Methods from Statistical Computing for Genetic Analysis of Complex Traits

BEHRANG MAHJANI
Abstract


The goal of this thesis is to explore, improve and implement some advanced modern computational methods in statistics, focusing on applications in genetics. The thesis has three major directions.

First, we study likelihoods for genetics analysis of experimental populations. Here, the maximum likelihood can be viewed as a computational global optimization problem. We introduce a faster optimization algorithm called PruneDIRECT, and explain how it can be parallelized for permutation testing using the Map-Reduce framework. We have implemented PruneDIRECT as an open source R package, and also Software as a Service for cloud infrastructures (QTLaaS).

The second part of the thesis focusses on using sparse matrix methods for solving linear mixed models with large correlation matrices. For populations with known pedigrees, we show that the inverse of covariance matrix is sparse. We describe how to use this sparsity to develop a new method to maximize the likelihood and calculate the variance components.

In the final part of the thesis we study computational challenges of psychiatric genetics, using only pedigree information. The aim is to investigate existence of maternal effects in obsessive compulsive behavior. We add the maternal effects to the linear mixed model, used in the second part of this thesis, and we describe the computational challenges of working with binary traits.

Keywords: Statistical Computing, QTL mapping, Global Optimization, Linear Mixed Models

Behrang Mahjani, Department of Information Technology, Division of Scientific Computing, Box 337, Uppsala University, SE-751 05 Uppsala, Sweden.

© Behrang Mahjani 2016

ISSN 1651-6214
urn:nbn:se:uu:diva-284378 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-284378)
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


V  Software as a Service in analysis of Quantitative Trait Loci, Behrang Mahjani, Salman Toor, Supporting material for article IV, 2016.

VI  Fitting Linear Mixed Models using Sparse Matrix Methods and Lanczos factorization, with applications in Genetics. Behrang Mahjani, Lars Rönnegård, Lars Eldén, Submitted.


Reprints were made with permission from the publishers.
Contents

1 Introduction ............................................................................................................... 7
  1.1 Aim and overview of the Work .............................................................................. 7
  1.2 Biological background ........................................................................................ 8
    1.2.1 Basic definitions .......................................................................................... 8
    1.2.2 Cell division ............................................................................................... 9
    1.2.3 Mendelian traits and QTL ............................................................................ 9
    1.2.4 Genetic maps ............................................................................................. 11
    1.2.5 Heritability and genetic value ..................................................................... 11
    1.2.6 Experimental populations ......................................................................... 12
    1.2.7 Genetic markers ....................................................................................... 12
  1.3 Research ethics ...................................................................................................... 13
    1.3.1 Collecting data ........................................................................................... 14
    1.3.2 Working with data ..................................................................................... 15

2 Optimization algorithms for QTL searches .............................................................. 16
  2.1 Basics of QTL analysis ....................................................................................... 16
  2.2 QTL search .......................................................................................................... 18
    2.2.1 Lipschitz optimization ................................................................................. 18
    2.2.2 DIRECT .................................................................................................... 20
    2.2.3 DIRECT and QTL search ......................................................................... 23
    2.2.4 PruneDIRECT ......................................................................................... 23
  2.3 Significance test .................................................................................................... 25
    2.3.1 PruneDIRECT for permutation testing ..................................................... 26
    2.3.2 Parallel frameworks for permutation testing .............................................. 26
  2.4 QTL as a Service: Cloud computing platform for QTL searches ....................... 27
  2.5 Outlook ................................................................................................................ 28

3 Sparse matrix techniques for linear mixed models .................................................. 29
  3.1 Linear mixed models ........................................................................................... 29

4 Computational challenges in modeling maternal effects in psychiatric disorders .................................................. 31
  4.1 Introduction psychiatric genetics ....................................................................... 31
  4.2 Linear mixed model with maternal effects ......................................................... 32

5 Summary of Attached Papers ................................................................................. 34

6 Svensk Sammanfattning ......................................................................................... 37
1. Introduction

1.1 Aim and overview of the Work

We are living in a complex world manifested by the vast amount of data coming to our life. Scientists and scholars in diverse disciplines are paving the way to analyze and interpret data. A new era has just started, marked with fast and huge generation and exchange of information. Much endeavor is needed to handle and analyze such amounts of data. As an example in life science, we already have a large collection of genetics data, together with other biological information and diagnoses. These data sets will grow even more with the new trend of cheaper and more accessible sequencing technologies, accompanied with almost live health monitoring systems, such as activity trackers with biological sensors. The question which arises here is how we should handle and analyze such large amounts of data.

