(x_i = 1) + \beta_2 I(x_i = 2), \tag{4.4}$$

we set $\beta_1 = \beta_2$ for the dominant model and $\beta_1 = 0$ for the recessive. For each model of the allele effect we compared the performance of our model with the corresponding version of the Armitage test. For example, using the notation of Figure 4.1, the Armitage test for the additive effect and number of controls equal to the number cases, is defined as (Sasieni (1997)):

	Number of alleles			
	0	1	2	Total
Case	r_0	r_1	r_2	N
Control	s_0	s_1	s_2	N
Total	n_0	n_1	n_2	$2N$

Table 4.1: Genotype distribution table when number of cases equals the number of controls.

$$\begin{aligned} T_A &= \frac{2N\{2N(r_1+2r_2)-N(n_1+2n_2)\}^2}{N(2N-N)\{2N(n_1+4n_2)-(n_1+2n_2)\}} \\ &= \frac{\{2(r_1+2r_2)-(n_1+2n_2)\}^2}{N(n_1+4n_2)-(n_1+2n_2)^2} \end{aligned} \tag{4.5}$$

Modifications of the Armitage test for recessive and dominant effects are given in the supplementary material of Sladek *et al.* (2007).

Figures 4.3, 4.4-4.9 show results of power simulations for the three common allele effect models, additive, dominant and recessive and two sample sizes, 250 and 1,000. The figures show that for all three scenarios and two different sample sizes the Bayesian test is more powerful for a broad range of values under the alternative hypothesis.

Figure 4.3 also demonstrates that optimal Bayesian protocol sets the prior distributions which are "realistic" in a sense that parameter values are distributed within the range where empirical power grows from 0 to 1. The data was generated from the logistic model for reasons explained below. It is an intriguing feature of our protocol that the "H_1 priors" are distributed quite tightly around the null value.

4.2 A decision-theoretic protocol for setting priors

As outlined above there is no subjectivity involved in setting our models. In this section we provide more details of the procedure for setting priors and discuss how it fits into the Bayesian decision theory.

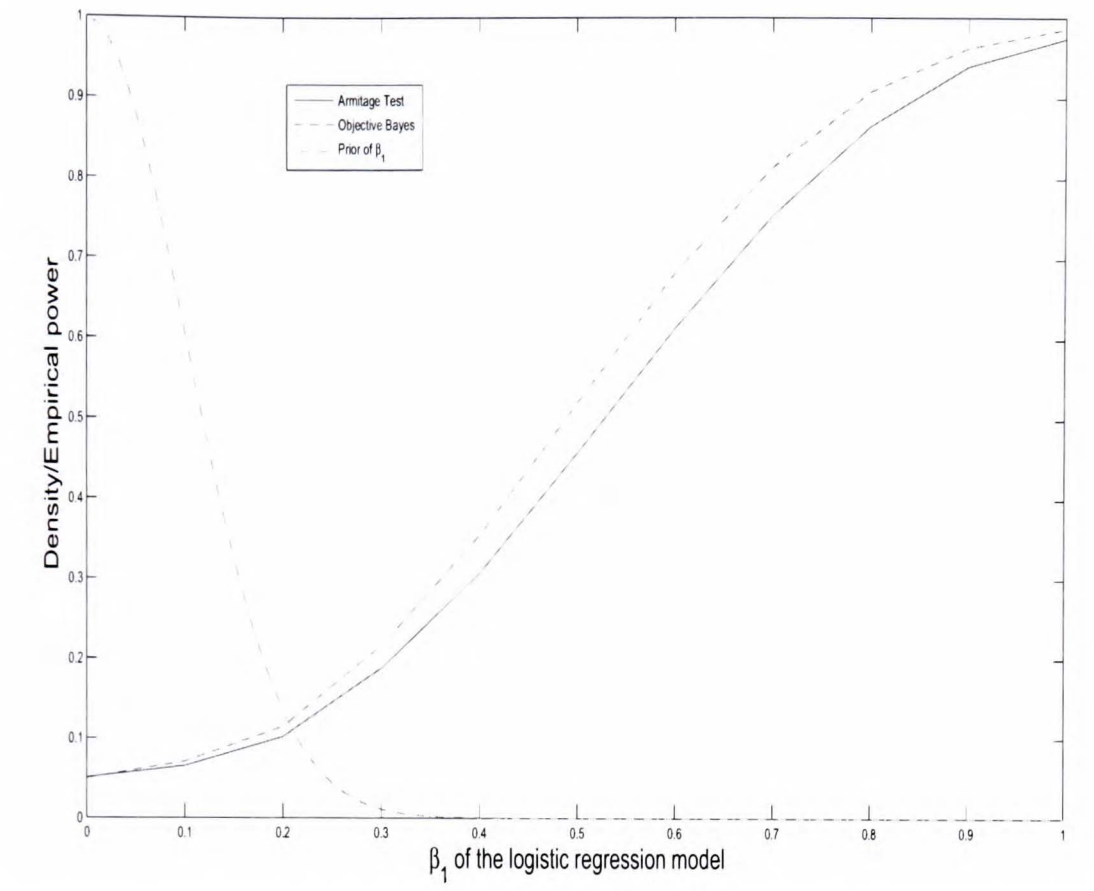


Figure 4.3: Prior distribution of β_1 under H_1 and power curves of Armitage test and optimal Bayesian test for logistic genetic effect, 250 cases and 250 controls. Prevalence 10%, MAF 20%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

The standard decision-theoretic formulation defines the *state of nature* θ from the set of possible values Θ, a decision rule $\delta(Y)$ that takes values in decision space A, and a loss function $L(\theta, \delta(Y))$. Then the risk function of the decision rule is:

$$R(\delta, \theta) = E\left(L(\delta(Y), \theta)\right),$$

where the expectation is taken over the distribution of Y. Frequentist decision theory seeks a decision rule that minimises the risk function. The Bayes risk further averages the risk function over possible values of θ using its prior distribution

$$r_\pi = \int_\Theta R(\delta, \theta) d\theta,$$

and the Bayes rule is the decision rule that minimises Bayes risk.

In our case the state of the nature is regression parameters β, the decision rule returns prior variance matrices $\Sigma^{(a)}$ (one for each hypothesis), and the loss function is one minus statistical power.

It is interesting to note that in the standard Bayesian decision theoretic approach, the optimal decision rule would provide a model that maximises power on average, including effect sizes that might not be of interest to the researcher. In our search for optimal decision rule (prior variance) we fixed state of nature β at a particular value. Namely, the one for which the Armitage test has power 0.5. This can be regarded as finding the Bayes rule when the prior on state of nature is a point mass function, because then the integral reduces to the value of the integrand at one point. Fortunately, the optimal decision rules displayed good properties for other states of nature β, i.e. provided a uniformly higher power. Alternatively, it can be seen as minimising the frequentist risk function for a known state of nature.

If one could solve analytically the minimisation problem for the Bayes or frequentist risk, this would be the best solution. We had to resort to a grid search for optimal values because of the complicated link between decision rule and the risk function in our setting. Under the null hypothesis of $\beta^0 = 0$ the distribution of our test statistics $T = BF(Y|\beta^0, \Sigma^{(a)})$ depends on the prior variance matrices $\Sigma^{(a)}$. This distribution has a cut-off value for rejecting the null at 0.05 Type I error, $t^*(\Sigma^{(a)})$, which is therefore a function of our decision rule. For a chosen $t^*(\Sigma^{(a)})$ and fixed state of nature under the alternative, $\beta^* > 0$, the expected loss is one minus the probability of the test statistic exceeding the cut-off value:

$$\begin{aligned} E(L(\Sigma^{(a)}, \beta^*)) &= 1 - P\left(BF(Y|\beta^*, \Sigma^{(a)}) > t^*(\Sigma^{(a)})\right) \\ &= 1 - \int_{t^*}^{\infty} f_T(t)dt \\ &= \int_{-\infty}^{t^*} f_T(t)dt \end{aligned}$$

For our likelihood and prior there is no analytic expression for $T = BF(Y|\beta^*, \Sigma^{(a)})$ and therefore for $E(L(\Sigma^{(a)}, \beta^*))$. Thus the problem is well defined within the decision-theoretic framework but can not be solved analytically.

N	MAF	β_1	Armitage	optimal Bayesian
500	0.5%	0.15	0.538	0.669
500	1%	0.090	0.470	0.567
250	5%	0.060	0.511	0.578
250	10%	0.042	0.493	0.535
250	15%	0.035	0.488	0.515
250	20%	0.032	0.445	0.466
250	25%	0.030	0.501	0.516
250	30%	0.029	0.512	0.524
250	35%	0.028	0.510	0.516

Table 4.2: Estimated power for different values of MAF. Data simulated from the additive model. Power and the null distribution estimated on 10,000 simulations, with disease prevalence of 10%.

4.3 Results

The 5% level threshold for rejecting the null hypothesis was set empirically by generating 100,000 datasets under the null hypothesis. For GWAS with many loci level of Type I error is often much lower but single SNP analysis is still a widely used approach (Balding (2006)). For example, Pharoah *et al.* (2007) conducted a large case-control association study in breast cancer using 710 SNPs in 120 candidate genes, and for each SNP they performed a test at a nominal 5% level. Throughout our simulations we kept the number of cases equal to the number of controls. We varied the minor allele frequency from 0.5% to 35%. We considered case-control studies of size 250, 500 and 1,000 and disease prevalence of 10%. For the sample size of 250 the protocol set the parameters as follows: $v_0^{(0)} = 10, v_1^{(0)} = 0.0001, v_0^{(1)} = 10, v_1^{(1)} = 0.01$. For the sample size of 1,000 all prior variance had the same values except a much smaller $v_1^{(1)} = 0.0014$, which has an intuitive explanation that for a larger sample size we expect to identify smaller genetic effects. Since the protocol goes through a large set of values and the time constraints of the project, we used the optimal prior found for $N = 250$ for datasets with $N = 500$, and optimal prior for $N = 1000$ for samples $N = 2000$. These optimal values were used for all subsequent simulations with different settings. Table 4.2 provides simulation results for different values of MAF.

More detailed results are in Figures 4.4-4.10. Note that since we did not optimise prior for samples of 2,000 cases the gain in power dicreased. This is not surprising given that for a fixed prior increasing sample size dilutes the influence of the prior.

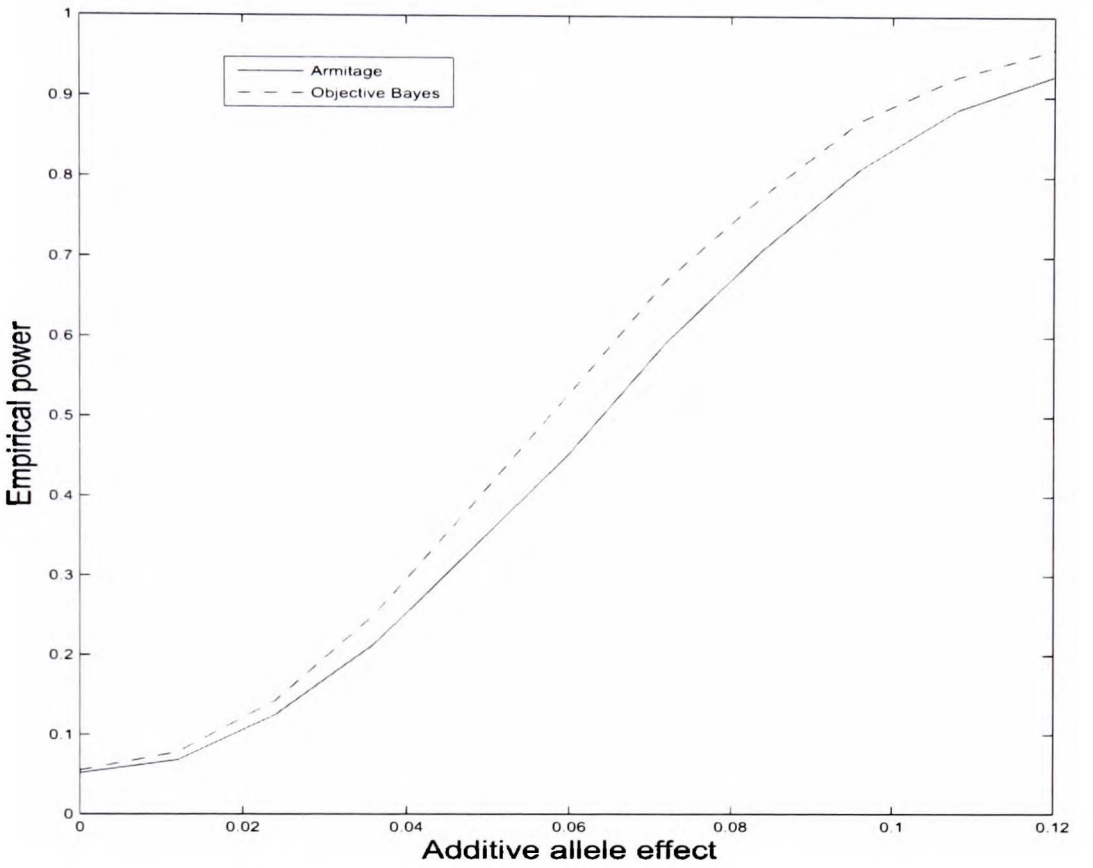


Figure 4.4: Empirical power of the additive Armitage and Objective Bayes tests for data simulated under additive allele effect, 1,000 cases and 1,000 controls. Prevalence 10%, MAF 1%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

4.4 Application to a colorectal cancer association study

In this section we apply the method presented above to testing for association in a study in colorectal cancer (CRC).

CRC is a major public health issue: it is the second to the fourth most common cancer in industrialised countries. In the USA, 145,000 people are diagnosed with CRC annually, and the global figure is around 875,000 (Potter (1999)). In the UK, the lifetime probability that a person will develop CRC is 6% and it contributes to 16,000 deaths per year (Quarini and Gosney (2009)).

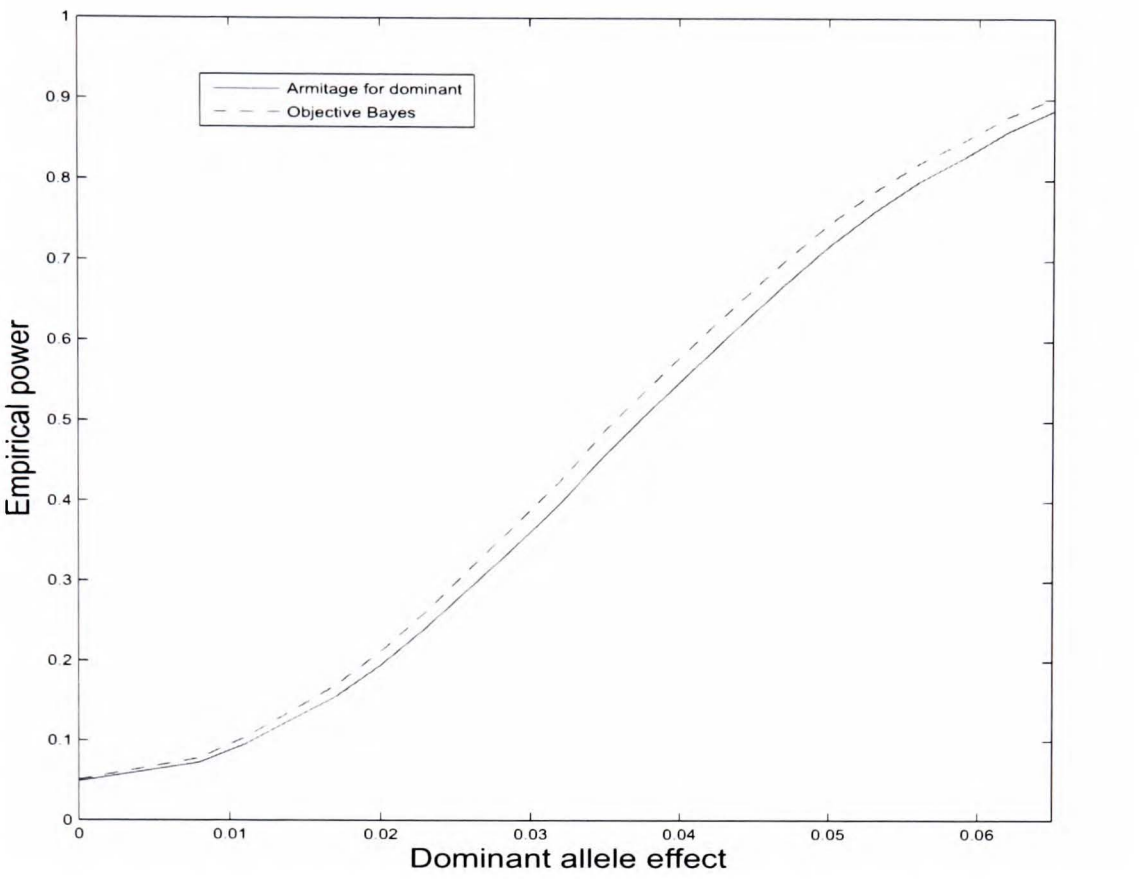


Figure 4.5: Empirical power of the dominant Armitage and Objective Bayes tests for dominant effect, 250 cases and 250 controls. Prevalence 10%, MAF 20%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

Genetics is thought to play a key role in the initiation and progression of CRC, and research into the genetic basis of CRC is considered to be leading the way in understanding cancer genetics (Chapelle (2004)). One of the reasons that CRC provides a useful model for other cancers is the fact that the various stages of tumour development can be relatively easily observed.

Considerable genetic basis of CRC is indicated by both case-control and twin studies. Lichtenstein *et al.* (2000) estimated that inherited susceptibility is responsible for $\sim 30\%$ of all CRC. Several high-*penetrance* genes (penetrance is defined as the probability that a particular phenotype is observed in individuals with a given genotype) which contribute to hereditary colorectal cancer have been identified (Aaltonen *et al.* (2007)). However, it is estimated that mutations in those genes together account for less than 5% of cases of CRC (Aaltonen *et al.* (2007)). Much of the remaining variation in genetic risk is attributed to more common, lower-penetrance

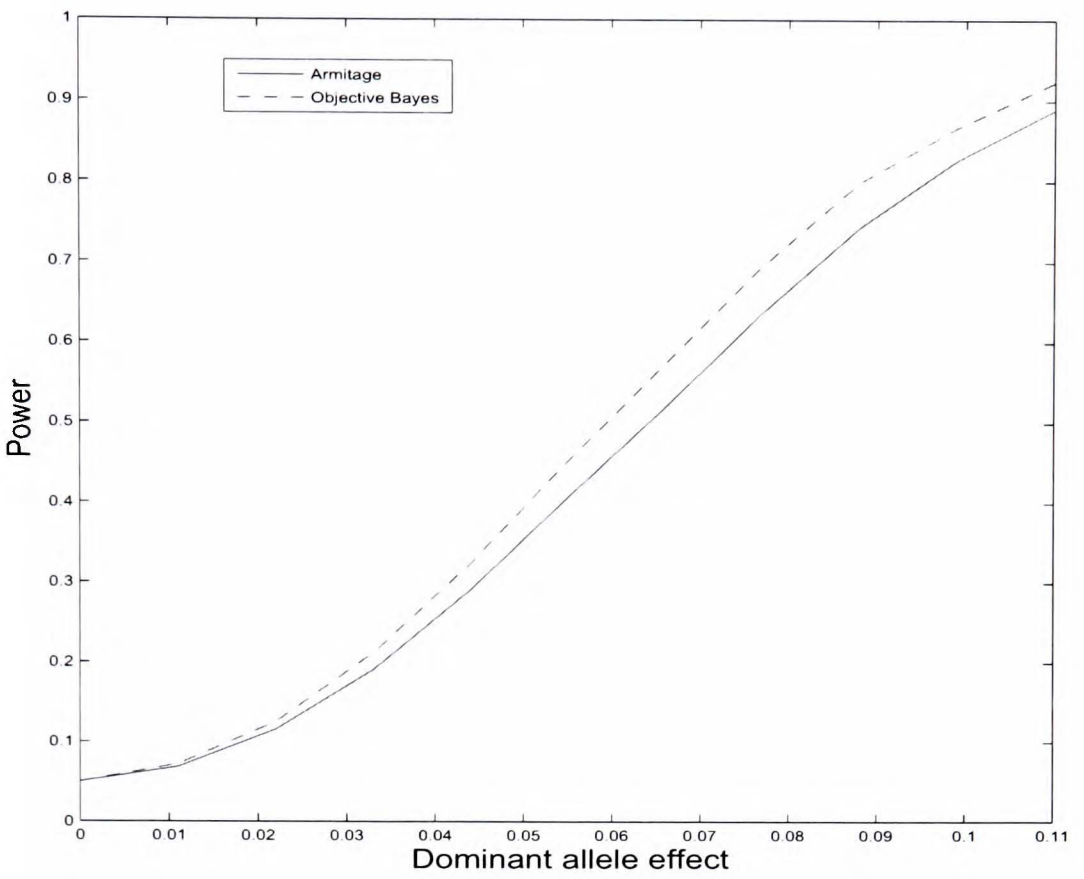


Figure 4.6: Empirical power of the dominant Armitage and Objective Bayes tests for dominant effect, 1,000 cases and 1,000 controls. Prevalence 10%, MAF 1%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

variants, and there is extensive ongoing research in identifying such alleles.

Here we applied our method to one of the datasets from the ongoing search for low-penetrance alleles associated with CRC. The dataset is a case-control sample from ColoRectal tumour Gene Identification (CoRGI) Study Consortium which has been collecting samples and data from families affected by CRC since 1997 (Kemp *et al.* (2006)). Samples from the consortium have been used to successfully identify several variants associated with CRC, e.g. reported in Broderic *et al.* (2007), Tomlinson *et al.* (2007), Tomlinson *et al.* (2008).

We used the case-control dataset for the SNP rs6983267 which was analysed in Tomlinson *et al.* (2007). This sample was used in the first phase of the study and contained 930 cases (620 CRC and 310 high-risk colorectal adenoma) and 960 controls. The most powerful test depends on the mode of inheritance but it is unknown at this locus, so Tomlinson *et al.* (2007) used the χ^2 test with 1 degree of freedom.

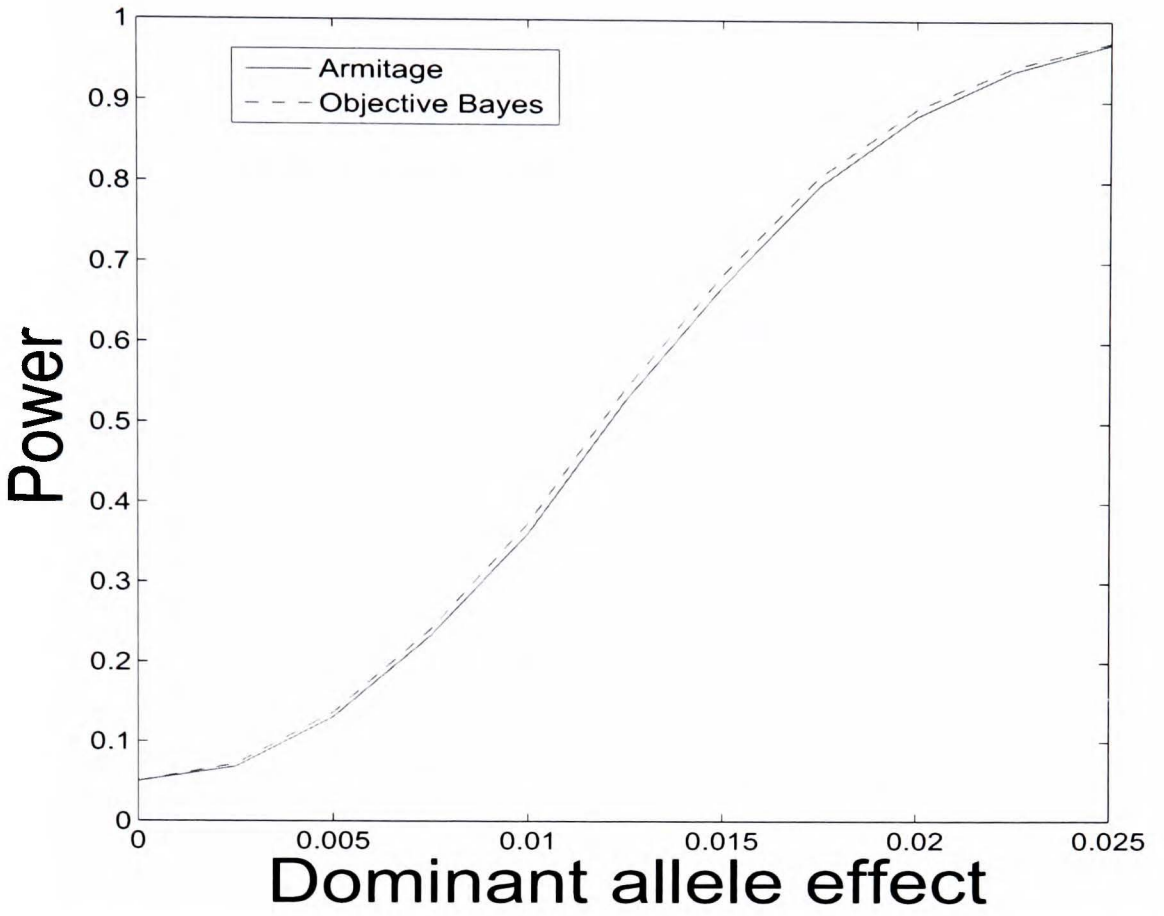


Figure 4.7: Empirical power of the dominant Armitage and Objective Bayes tests for dominant effect, 2,000 cases and 2,000 controls. Prevalence 10%, MAF 20%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

The same SNP was reported in Zanke *et al.* (2007) who used the maximum of dominant, recessive and additive statistics for the same reason of unknown inheritance. We resorted to the same technique and ran 10,000 simulations to get the null distribution of the maximum of the three proposed Bayesian models, dominant, recessive and additive. We rejected the null hypothesis of no association for this SNP based on the simulated null distribution, at the 5% type I error rate. Tomlinson *et al.* (2007) and Zanke *et al.* (2007) also reported this SNP as associated with CRC.

Application to this real dataset highlighted one area for further development and improvement of the method. Unlike some of the commonly used tests, the Bayesian model presented above requires separate optimisation of the parameters for each sample size. If a theoretical null distribution was known the separate optimisation would not be needed. Knowing a theoretical null distribution would also alleviate the computational burden when one needs a lower type I error rate: for the presented

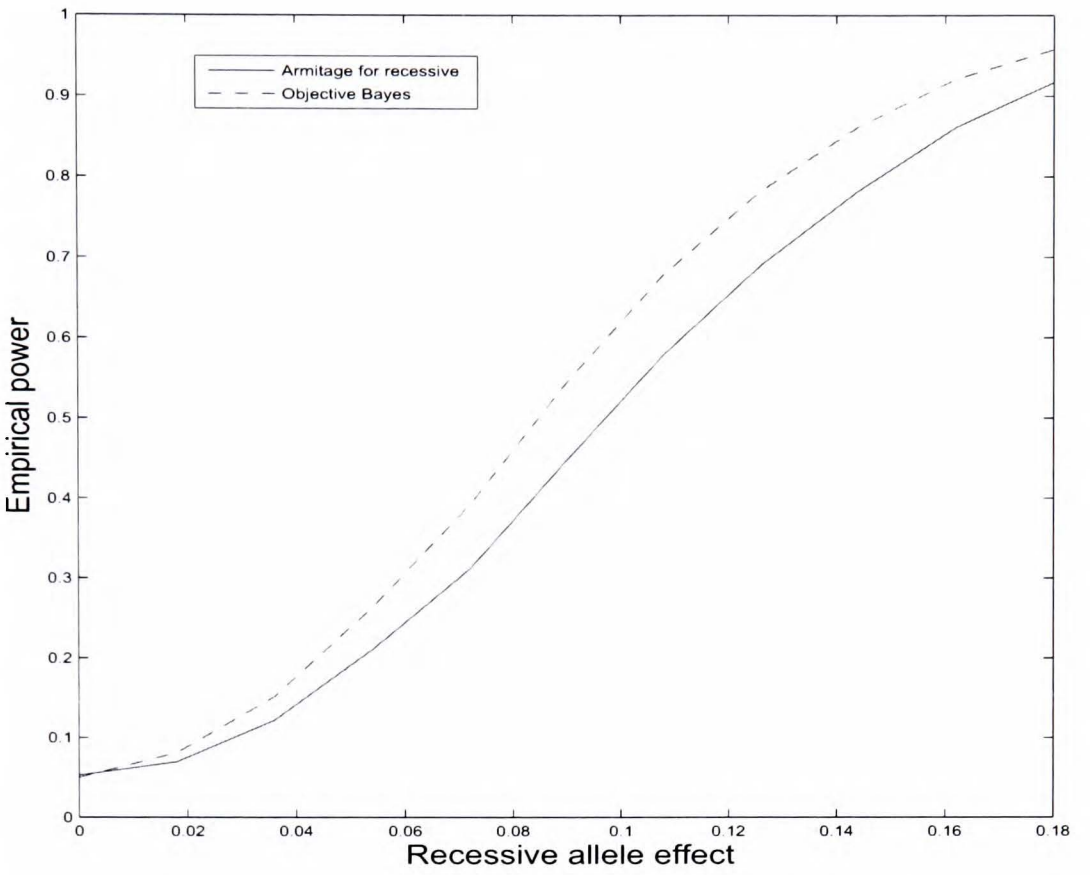


Figure 4.8: Empirical power of the recessive Armitage and optimal Bayesian tests, for 250 cases and 250 controls. Prevalence 10%, MAF 20%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

model testing at a lower nominal level would require many more simulations to get a more precise null distribution in the tail.

4.5 Comparison to similar models

One of the most prominent recent GWAS studies of case-control data using Bayesian approaches was WTCCC (2007). The detailed description of the models used in the analysis of all seven GWAS carried out as part of WTCCC is provided in Marchini *et al.* (2007), where it is stated that authors searched for models with good ability to predict the causal variants rather than focusing on difference in power between various tests. Nevertheless, given the prominence of the study and the fact that the models were used for detecting possible causal alleles (even if only for selection for further investigation) it is of interest to compare power properties of our model

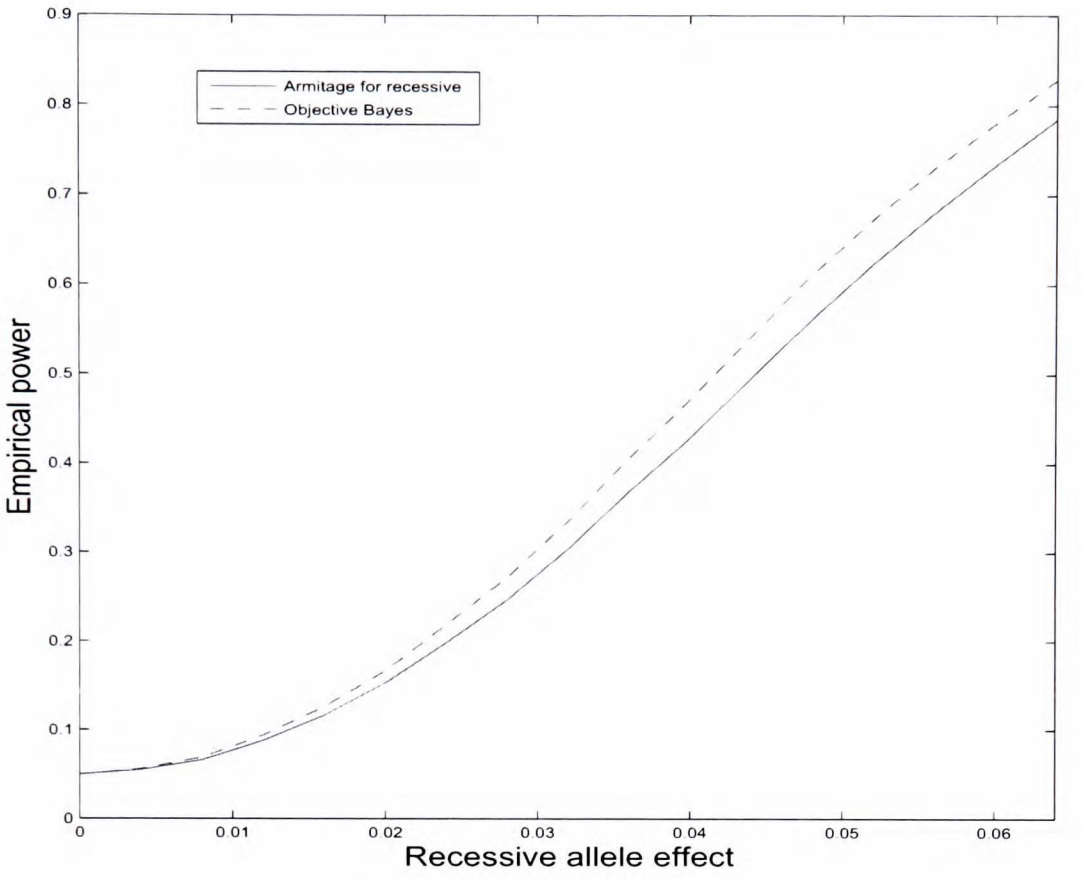


Figure 4.9: Empirical power of the recessive Armitage and optimal Bayesian tests, for 1,000 cases and 1,000 controls. Prevalence 10%, MAF 20%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

with the models used in that large project. Also, the priors for the Bayes factors in Marchini *et al.* (2007) were carefully selected using prior knowledge of possible effect sizes.