An important starting point can be that we set our goal for what we are looking for, before we start analyzing a data set. In statistical language, the central dogma of statistics says that data should be viewed as realizations of random variables [22]. Therefore, one should construct a hypothesis to verify if a specific random process generated the data. This means that one should formulate a "meaningful" question (hypothesis), before analyzing a data. Simulation methods can be of great help to find a meaningful question.

After formulating a hypothesis, we need to verify it. Many of the classical approaches towards modeling data fails when they deal with large data sets, because of new sources of uncertainty, in addition to higher computational errors and larger computational demands [14, 43].

It is crucial to mention that uncertainty and randomness are an inseparable part of real data sets, regardless of the existence of the underlying deterministic or seemingly undeterministic process. This uncertainty can influence the calculations in different ways. For example, it is not meaningful to enforce that the numerical error in solving a statistical model must be significantly smaller than the error in the data. Another example is that you can never fit a model perfectly to the data, since there are always unknown sources of randomness. Therefore, it is very challenging to find the best model explaining a real data set. These concerns become more critical when dealing with larger data sets.

Computational methods come to help statistics in different ways. Computational methods are used to drive numerical solutions of statistical problems. Here, efforts on developing numerical linear algebra, developing algorithms,
parallelizing algorithms, data structures and modern infrastructures such as clouds are all needed. A former president of the International Association for Statistical Computing, Carlo Lauro, has defined the term Computational Statistics as the application of Computer Science to Statistics. We propose to extend the definition to the application of Scientific Computing and Computer Science to Statistics. This definition is consistent with the work presented in this thesis.

The aim of the thesis is to explore and enhance a number of the methods in statistical computing in Life Science. The work is divided in three parts. First, we study likelihoods for genetics analysis of experimental populations. One can consider at maximum likelihood as a computational global optimization problem. We introduce a faster optimization algorithm and explain how one can parallelize it using modern computer infrastructures. We also explain the concept of Software as a Service and how one can benefit from cloud computing.

In the second part, we consider applications of sparse matrices in linear models. One can use the sparsity of the matrices in a model to make the calculation faster and decrease memory demands when working with large matrices.

In the last part, we study psychiatric genetics using only pedigree information. We discuss how one can formulate a meaningful question using simulation studies. We try to use part of our new optimization algorithm from part one, together with our sparse matrix techniques from part two, to analyze the model.

As an overarching theme, in all parts of this work, we are modifying and enhancing classical algorithms in scientific computing, adapting them for statistical problems. As an example, one can use a classical derivative free optimization algorithm to find a global optimum for an undeterministic function. We modify one of these classical methods by understanding the randomness in the data to make it more efficient.

1.2 Biological background
In this section we cover a basic introduction to genetics.

1.2.1 Basic definitions
Our inherited biological characters are coded in *Deoxyribonucleic Acid* (DNA). DNA has a special structure in cells which is called a *chromosome*. DNA is a polymer that consists of building blocks called *nucleotides*. There are four different types of nucleotides which can be distinguished by their bases, cytosine (C), adenine (A), guanine (G) and thymine (T). We define a *gene* as a particular segment of DNA that specifies the structure of a protein. In other
words, a gene describes the characteristics of an offspring, which are inherited from parents. A locus (plural loci) is a specific location along a chromosome. A locus might clearly map to a gene.

The genetic makeup of an organism is called genotypes. On the other hand, phenotypes are the description of actual physical characteristics, such as eye color or height. Genotypes, epigenetic factors, and non-inherited environmental factors are the elements that control a phenotype. Epigenetic factors are factors that affect how cells read genes.

Each gene can exist in alternative forms, alleles. In humans and most animals, cells other than sex cells, have two sets of each chromosomes, they are therefore called diploid cells. If a cell contains only a single set of chromosomes, such as a human sex cell, then it is haploid. If both copies of a gene in a diploid cell are similar, then the individual is called homozygous for that specific gene, otherwise, it is called heterozygous.

A diploid locus with two alleles \( a \) and \( A \) can have three possible genotypes \( AA, aa, Aa \). If the effect of one allele is dominated by the effect of the other allele, then we call the dominating allele dominant and the other allele recessive. As an example, if \( A \) is dominant, then the phenotype of the heterozygous \( Aa \) is similar to \( A \). Brown eye color is dominant over blue.

1.2.2 Cell division

One of the main causes of genetic variation is genetic recombination. Genetic recombination is the process where genetic material breaks and join other genetic material. Long regions are exchanged, hundreds of genes between sister chromatids. A sister chromatid is any of the two identical copies formed by the replication of a single chromosome.

There are two types of cell divisions, meiosis and mitosis. Mitosis is when a mother cell divides into two genetically identical daughter cells. On the other hand, meiosis is a reproductive cell division and is one of the main sources of genetic diversity. During meiosis, a mother cell divides into four cells, called gametes. Each of the gametes carry only half the genetic material of the mother. These procedures are illustrated in Figure 1.1.

Chromosomal crossover is the final stage of genetic recombination during meiosis. In genetic crossover for two strands of DNA (in the simplest model), each strand of DNA breaks and rejoins to the other strand of DNA. In other words, genetic recombination is a kind of "exchange of genetic materials" between two sister chromatids.

1.2.3 Mendelian traits and QTL

Mendel’s work, published in 1866, was about applying artificial fertilization on pea plants in order to obtain new variations in their colors [1]. This work
Figure 1.1. Mitosis and meiosis. Mitosis create identical cell while meiosis creates the reproduction cells. Chromosomal crossover is the final stage of genetic recombination during meiosis.

was the beginning of explaining hereditary traits, from parents to offsprings, which was later called Mendelian inheritance or Mendel’s laws.

Mendel’s law consists of two parts, the law of segregation and the law of independent assortment. Mendel’s first law, the law of segregation, contains four parts:

- Variations in inherited characteristics are caused by different alleles.
- Each offspring gets one allele from the father and one allele from the mother.
- For each offspring, if two alleles (inherited from the parents) are different, one of them will be dominant (results in a specific physical characteristic) and the other will be recessive (does not result in a specific physical characteristic).
- The two alleles (inherited from the parents) segregate during sex cell (gamete) production.

Mendel’s second law (the law of inheritance) indicates that different characteristics are inherited independently.
Mendelian traits are such traits that exhibit only dominant or recessive contribution from a single gene. In contrast, non-Mendelian traits do not follow Mendel’s laws. Two examples of non-Mendelian traits are *incomplete dominance* and *co-dominance*. Incomplete dominance within individual genes is when a heterozygous organism has phenotypes of both the dominant and the recessive allele. Co-dominance is when both alleles contribute equally to the traits. Most of the complex traits are non-Mendelian traits, but mainly due to mainly polygenetic nature.

The traits that fall into distinct classes are called *discontinuous traits*. The traits that have continuous distribution are referred to as *quantitative traits*.

Understanding the relation between genes and traits is a fundamental problem in genetics. Such knowledge can lead to e.g. the identification of possible drug targets, treatment of heritable diseases, and efficient designs for plant and animal breeding. The aim of quantitative trait loci (QTL) analysis is to locate regions in the genome which can be associated to quantitative traits. Many such traits are affected by both genetic and environmental factors, as well as their interactions.

### 1.2.4 Genetic maps

There is higher probability for loci that are close to each other to be together during meiosis. These loci are then linked. We can measure the distance between two genes in terms of *recombination frequencies*. The recombination frequency is the frequency that a crossover take place between two genes during meiosis and it is measured in centimorgan (cM). 1 cM is a recombination frequency of 1 percent. A list of loci based on genetic distance is called *genetic map* or *linkage map*. When genetic distance between two genes are larger, then the chance that they will not get inherited together is higher.