We carried out a study under the conditions specified in Marchini *et al.* (2007): simulated datasets of 2,000 cases, with $MAF = 1\%$ (though they analysed more common alleles as well). WTCCC data consisted of 1,924 cases and 2,938 controls, but our experimentation (results not shown here) with increasing the number of controls relative to the number of cases showed that power changes very slowly when the number of controls grows. This effect is probably due to the relatively small amount of information contained in controls. Thus here we compare the models on datasets with equal number of cases and controls, 2000. WTCCC model for H_1 used prior with standard deviation equal 0.2 for the genetic effect parameter $\beta_1 \sim N(0, 0.2)$. Their null model had only one parameter, the intercept. To reproduce

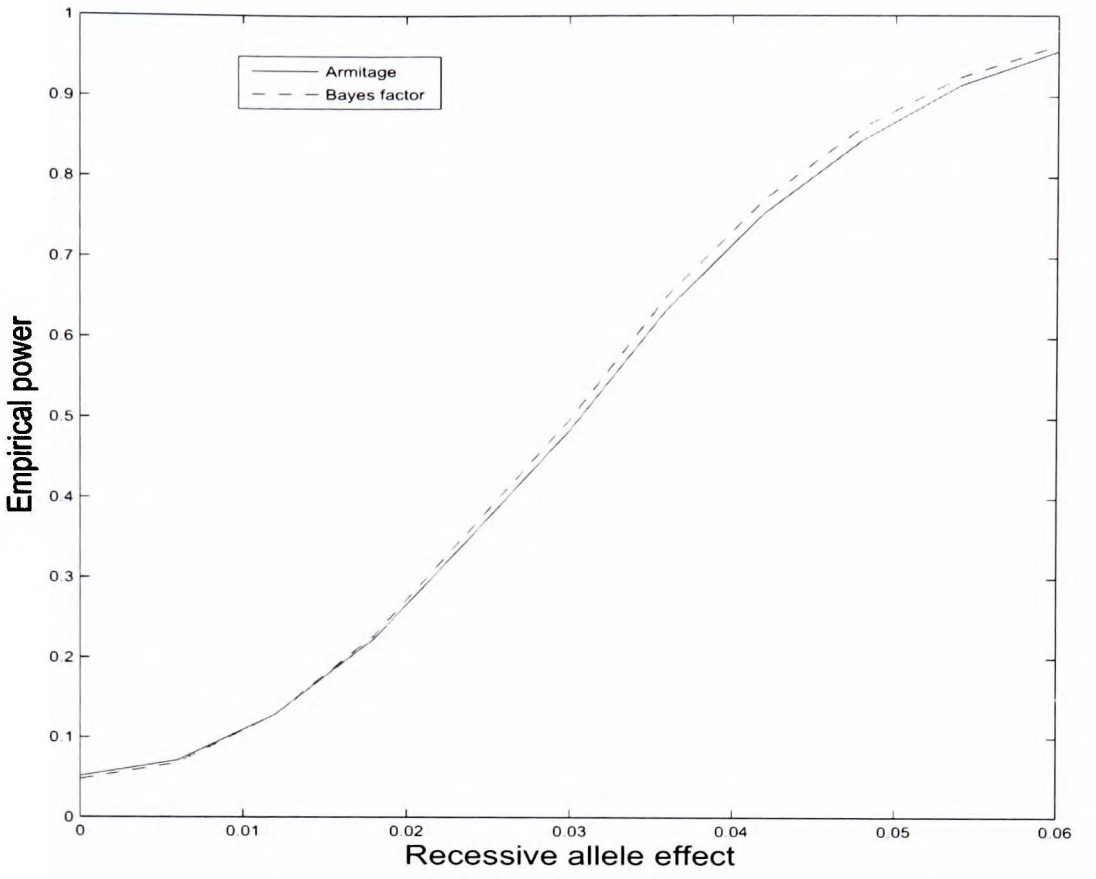


Figure 4.10: Empirical power of the recessive Armitage and optimal Bayesian tests, for 2,000 cases and 2,000 controls. Prevalence 10%, MAF 20%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

the null model we set a very small variance of 0.00001 on β_1 for our H_0 model in (4.2).

We also investigated, as in the previous chapter, the relative contributions of optimal parameters and parameter constraints. Figure 4.12 illustrates that similarly to the previous models, optimisation of priors is more important.

4.6 Discussion

Many of the arguments from the previous chapter explaining the reasons why the proposed models have a power advantage apply here as well: we utilised information from both cases and controls, identified and took into account the correct parameter space. Also, all other reasons for using Bayesian framework outlined in that chapter could be reiterated. The important result of this chapter is the proposal of models

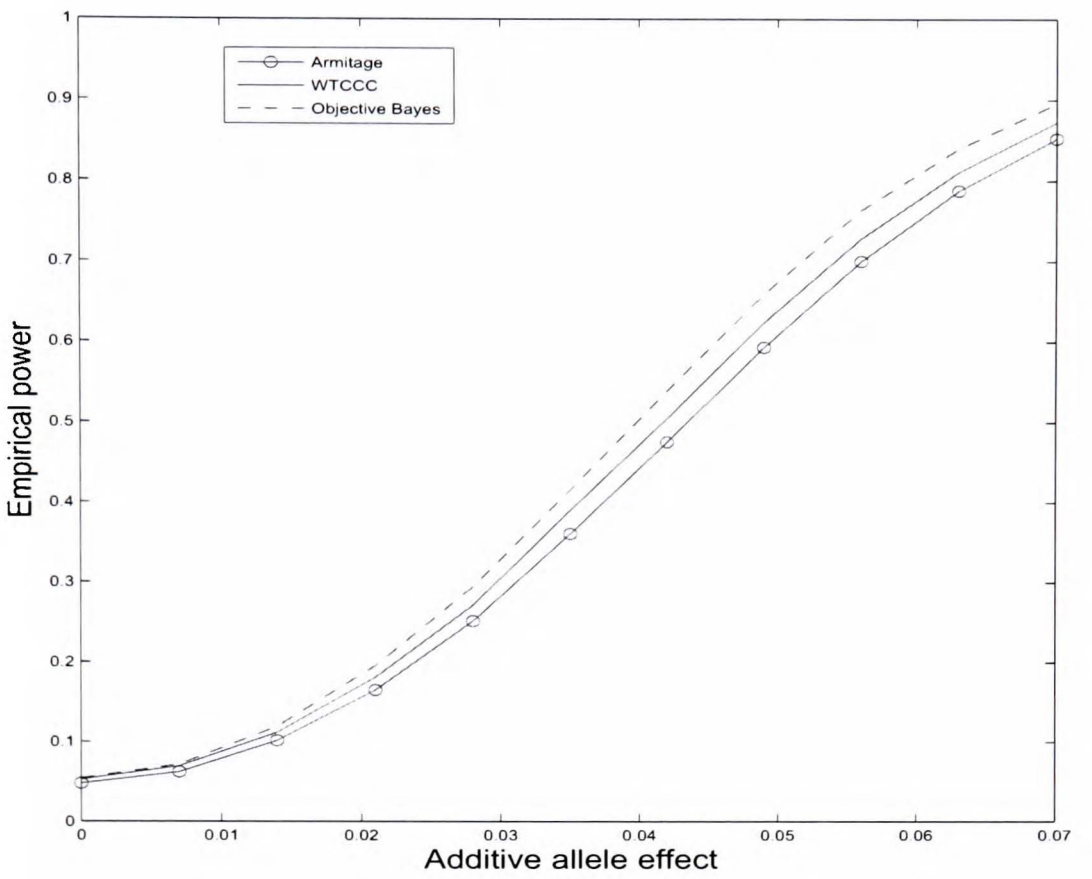


Figure 4.11: Power curves of WTCCC Bayes factor, Objective Bayes and the Armitage tests for data simulated under additive allele model, with 2,000 cases and 2,000 controls, MAF = 1%, prevalence 10%. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

that are more powerful *for all three* genetic effect models, computationally efficient and allow incorporating additional parameters for the GWAS with unknown MAFs. Such studies are at the focus of the current developments in quantitative genetics and we believe that the proposed model can enhance advances in localising disease-predisposing genes.

Sometimes loci are ranked in GWAs based on their Bayes factor values. Bayes factors can be approximated by the method suggested in Wakefield (2007), where the approximation has the simple form:

$$ABF = \frac{1}{\sqrt{1-r}} exp\left[-\frac{Z^2}{2}r\right], \qquad r = W/(W+V), \tag{4.6}$$

where W is prior variance and V is variance of the MLE estimate. It might seem logical to conclude from the expression that the ranking of loci does not depend

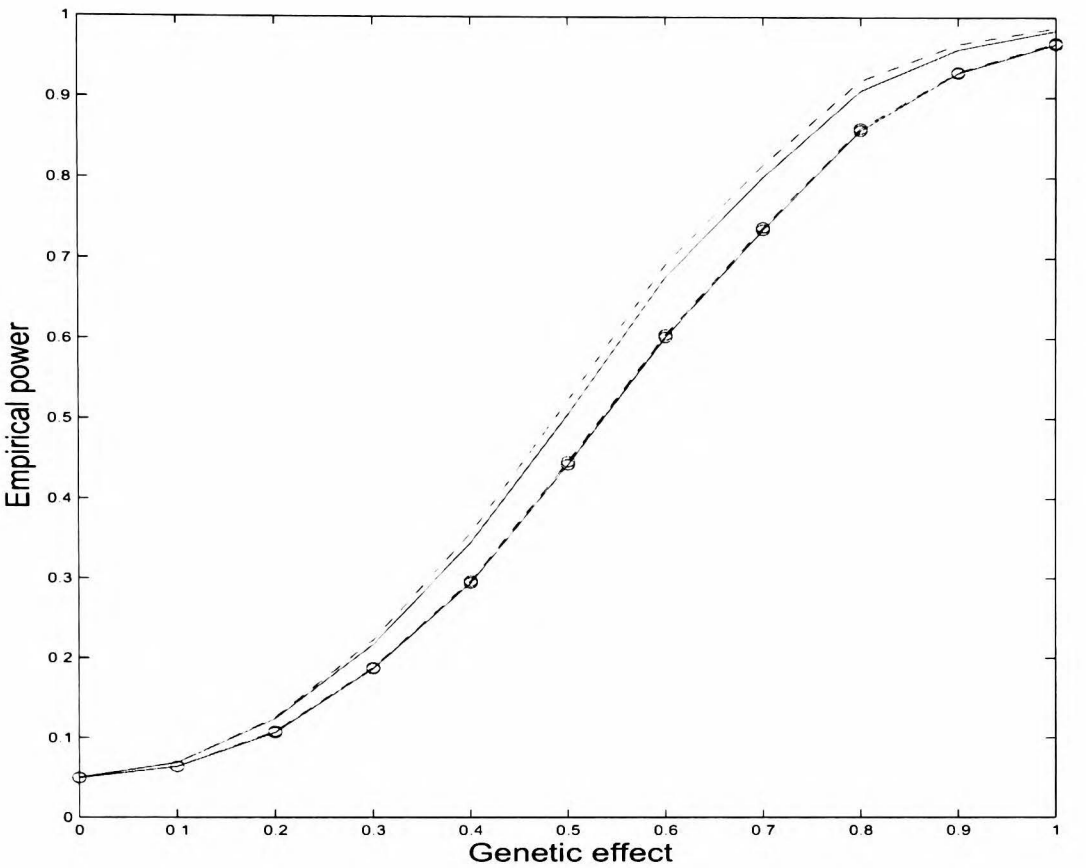


Figure 4.12: Comparison of statistical power for optimal $(\widehat{v_1^{(0)}} = 0.0001, \widehat{v_1^{(1)}} = 0.01)$ versus unoptimised priors $(v_1^{(0)} = 0.01, v_1^{(1)} = 1)$; and incomplete integration versus full integration. Dashed lines represents improvement due to incomplete integration. The lines with "o" are for the unoptimised model. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

on the choice or prior, since for two different values of Z (for two different loci), regardless of choice of shrinkage ratio r, one approximate Bayes factor will be larger than the other (Figure 4.13). Therefore we might conclude that the choice of prior does not change ranking. However, the shrinkage ratio depends on the marker values (design matrix X) at each locus and can be different even for two genes with the same MAFs of their alleles. For illustration we plotted distribution of the shrinkage ratio for a fixed MAF in Figure 4.14.

This variability implies that both Z and r in (4.6) can differ between loci with equal MAFs. Therefore it does not easily follow from the approximation that ranking is invariant. Furthermore, it is possible to construct a counterexample where ranking changes if we vary the prior. Ranking invariance could have contradicted our findings

where power depends heavily on the prior.

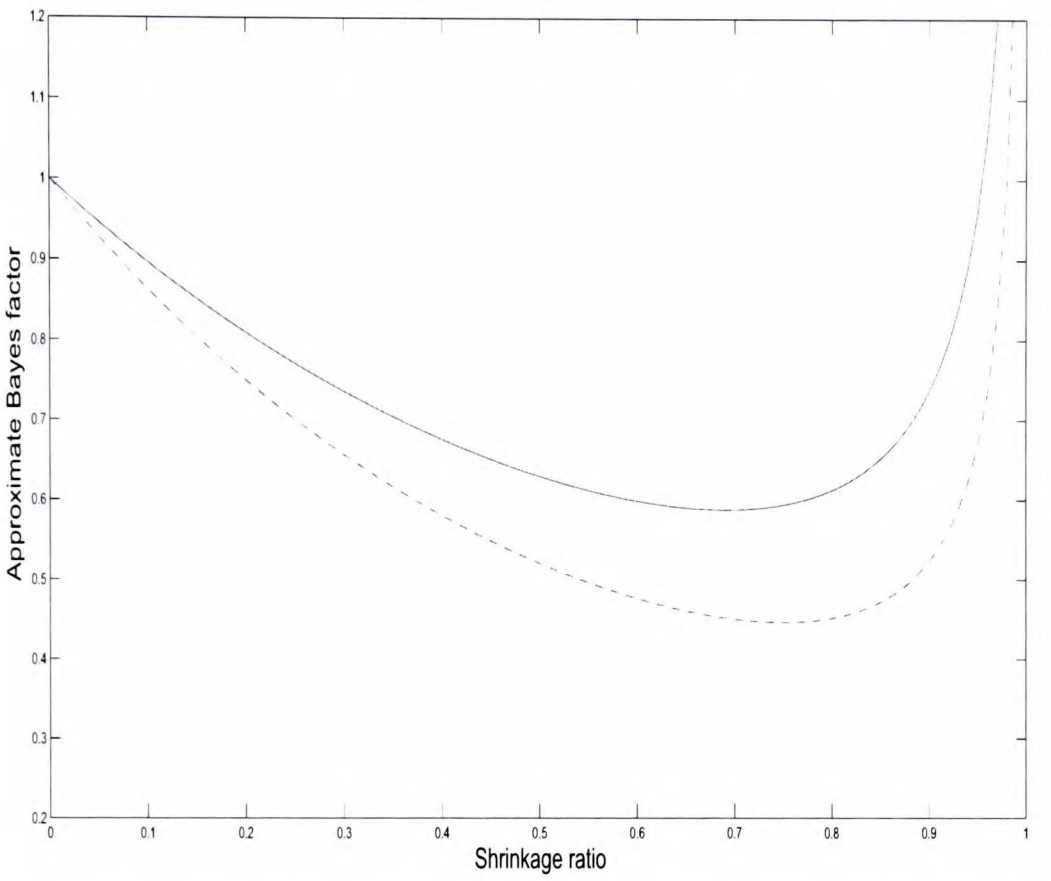


Figure 4.13: Dependency of the approximate BF on the shrinkage ratio for two loci. For one loci (solid line) Z score is 1.8, and 2 for the other (dashed line). The lines do not intersect but V and therefore $r(V, W)$ are different for the two loci.

Possible further investigations could include searching for a test statistics that would have good power and robustness characteristics under all three models, additive, recessive and dominant. Such a test would be very much welcomed by the scientific community given that there is often no prior knowledge of the type of the allele effect. Another interesting direction of research is using models for the case-control design with unknown MAFs via retrospective likelihood. Since sampling is carried out based on the disease status, it might be advantageous to design models that directly capture this scheme rather than rely on the finding that parameter estimates from prospective and retrospective likelihoods are asymptotically equivalent.

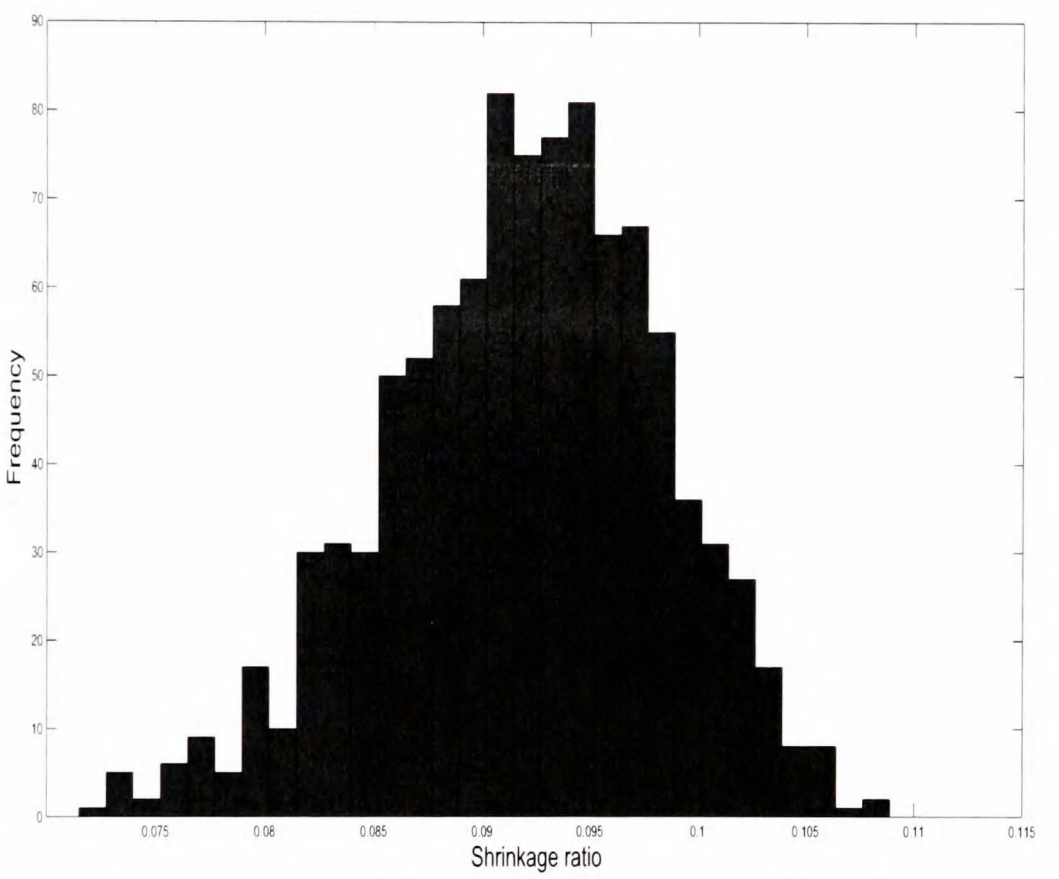


Figure 4.14: Distribution of the shrinkage factor. From one thousand simulations of samples of 250 cases and 250 controls from a population with $MAF = 0.05$, prevalence 10% and no genetic effect, $\beta_1 = 0$, prior variance $W = 0.03$.

Appendices

Appendix A

Notation

We have tried to use standard notation throughout this thesis. Random variables have been denoted with uppercase letters and the realised values of the variable denoted by the corresponding lowercase letter. Estimated parameters have been usually denoted by β (for regression models) and θ. A few well-known distributions and statistical concepts have been mentioned using the following notation:

$N(\mu, \sigma^2)$	Normal distribution with the expectation μ and variance σ^2
$MN(m, \alpha_1, ..., \alpha_p)$	Multinomial distribution with m trials and cell probabilities $\alpha_1, ..., \alpha_p$
$Bin(n, p)$	Binomial distribution, where p is the probability of "sucess" and n is number of trials
$L(.)$	The likelihood function
$l(.)$	The log-likelihood function

Appendix B

Computer programmes

This appendix lists the key functions used to produce the results presented in this thesis. All the programmes were implemented in Matlab computing language and environment, version 7.4.0.287 (R2007a).

For Chapter 2 we list the more generic version of several functions used in the chapter, the algorithm for finding maximum-a-posteriori (for binary response, Section 2.4.2):

```
function result = EMRunBinMAP( its, initBeta, initProbs, combinations, y,
prior_mu, prior )
% The function returns maximum a posteriori of betas;
% There is no stopping rule at the moment, so need to choose appropriate
% number of iterations 'its'.
% initBeta - 1xd vector of starting values
% initProbs - NxS, where S - number of possible combinations
% combinations - all the possible values vector x_j can take, without intercept
% y - binary response
% prior_mu - prior precision of the intercept
% prior - prior precision of other betas
betaDim = size(initBeta, 2);
ysize = size(y,1);
```

```
comSize = size(combinations,1);
% Add intercept
combinations = [ones( comSize, 1 ), combinations];
xFull = repmat( combinations, ysize, 1 );
yFull = y(1) * ones(comSize,1);
for i= 2:ysize
yFull = vertcat( yFull,y(i)*ones(comSize,1) );
end
betas = zeros(its,betaDim);
betas(1,:) = initBeta;
% Joint loop
for n=2:its
currentBetas = betas(n-1,:)';
% ====== E-step =====
wVector=zeros(0,1);
for i=1:1:ysize
denom = sum( exp( y(i)*combinations*currentBetas-...
log( 1.+exp(combinations*currentBetas) ) ).*initProbs(i,:)' );
wVector = [wVector; exp( y(i)*combinations*currentBetas-...
log( 1.+exp(combinations*currentBetas) ) ).*initProbs(i,:)'/denom];
end
% ====== M-step ============
[loglik, b, Var_H0, Prec, step]= irls_logistic_mu( xFull, yFull, prior_mu, prior, ...
wVector );
betas(n,:)= b';
end
result = betas;
% End of function EMRunBinMAP
```

For the models of Chapter 3, the core function is the computation of the incomplete integral of (3.3). The first part is calculated as follows (including the constants not shown in the formula):

```
function res = BF_incr( N, r1, r0, a0, b0, a1, b1, p0 )
% parameters of the function correspond to the notation of Section 3.2
res = ...
log( betainc( p0, r0 + a0, N - r0 + b0 ) )...
- gammaln(N + a0 + b0) + gammaln(r0 + a0) + gammaln(N - r0 + b0) + ...
log( ( 1 - betainc( p0, r1 + a1, N - r1 + b1 ) ) )...
- gammaln( N + a1 + b1 ) + gammaln( r1 + a1 ) + gammaln( N - r1 + b1 ) ...
-( r1 + r0 + a1 + a0 - 1 )*log( p0 ) ...
-( 2*N - r1 - r0 + b1 + b0 - 1 )*log( 1 - p0 );
% End of function BF_incr
```

In Chapter 4, after the maximum-a-posteriori is calculated via the IRLS algorithm, the marginal is approximated using the following function:

```
function res = BayesTest(X, Y, B_H1, Var_H1, prec10, prec11, Var_H0, B_H0,...
prec00, prec01)
% X - genotypes
% Y - phenotypes
% B_H1 - MAP under the alternative hypothesis
% Var_H1 - Hessian of the approximation under the alternative
% B_H0 - MAP under the null hypothesis
% Var_H0 - Hessian of the approximation under the null
% prec10, prec11 - precision (1/var) for beta0 and beta1 under the
% alternative
% prec00, prec01 - precision (1/var) for beta0 and beta1 under the
% null
```

```
d = size( X, 2);
res = ...
log( (2*pi)^(d/2)*sqrt( det( Var_H1 ) ) ) + loglik( Y, X, B_H1 ) + ...
log( mvnpdf( B_H1', [0 0], diag([1/prec10, 1/prec11]) ) ) - ...
( log( (2*pi)^(d/2)*sqrt( det(Var_H0) ) ) + loglik( Y, X, B_H0 ) + ...
log( mvnpdf( B_H0', [0 0], diag( [1/prec00, 1/prec01] ) ) ) );
% End of function BayesTest
```