*LOD score*, logarithm of the odds, is the likelihood of observing two loci are linked, to the likelihood of observing the same data purely by chance. This is an estimate to check if two genes are likely to be close to each other. A LOD score larger than 3 is considered evidence for linkage, and a LOD score smaller than -2 is an evidence of no linkage [8].

### 1.2.5 Heritability and genetic value

The *phenotypic value* of an individual is the sum of its *genotype effect*, $G$, and the *environmental effects*, $E$. The variance of the phenotypic values can be written as:

$$\text{var}(P) = \text{var}(G) + \text{var}(E)$$ (1.1)

Genotypic variation can be divided into *additive variation*, $A$, and *dominance variation*, $D$. Additive variation is the cumulative effect of loci, while dominance variation is the variation from the interaction between alleles.
some cases, there is also interaction between different genes which is called epistatic effect, I. We can write the total variation in the phenotypes value as (neglecting the interaction between environment and genetic values):

\[ \text{var}(P) = \text{var}(A) + \text{var}(D) + \text{var}(E) + \text{var}(I) \] (1.2)

From the above definition we can estimate how much of the variation in a phenotype for a trait is inherited (genetic effects) and how much of it is due to environmental factors. Heritability, \( H \), is defined as how much of the variation in phenotype can be explained by genetic effects:

\[ H = \frac{\text{var}(G)}{\text{var}(P)} \] (1.3)

Consider a gene with two alleles \( A \) and \( a \). Each genotype gives a different value to the trait [8]:

\[ G = \begin{cases} 
    c, & \text{AA} \\
    d, & \text{Aa} \\
    -c, & \text{aa} 
\end{cases} \] (1.4)

Animal and plant breeders use a measure called breeding value, in addition to heritability. Breeding value is the sum of the average effects of the alleles. If there is no dominance then the genetic value is equal to the breeding value.

1.2.6 Experimental populations

Experimental populations help geneticists in animal and plant breeding to get a better understanding of genetics. Two of the most important experimental crosses are inbreed backcross and inbred intercross. Both of these population structures start with two diploid parents, where each parent has only one allele for each gene. In other words, the parents are completely inbred. As an example, in Figure 1.2, one parent has AA and the other parent has BB. The parents mate and create population F1. In this example, population F1 has AB.

In a backcross, the offspring in population F1 mates with another offspring. The result can be AB or BB. In intercross, the offspring in population F1 mate with one of the parents. The results can be AB, BB or AA, see Figure 1.3. The main difference between intercross and backcross populations is that in intercross populations we have both homozygous aa and AA.

1.2.7 Genetic markers

A genetic marker is a known physical position on a chromosome. Genetic markers can help to link a trait with the responsible genes. Today, dense genetic markers are available for many species. Some of the molecular genetic markers are:
Figure 1.2. Backcross. No heterozygous

- SNP (Single Nucleotide Polymorphism)
- RFLP (Restriction Fragment Length Polymorphism)
- SSR (Simple Sequence Repeat, or Microsatellite)
- AFLP (Amplified Fragment length Polymorphism)
- CAPS (Cleaved Amplified Polymorphic Sequence)
- dCAPS (Derived Cleaved Amplified Polymorphic Sequence)
- RAPD (Ramdomly Amplified Polymorphic DNA)

1.3 Research ethics

Society has already benefited from genetic research, and we are going to benefit even more from it in the future. While collecting and working with genetic information is useful for the society, it raises ethical issues as well. It is of great importance to be aware of these issues. Thereby, we review a few of the ethical concerns most clearly related to collecting and working with genetic data.

It is important to mention that each country has its own legal and ethical protocols for collecting and working with genetic data. Our aim here is not to discuss governmental protocols, but to address self-awareness for individual researchers beside the governmental routines. All ethical issues are not always covered by laws and governmental ethical protocols. We should be aware that having ethical approvals does not remove our own individual responsibility
1.3.1 Collecting data

Genetic data is an exceptional type of biological information since it is not only related to the specific individual, but also family information. It should be clear where and how the data is going to be used. It is also critically important to clarify who owns the data.