Appendix C

Detailed results for Chapter 3

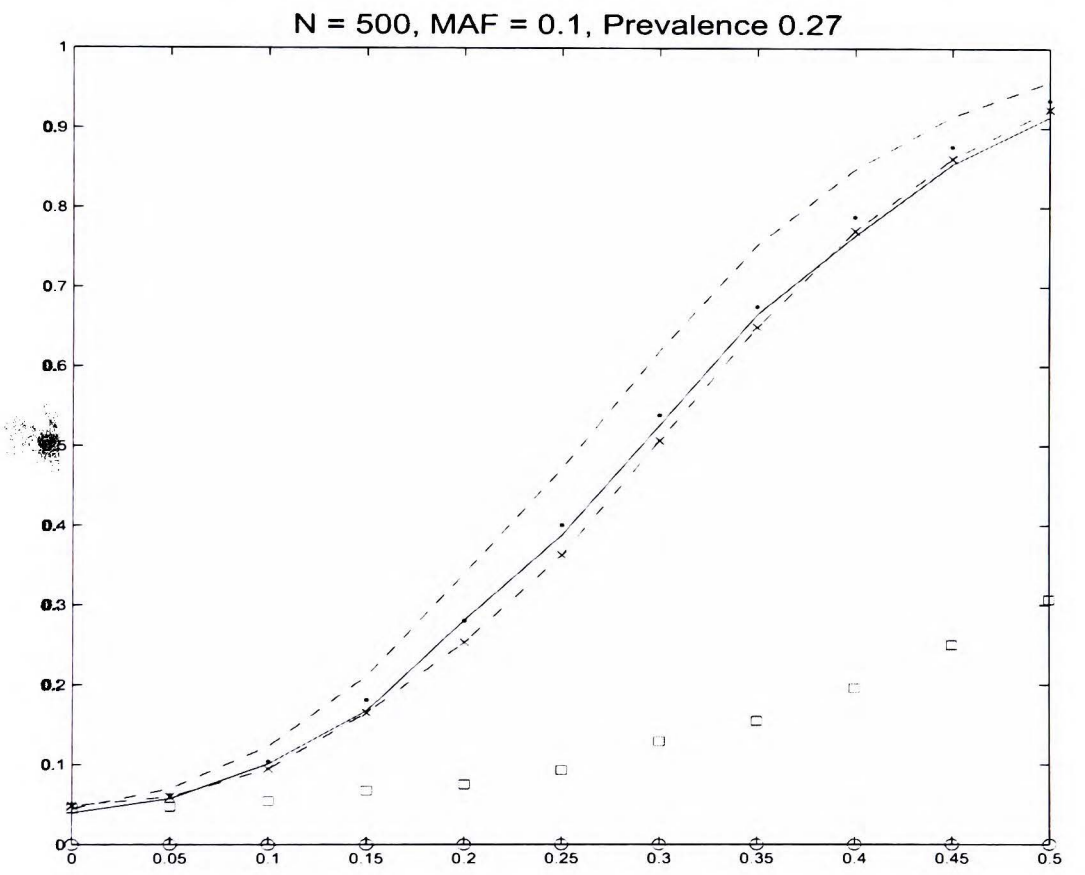


Figure C.1: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

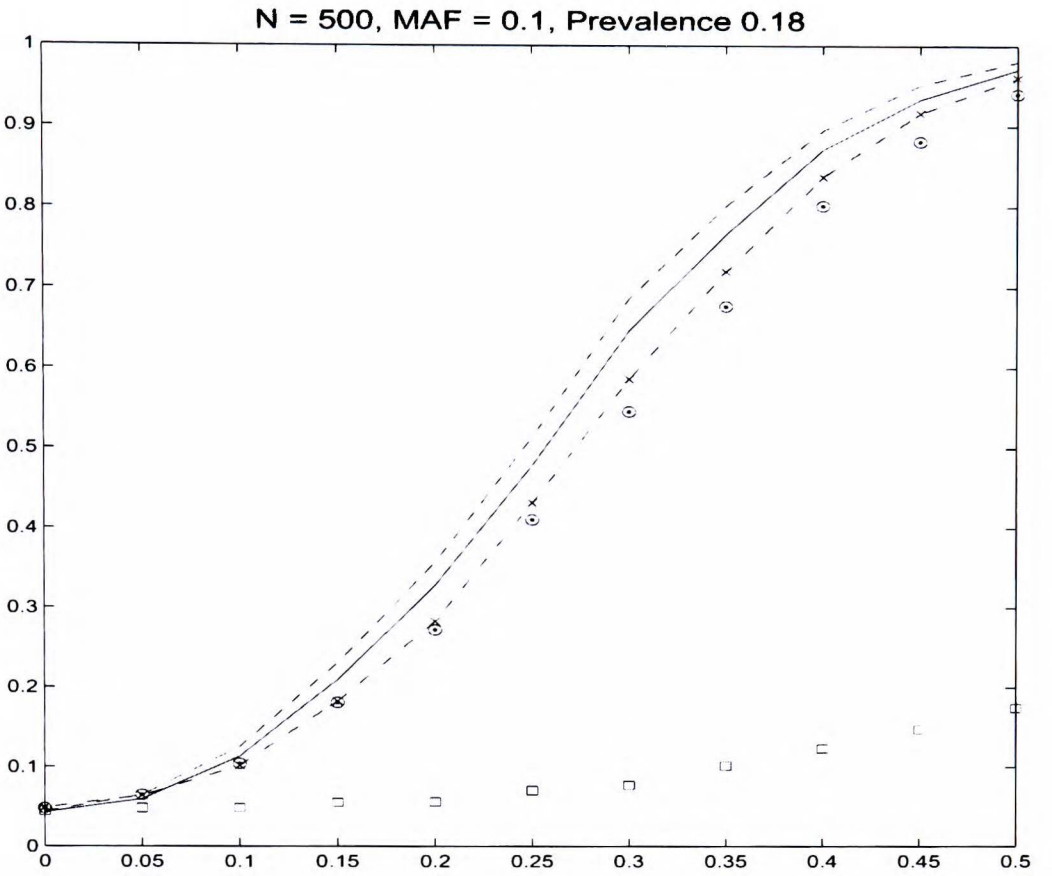


Figure C.2: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

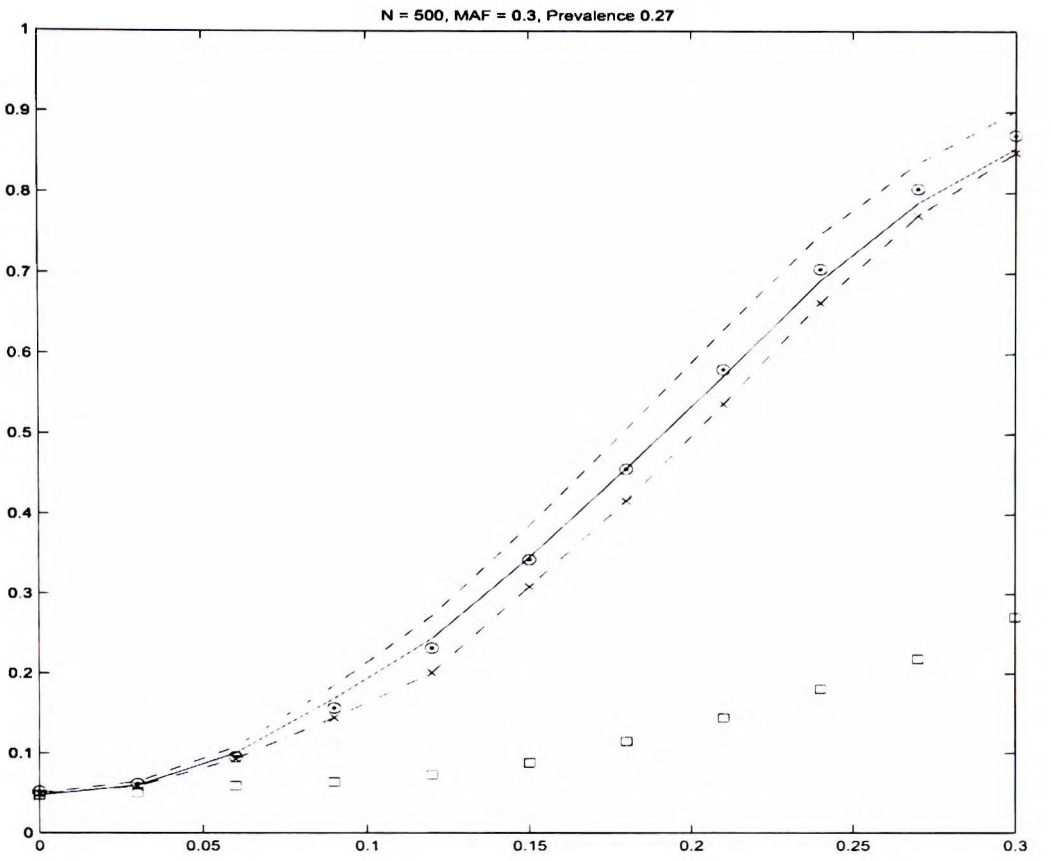


Figure C.3: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

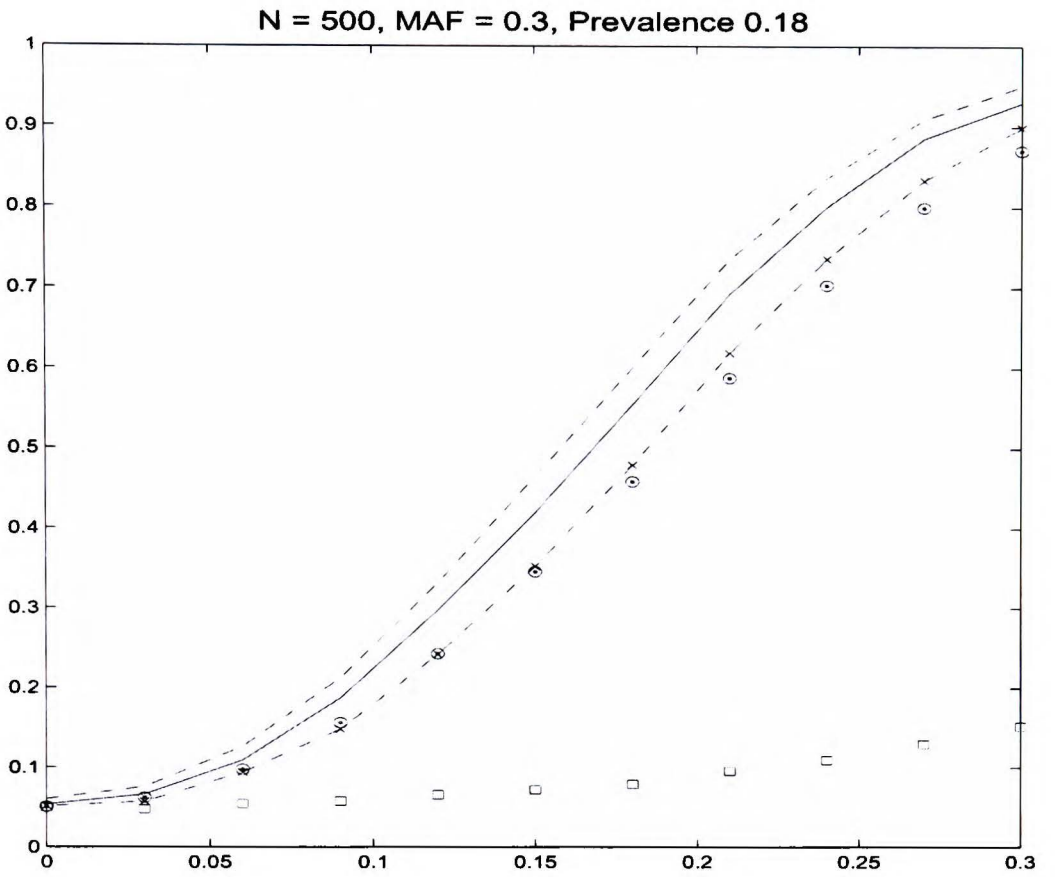


Figure C.4: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

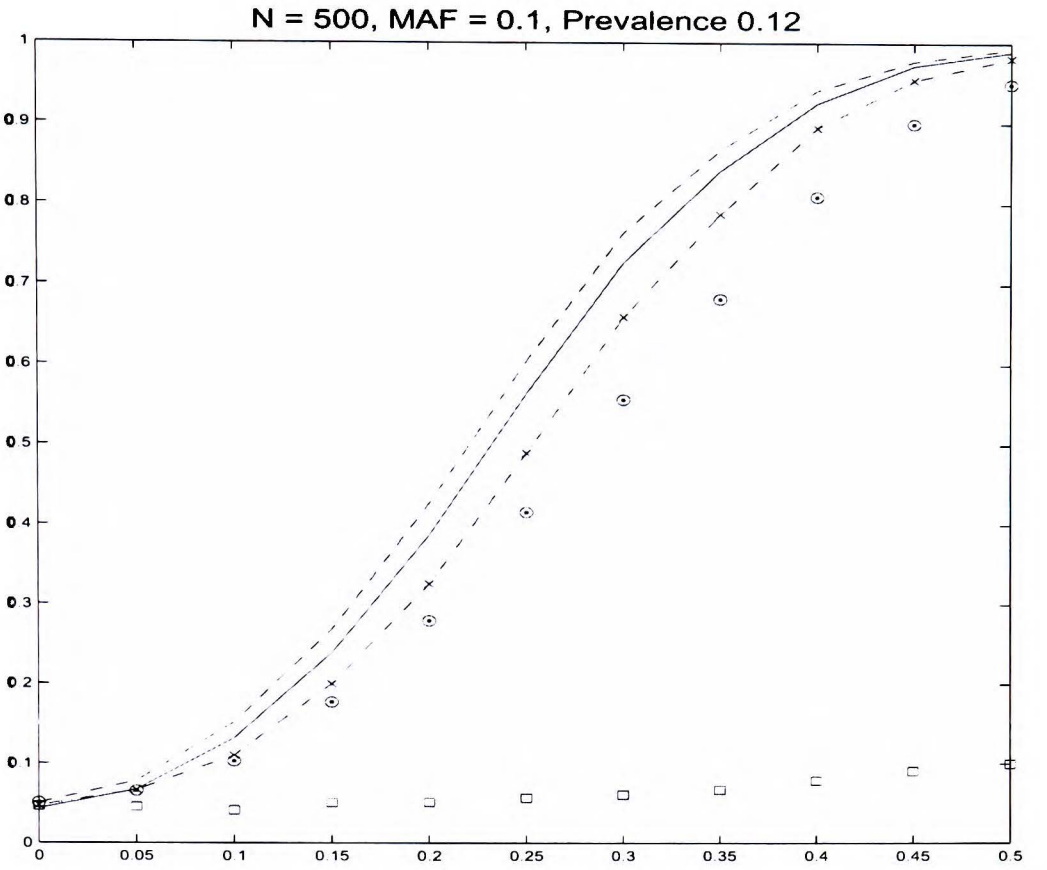


Figure C.5: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

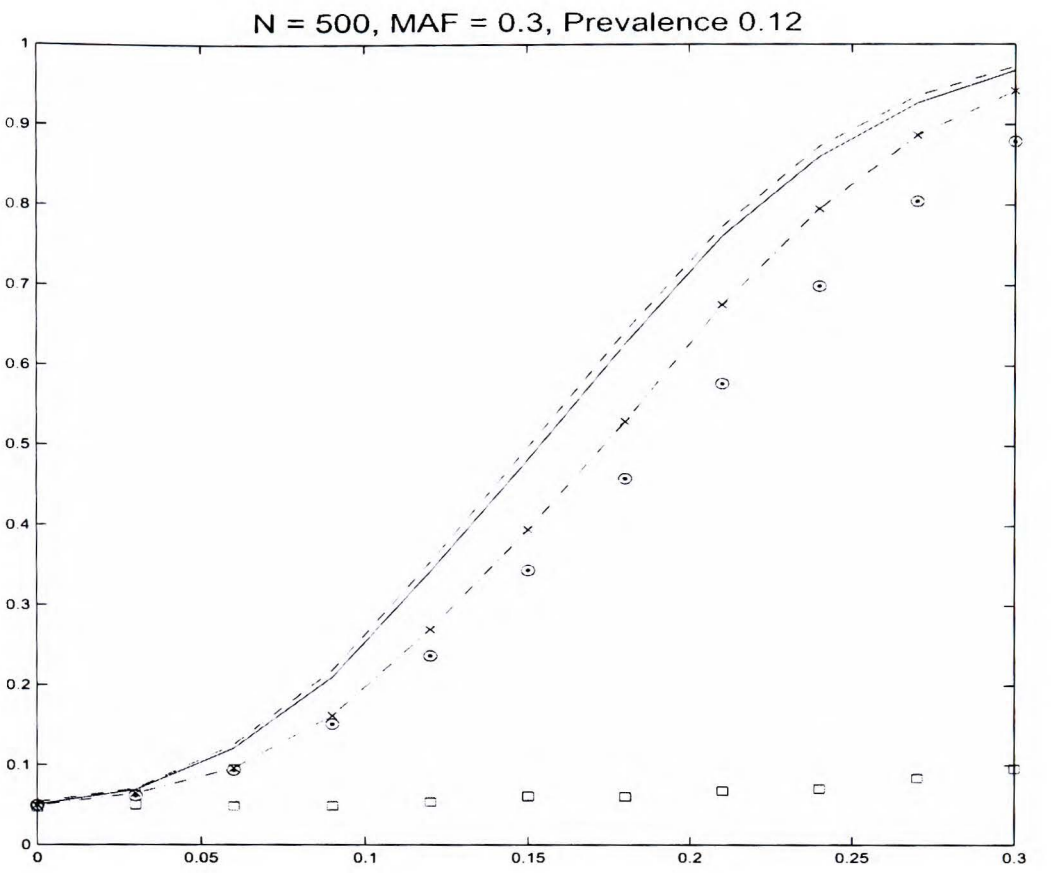


Figure C.6: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

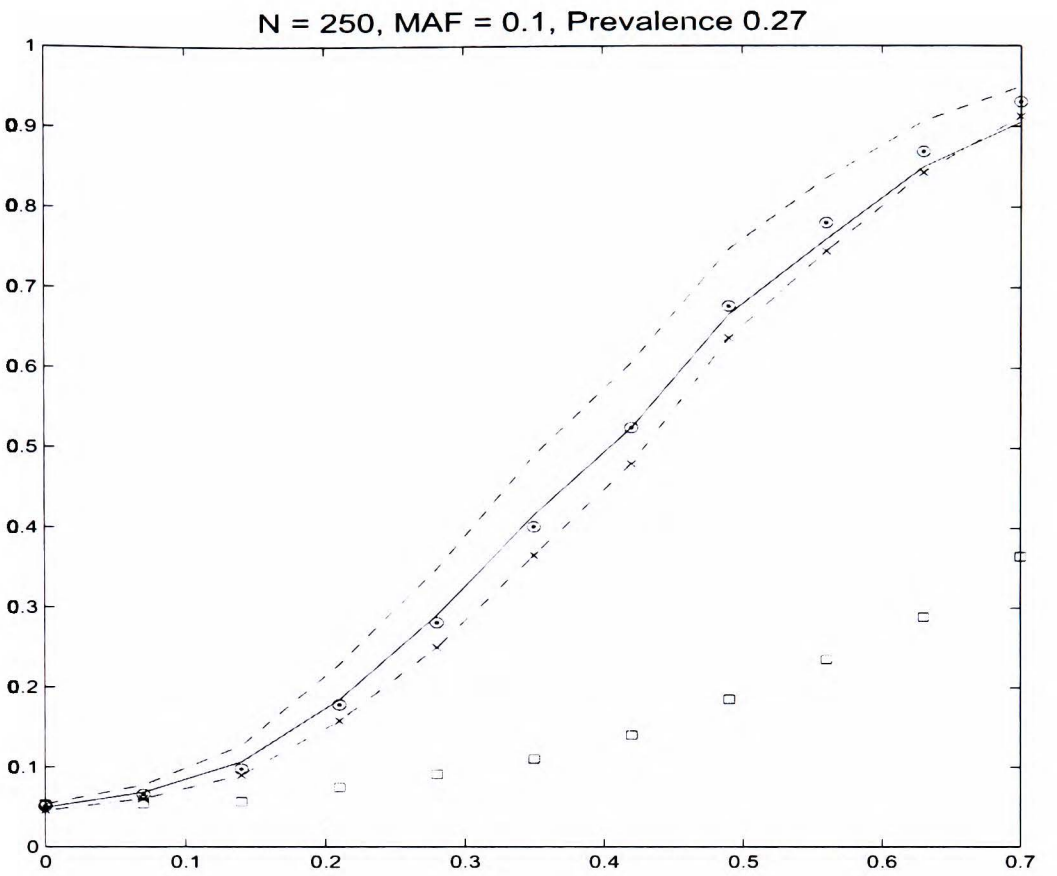


Figure C.7: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

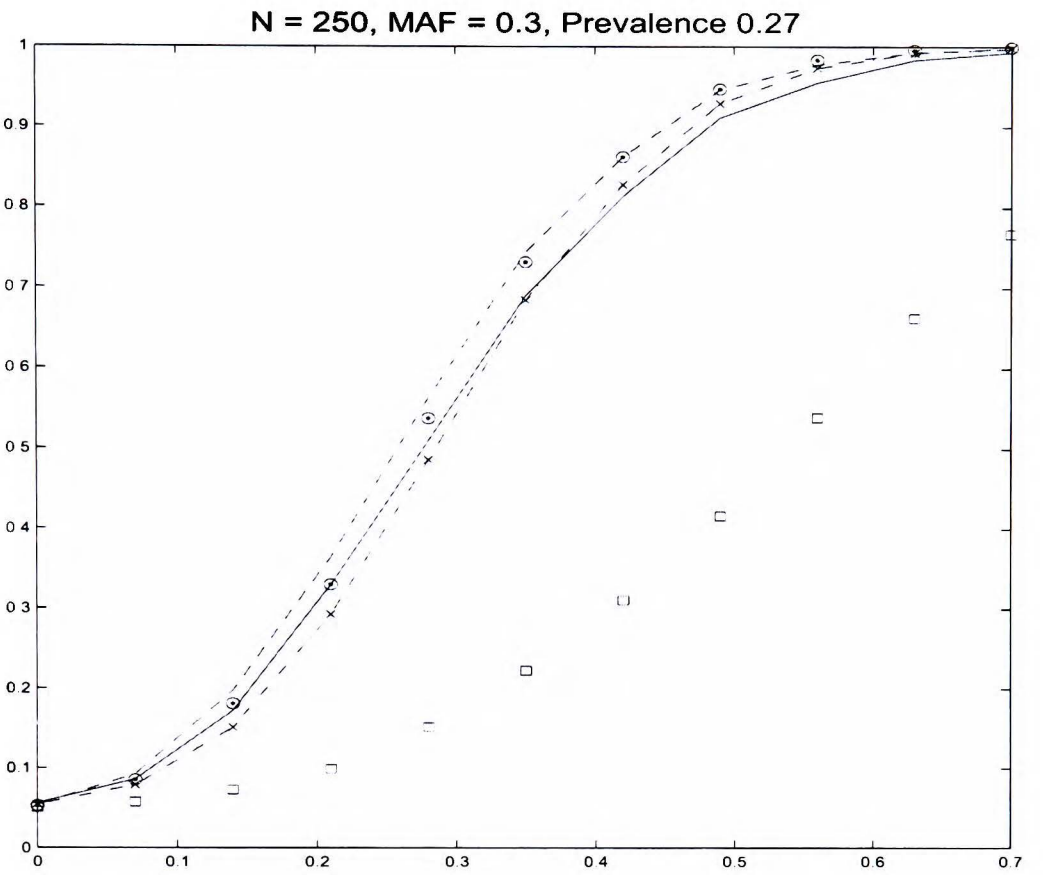


Figure C.8: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

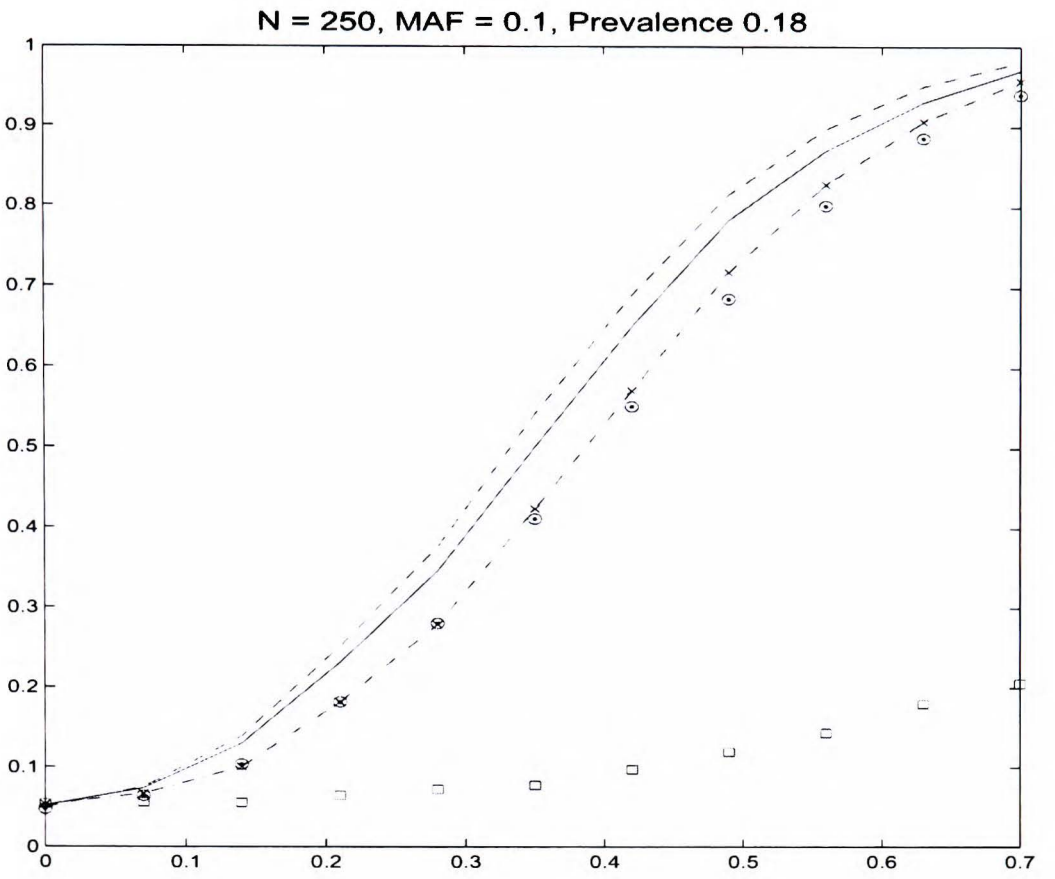


Figure C.9: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

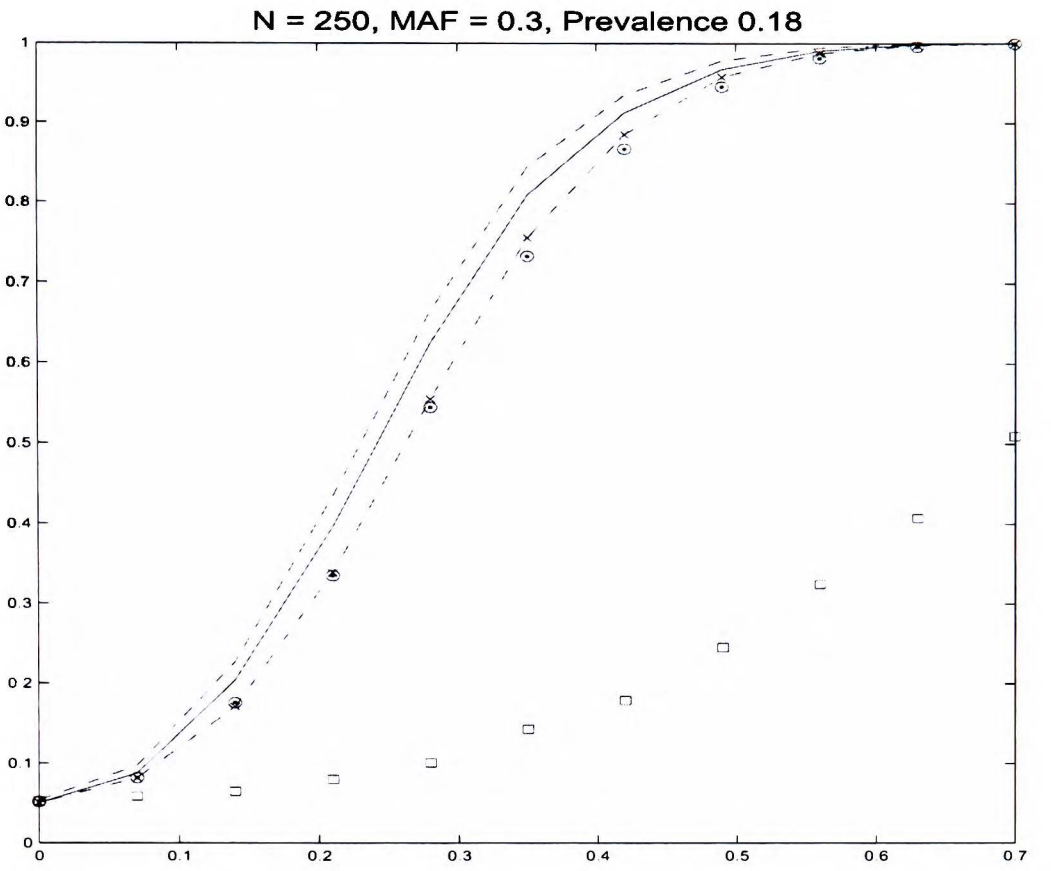


Figure C.10: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

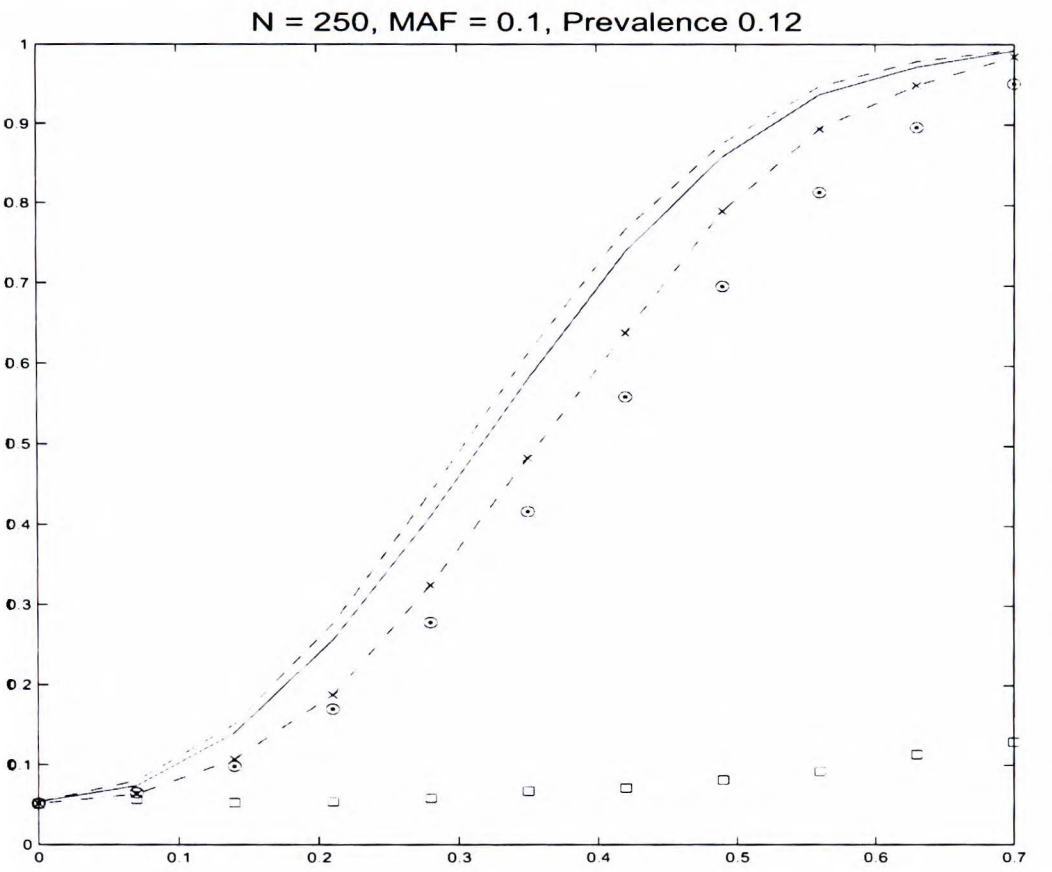


Figure C.11: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

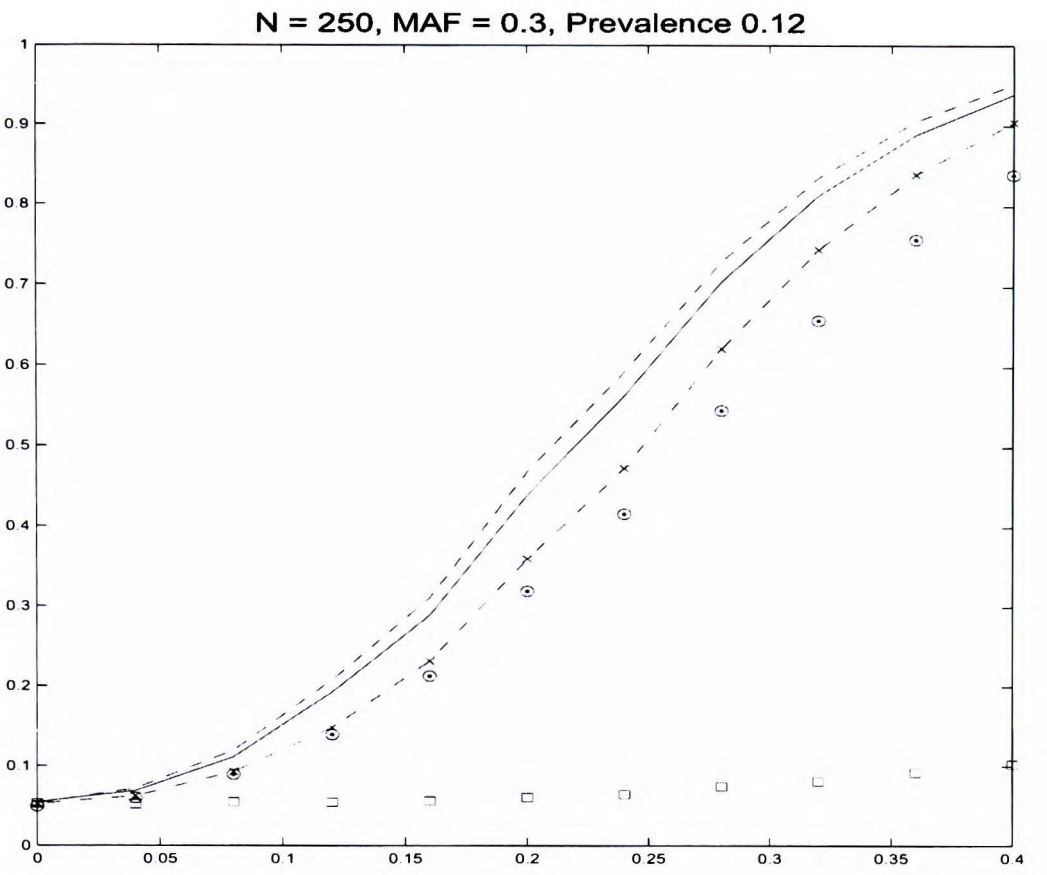


Figure C.12: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

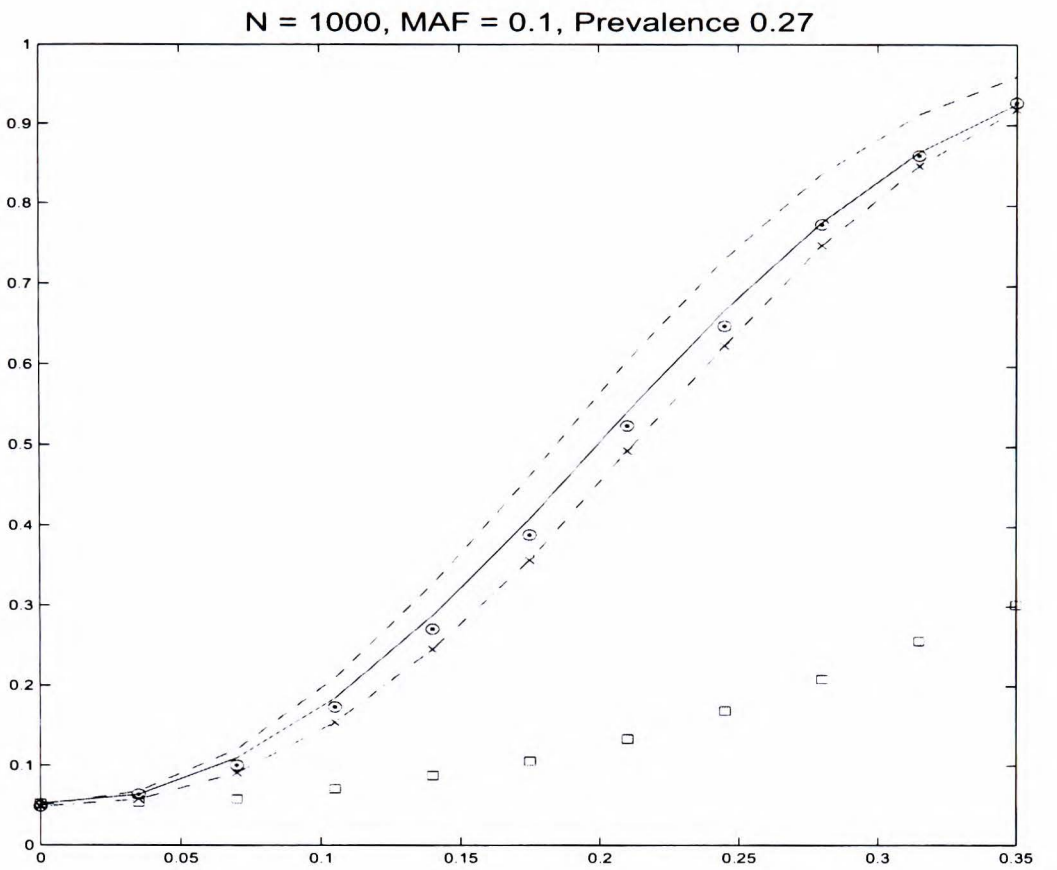


Figure C.13: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

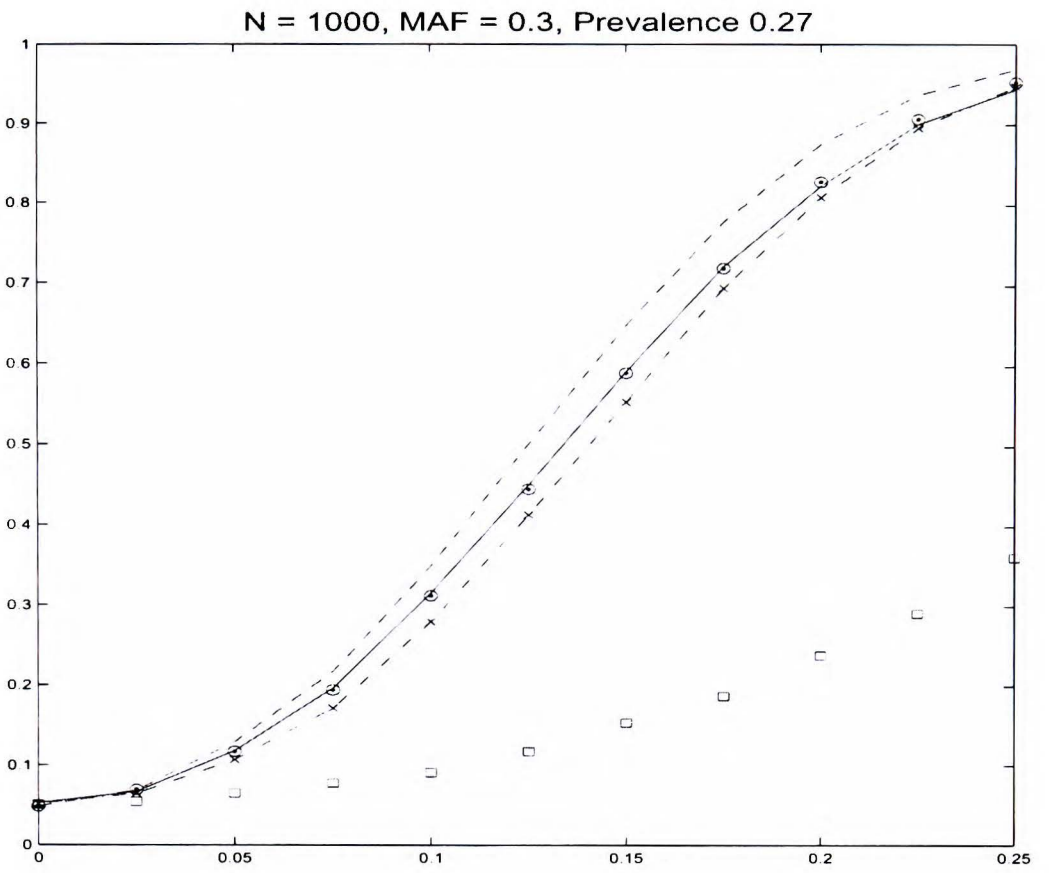


Figure C.14: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

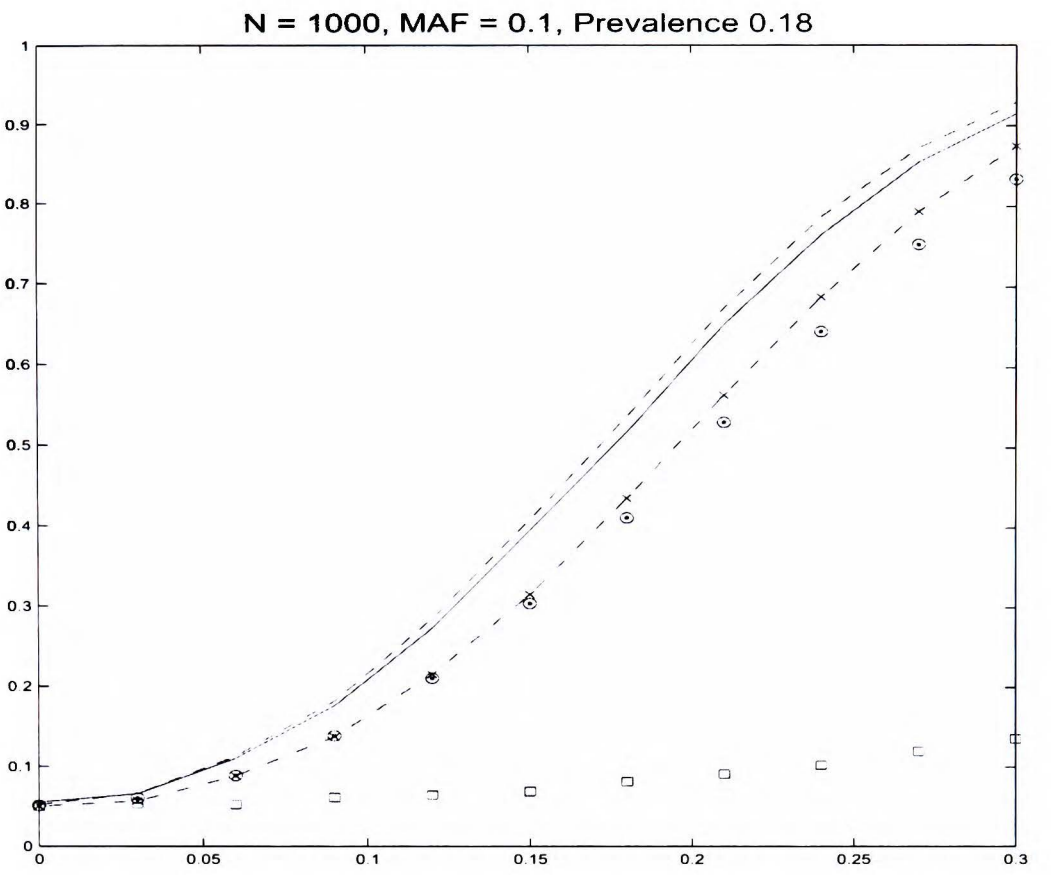


Figure C.15: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

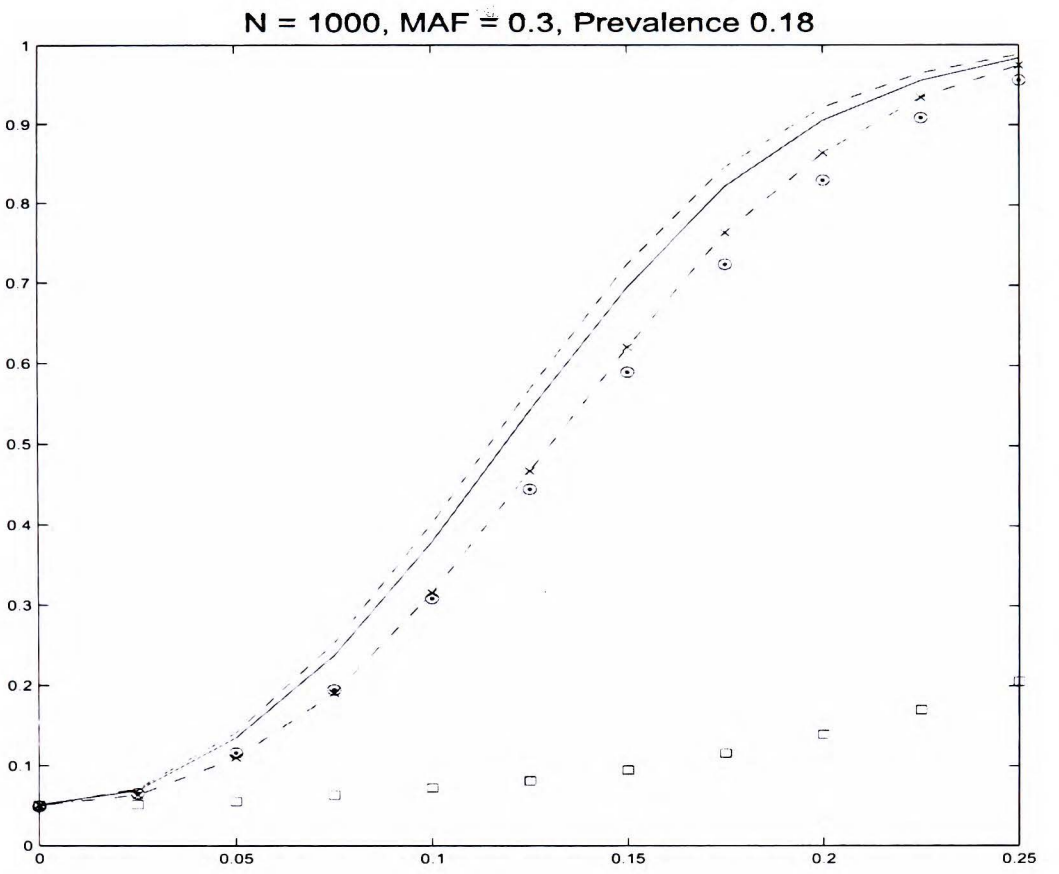


Figure C.16: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

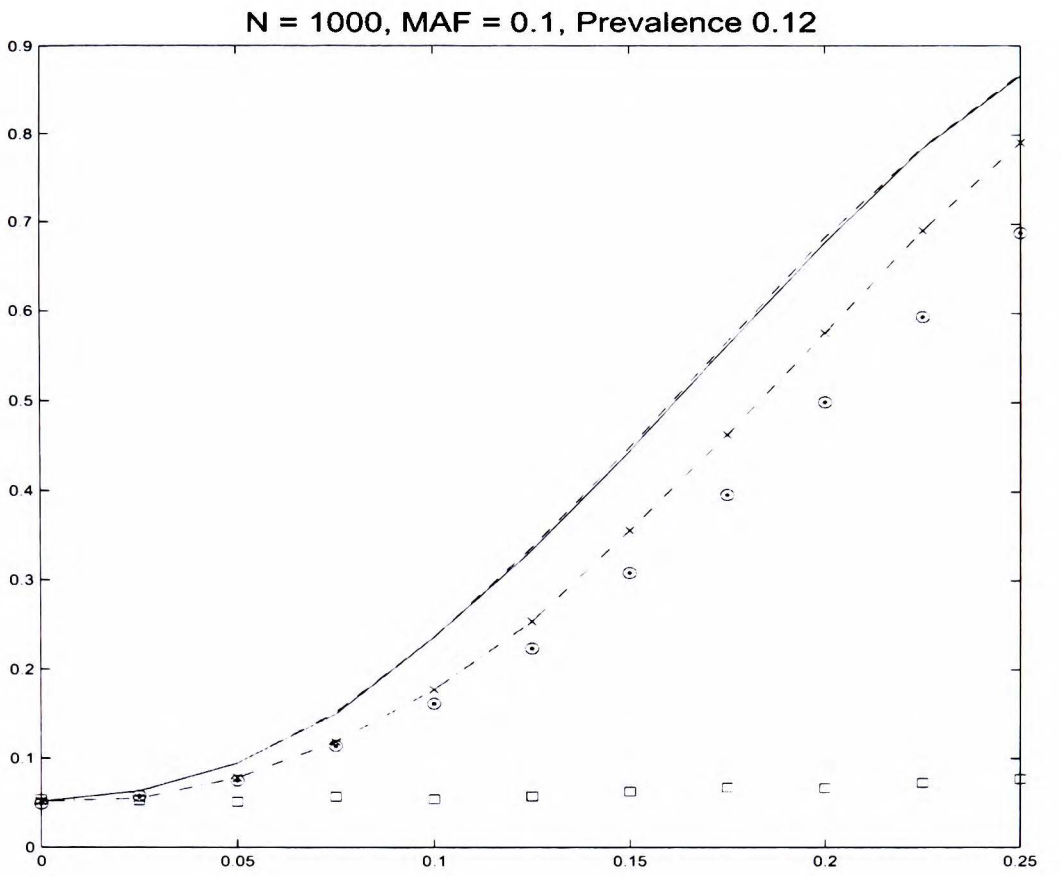


Figure C.17: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

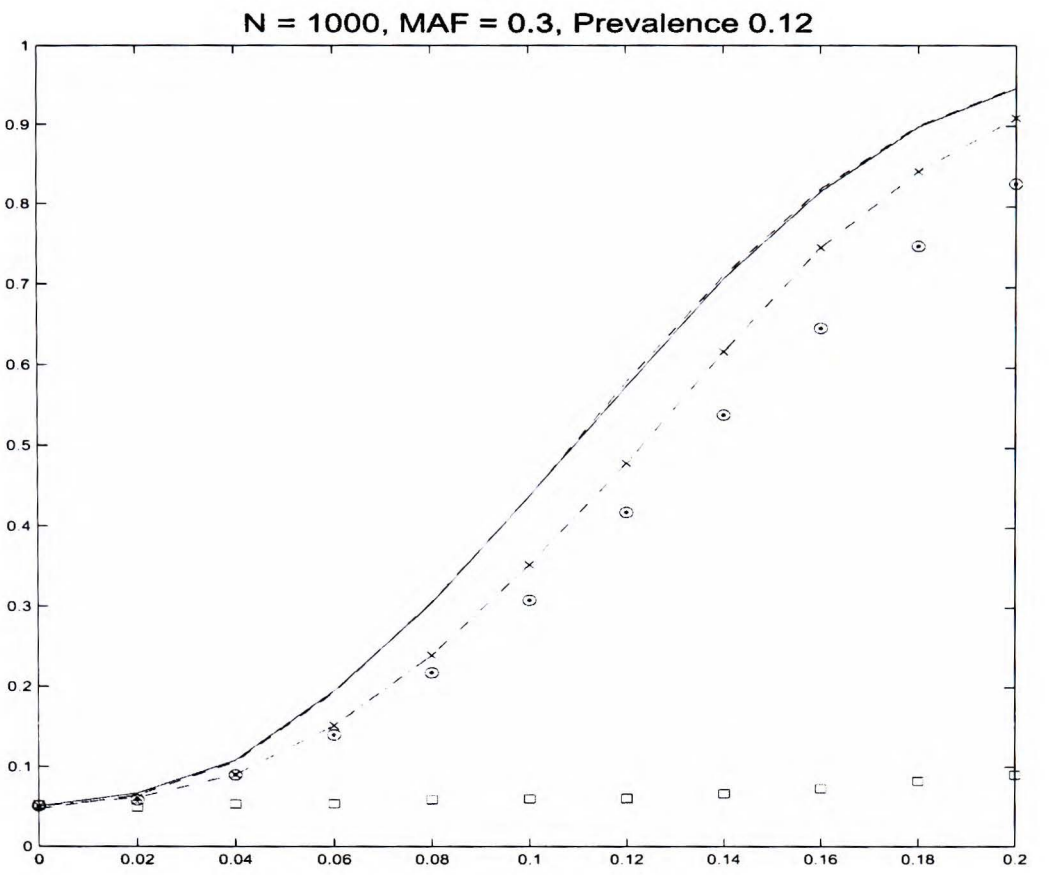


Figure C.18: Estimated power of the six frequentist tests and Bayes test. Dashed line represents the Bayesian test, solid line is case-only test, circles - logistic, dots - test of homogeneity, crosses - Pearson's test, dashes and dots - two parameter cases and controls test, and squares - control-only test. Type I error rate 5%, 10,000 simulations, estimation standard error ≈ 0.002

References

Aaltonen, L., Johns, L., Jarvinen, H., Mecklin, J.P., Houlston, R. (2007). Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clin. Cancer Res*, **13**: 356-361.

Antonyuk, A., Holmes, C. (2009). On testing for genetic association in case-control studies when population allele frequencies are known. *Genetic Epidemiology*, **33(5)**: 371-378.

Ambrosius, W.T., Lange, E.M., Langefeld, C.D. (2004). Power for genetic association studies with random allele frequencies and genotype distributions. *Am J Hum Genet*, **74(4)**: 683-93.

Asslaber, M., Zatloukal, K. (2007). Biobanks: transnational, European and global networks. *Briefings in Functional Genomics and Proteomics*, October 4, 2007.

Austin, M.A., Harding, S., McElroy, C. (2003). Genebanks: a comparison of eight proposed international genetic databases. *Community Genetics*, **6**: 37-45.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet*, **7**: 781-791.

Balding, D.J., Bishop, M., Cannings, C., editors (2003). *Handbook of Statistical Genetics*, 2nd edition. John Wiley and Sons.

Bernardo, J.M., Smith, A.F.M. (1994). *Bayesian Theory.* Wiley, New York.

Berger, J.O. (1993). *Statistical decision theory and Bayesian analysis. 2nd edition.* Springer, New York.

Bickel, P.J., Doksum, K.A. (2001). *Mathematical Statistics, Basic Ideas and Selected Topics.* Prentice Hall.

Bingham, S., Riboli, E. (2004) Diet and cancer the European prospective investigation into cancer and nutrition. *Nature Rev Cancer*, **4**: 206-215.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**: 362.

Boos, D.D., Brownie, C. (1986). Testing for a treatment effect in the presence of nonresponders. *Biometrics*, **42**: 191-197.

Breslow, N.E. (1996). Statistics in Epidemiology: The Case-Control Study. *J Amer Statist Assoc*, **91**: 14-28.

Broderick, P. *et al.* (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet*, **39**: 1315-1317.

Broman, K., Speed, T. (1999). A review of methods for identifying QTLs in experimental crosses. *Proceedings of the Joint Summer Research Conference on Statistics in Molecular Biology*, edited by Seillier-Moiseiwitsch F, Seattle.

Casella, G., George, E.I. (1992) Explaining the Gibbs sampler. *American Statistician*, **46**: 167-174.

Cavalli-Sforza, L.L., Menozzi, P., Piazza, A. (1994). The History and Geography of Human Genes. Princeton University Press.

Chapelle, A. (2004). Genetic predisposition to colorectal cancer. *Nature reviews. Cancer*, **Vol. 4, No. 10**: 769–780.

Cockerham, C.C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, **39**: 859–882.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*, **38**: 75-81.

Cornfield, J. (1951). A method for estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute*, **11**: 1269–1275.

Cover, T.M., Thomas, J.A. (1991). *Elements of information theory.* John Wiley and Sons.

Cramer, H. (1946). *Mathematical Models of Statistics.* Princeton University Press.

Darvasi, A., Soller, M. (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, **141**: 1199–1207.

Davis, R. L., Khoury, M. J. (2007). The Emergence of Biobanks: Practical Design Considerations for Large Population-Based Studies of Gene-Environment Interactions. *Community Genetics*, **10**: 181–185.

Day, N.E. (1969). Estimating the components of a mixture of normal components. *Biometrika*, **56**: 464–474.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J Roy Statist Soc B*, **39**: 1–38.

Eaton, M.L., Giovagnoli, A., Sebastiani, P. (1996). A predictive approach to the Bayesian design problem with application to normal regression model. *Biometrika*, **83**: 111–125.

Ewens, W.J., Spielman, R.S. (2003). *Handbook of Statistical Genetics 2nd edn*: 961–972. Wiley, New York.

Feng, Z., McCulloch, C. (1994). On the Likelihood Ratio Test Statistic for the

Number of Components in a Normal Mixture with Unequal Variances. *Biometrics*, **50**: 1158–1162.

Flint, J., Valdar, W., Shifman, S., Mott, R. (2005). Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Rev Genet*, **6**: 271–286.

Fowlkes, E.B. (1979). Some methods for studying the mixture of two normal (log-normal) distributions. *J Amer Statist Assoc*, **74**: 561–575.

Fryer, J.G., Robertson, C.A. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika*, **59**: 639–648.

Geman, S., Geman, D. (1984) Stochastic relaxation, gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intell.*, **6**: 721–741.

Goffinet, B., Loisel, P., Laurent, B. (1992). Testing in normal mixture models when the proportions are known. *Biometrika*, **79**: 842–846.

Green, P.J. (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J Royal Statist Soc* B, **46**: 149–192.

Guedj, M., Nuel, G., Prum, B. (2008) A note on allelic tests in case-control association studies. *Annals of Human Genetics*, **72**: 407–409.

Hagen, H.E., Carlstedt-Duke, J. (2004). Building global networks for human diseases: genes and populations. *Nature Medidicne*, **10**: 665–667.

Haldane, J.B.S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, **8**: 299–309.

Helgason, A., Nicholson, G., Stefansson, K., Donnelly, P. (2003). A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci for relative homogeneity caused by genetic drift. *Ann. Hum. Genet.*, **67**: 281–297.

Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., Stefansson, K. (2005). An Icelandic example of the impact of population structure on association studies. *Nat Genet*, **37**: 90–95.

Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., McKeigue, P.M.. (2004). Design and analysis of admixture mapping studies. *Am J Hum Genet*, **74(5)**: 965–978.

Holmans, P. (2003). *Handbook of Statistical Genetics, 2nd edition*: 919–938. Wiley, New York.

Horner, M.J. (2006). SEER Cancer Statistics Review, 1975–2006. *National Cancer Institute, USA*.

Hosmer, D.W., Lemeshow, S. (2000). *Applied logistic regression, 2nd edition.* Wiley, New York.

Ibrahim, J.G. Chen, M-H. Lipsitz, S.R, and Herring, A.H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. *J Amer Statist Assoc*, **100**: 332–345.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437(7063)**: 1299–1320.

Jamshidian, M., Jennrich, R.I. (2000). Standard errors for EM estimation. *J. R. Stat. Soc., Ser. B*, **62**: 257–270.

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

Jansen, R.C. (2001). Quantitative trait loci in inbred lines, *Handbook of Statistical Genetics*, pp. 567–597. Wiley, New York.

Kao, C. and Zeng, Z. (1997). General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics*, **53**: 653–665.

Kao, C., Zeng, Z., Teasdale, R. (1999). Multiple interval mapping for quantitative trait loci. *Genetics*, **152**: 1203–1216.

Kao, C. (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics*, **156**: 855–865.

Kass. R., Raftery, A. (1995). Bayes Factors. *J Amer Statist Assoc*, **90**: 773–795.

Kemp, Z., *et al.* (2006). Evidence for a colorectal cancer susceptibility locus on chromosome 3q21-q24 from a high-density SNP genome-wide linkage scan. *Hum Mol Genet*, **15**: 2903–2910.

Knapp. S.J.. Bridges, W.C, JR., Birkes, D. (1990). Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics*, **79**: 583–592.

Knott. S.A., Haley, C.S. (1992). Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetics Research*, **60**: 139–151.

Korstanje, R., Paigen, B. (2002). From QTL to gene: The harvest begins. *Nat Genet*, **31**: 235–236.

Krzanowski. W.J. (2007). *Statistical principles and techniques in scientific and social investigations.* Oxford University Press, New York.

Lander. E., Botstein, D.(1989). Mapping mendelian factors underlying quantitative traits using r p linkage maps. *Genetics*, **121**: 185–197.

Lichtenstein, P. *et al.* (2000). Environmental and heritable factors in the causation of cancer-analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*, **343(2)**: 78–85.

Little, R.J.A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Wiley, New York.

Lubin, J.H., Gail, M.H. (1984) Biased selection of controls for case-control analyses of cohort studies. *Biometrics*, **40**: 63–75.

Lunn, D.J. , Whittaker, J.C., Best, N.A. (2006). Bayesian toolkit for genetic association studies. *Genet Epidemiol*, **30**: 231–247.

Manolio, T.A.. Bailey-Wilson J.E., Collins, F.S. (2006) Genes, environment and the value of prospective cohort studies. *Nat Rev Genet*, **7**: 812–820.

Marchini, J.. Howie, B.. Myers, S., McVean, G., Donnelly, P. (2007). A new multipoint method for genome-wide association studies via imputation of genotypes. *Nature Genetics* **39**: 906–913.

Marshall, R.J. (1988). Bayesian analysis of case-control studies. *Statistics in Medicine* **7**: 1223–1230.

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

Mayer, M. (2005). A comparison of regression interval mapping and multiple interval mapping for linked QTL. *Heredity*, **94**: 599–605.

McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, **36**: 318–324.

McLachlan, G.J., Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

McLachlan, G.L., Basford, K.E. (1988). *Mixture Models.* New York: Marcel Dekker.

Meilijson, A. (1989). A fast improvement to the EM algorithm on its own terms. *J. Roy. Statist. Soc., Ser. B.*, **51**: 127–138.

Meng, X.-L., Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**: 267–278.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**: 1087–1092.

Minka, T.P. (1998). Expectation-maximization as lower bound maximization. Online article. http://citeseer.nj.nec.com/minka98expectationmaximization.html.

Morris, A. (2005). Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modeling of haplotypes. *Genet Epidemiol*, **29**: 91–107.

Mott, R. (2006). Finding the molecular basis of complex genetic variation in humans and mice. *Philos Trans R Soc Lond B Biol Sci*, **361(1467)**: 393–401.

Mott, R., Flint, J. (2002) Simultaneous Detection and Fine Mapping of Quantitative Trait Loci in Mice Using Heterogeneous Stocks. *Genetics*, **160**: 1609–1618.

Mott, R., Talbot, C.J., Turri, M.G., Collins, A.C., and Flint, J. (2000). A new method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci.*, **97**: 12649–12654.

Mukherjee, B., Sinha, S., Ghosh, M. (2005). Bayesian Analysis for case-control studies: A review article. *Handbook of Statistics*, **25**: 793–819.

Müller, P., Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**: 523–537.

Munday, P., Holmes, C.C. (2004). *QTL Mapping using a Maximum-Likelihood Approach.* Technical Report. Department of Statistics, University of Oxford.

Niu, T., Ding, A., Kreutz, R., Lindpaintner, K. (2005). An Expectation-Maximization Likelihood-Ratio Test for Handling Missing Data: Application in Experimental Crosses. *Genetics*, **169**: 1021–1031.

Nurminen, M., Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics* **14**: 67–77.

Padmanabhan, S., Melander, O., Hastie, C., Menni, C., Delles, C., Connell, J.M., Dominiczak, A.F. (2008) Hypertension and genome-wide association studies: combining high fidelity phenotyping and hypercontrols. *Journal of Hypertension*, **26(7)**: 1275–1281.

Pearson, K. (1894). *Contributions to the mathematical theory of evolution.* Phil. Trans. Roy. Soc.. London. **A 185**: 71–110.

Pharoah, P.D., Tyrer, J., Dunning, A.M., Easton, D.F., Ponder, B.A. (2007). Association between common variation in 120 candidate genes and breast cancer risk. *PLoS Genet*, **3**: e42.

Picard. F. (2007). An introduction to mixture models. Research report, http://genome.jouy.inra.fr/ssb/

Potter, J. D. (1999). Colorectal cancer: molecules and populations. *J. Natl Cancer Inst*, **91**: 916–932.

Pritchard. J.K.. Stephens, M., Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**: 945–959.

Pritchard, J.K., Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, **65**: 220–228.

Quarini, C.. Gosney, M. (2009) Review of the evidence for a colorectal cancer screening programme in elderly people. *Age Ageing*, **July 6** (Epub ahead of print).

Rao, D.C., Gu, C. (2008) Genetic dissection of complex traits. Academic Press, San Diego.

Rapp, J.P. (2000). Genetic analysis of inherited hypertension in the rat. *Physiol Rev*, **80**: 135–72.

Rebbeck, T.R., Ambrosone, C.B., Shields, P.G. (2008). Molecular epidemiology. Informa Healthcare, New York.

Reich, D.A., Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends Genet*, **17**: 502–510.

Risch, N.J. (2000). Searching for genetic determinants in the new millennium. *Nature*, textbf405: 847–856.

Robert, C.P. (2001). *The Bayesian choice, 2nd edition.* Springer, New York.

Rodolphe, F., Lefort-Buson, M. (1993). A multi-marker model for detecting chromosomal segments displaying QTL activity. *Genetics*, **134**: 1277–1288.

Satagopan, J.M., Yandell, B.S. (1996) Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases, Biometric Section, Joint Statistical Meetings, Chicago, IL. (available at ftp://ftp.stat.wisc.edu/pub/yandell/revjump.html)

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

Sasieni, P.D., (1997). From genotypes to genes: doubling the sample size. *Biometrics*, **53**: 1253–1261.

Seaman, S.R., Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case–control studies. *Biometrika*, **91**: 15–25.

Sen, S., Churchill, G.A. (2001). A statistical framework for quantitative trait mapping. *Genetics*, **159** : 371–387.

Servin, B., Stephens, M. (2007). Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLOS Genetics*, **3**: e114.

Sillanpaa, M., Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, **148**: 1373–1388.

Sladek, R., *et al.* (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**: 881–885.

Slager, S.L., Schaid, D.J. (2001) Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Hum Hered*, **52**: 149–153.

Smith, A.F.M., Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *J. Roy. Statist. Soc., Ser. B*, **55**: 3–23.

Smith, M.W., O'Brien, S.J. (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Rev Genet*, **6**: 623–266 .

Speed, T.P. (2005). Genetic map functions. In: Encyclopedia of Biostatistics. Wiley, New York

Stephens, D.A., Fisch, R.D. (1996) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Technical report. (available at http://www.ma.ic.ac.uk/statistics/techrep.html)

Stephens, D.A., Smith, A.F.M. (1993) Bayesian inference in multipoint gene mapping. *Ann. Human Genet.*, **57**: 65–82.

Tai, J.J. (1989) Application of Bayesian decision procedure to the inference of genetic linkage. *J. Am. Statist. Assoc.*, **84**: 669–673.

Talbot, C.J., Nicod, A., Cherny, S.S., Fulker, D.W., Collins, A.C., Flint, J. (1999) High-resolution mapping of quantitative trait loci in outbred mice, **2**: 305 – 308.

Tan, W.Y., Chang, W.C. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *J Amer Statist Assoc*, **67**: 702708.

Terwilliger, J.D., Weiss, K.M. (1998). Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol*, **9**: 578–594.

Thode, H.C., Finch, S.J., and Mendell, N.R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for mixture of two normals. *Biometrics*, **44**: 1195–1201.

Thomas, D.C., Cortessis, V. (1992) A Gibbs sampling approach in linkage analysis. *Hum. Hered.*, **42**: 6376.

Thomas, D.C. (2004). *Statistical Methods in Genetic Epidemiology.* Oxford Univ Press.

Titterington, D.M., Smith, A.F.M., Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Tomlinson, I. *et al.* (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet*, **39**: 984–988.

Tomlinson, I. *et al.* (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet*, **40(5)**: 623–630.

Valdar, W. (2003). *Detection of Epistatic Quantitative Trait Loci in Murine Heterogeneous Stock Recombinant Inbred Lines.* MSc thesis. Department of Statistics, University of Oxford.

Wakefield, J. (2007). A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet*, **81**: 208–226.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**: 661–78.

Wolfe, J.W. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**: 329–350.

Yalcin, B., Fullerton, J., Miller, S., Keays, D.A., Brady, S., Bhomra, A., Jefferson, A., Volpi, E., Copley, R.R., Flint, J., Mott, R. (2004) Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc Natl Acad Sci* **101**: 9734–9739.

Zanke, B.W. *et al.* (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **39**: 989–994.

Zeggini, E. et al. (2005). An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nature Genet*, **37**: 1320–1322.

Zeng, Z.-B. (1993a). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci.*, **90**: 10972–10976.

Zeng, Z-B. (1993b). Precision mapping of quantitative trait loci. *Genetics*, **136**: 1457–1468.

Zondervan, K.T., Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. *Nat Rev Genet*, **5**: 89–100.