One of the main actions towards avoiding ethical issues is to have pre-test consulting with the participants. In such consultations, the interviewers should be aware of psychological factors of such interviews. If the individual have the right to know the result of the research on his or her data, then interviewer should get sure that the patient understand the result, and she or he is psychologically prepared for it. Interpreting results from genetic research is always challenging. Having a gene associated to a disease does not necessarily mean that the individual will get the disease. There are, in general, many environmental factors involved. The researcher should ensure that the data collectors are providing the participants with correct and understandable information.

Collecting animal or plant genetic data also involves important ethical issues. In case of animals, it is important to get sure that the test subjects are treated correctly. In plant genetics, the environmental factors of farming the plants should be thoroughly investigated. In such works one should also be
aware of the ethical issues of the impact of the project. One of the important impacts is the risk for environment and wildlife. As an example, if herbicide-resistant genes find their ways in weeds, there is a high risk for damaging crops. Another important issue is that biodiversity could be threatened if genetically modified plants breed with wild species.

Another aspect of ethical issues of collecting genetic data which is more related to this thesis is about security of the data when storing it electronically. It is important that the researcher checks if the necessary steps has been taken to keep the data secure from the very beginning. One should define different security levels and specify who can access which part of the data. One might consider making the data anonymous before storing it electronically.

1.3.2 Working with data

Working with genetic data also raises different ethical questions. The researcher should always think about the possible consequences of the research outcome. There might be cases where the outcome of a research work might causes harm to a specific group of individuals. Since the result of this kind of research can easily be misunderstood, it is very important for the researcher to be aware of with whom the information is shared during the research. Also, after finishing the research, it is crucial to make sure that the result of the work is reflected upon and the information correctly given to the society.

A very common concept in genetic research is biobanking. A biobank is a biorepository that stores biological samples. One of the main issues with biobanks is the secondary non-planned uses of the samples. A sample from an individual was collected for a specific disease some time ago. It is not always very clear to know if it okay to use such samples for genetic tests for other diseases [6].

Working with genetic data is often computationally very demanding. Therefore, one might use different computer infrastructures to gain better performance. The researcher should always check the security of the resources which he or she is using. Another challenging topic now a days is using clouds for storing and analyzing genetic data. While we are concerned about storing genetic data outside our research institute, sometimes the outside cloud providers are significantly more secure than the institute internal network. It is important to check privacy laws of the providing cloud infrastructures. One might also take a step towards encrypting the information to make it more difficult of misuse.
2. Optimization algorithms for QTL searches

Quantitative Local Traits (QTL) are traits which show continuous distribution, such as height. Different statistical methods have been developed for QTL mapping, the standard approach is interval mapping [27]. Permutation testing is commonly used to calculate the significance level of a QTL. In this chapter we give an overview of the statistical models used. We then explain how one can interpret at maximum likelihood as a global optimization problem.

We have developed a new global optimization algorithm, PruneDIRECT, which is a derivative-free optimization algorithm based on DIRECT [35]. We explain how to use PruneDIRECT for QTL mapping instead of exhaustive search. We have released an implementation of PruneDIRECT as an open source R package [4].

We also explain how one can use the map-reduce programming framework for massive parallel permutation testing of QTL points. We have provided this framework as a Software as a Service (SaaS) implementation using Cloud Computing.

Some basic resources for studying QTL analysis are [8, 23, 24, 41, 40, 37]. Articles I, II, III, V and IV are the supporting material for this chapter.

2.1 Basics of QTL analysis

In QTL analysis we usually talk about flanking markers. A flanking marker is an identifiable region of DNA located on either side of a locus.

Below we consider a model which assumes that two interacting QTL dominate the genetic effect on the phenotype. The QTL can be on the same chromosome, or two different chromosomes. For simplicity, we choose a backcross population. Assuming there is a putative QTL at position \((x_{i1}, x_{i2})\) for individual \(i\). Then one can model the phenotypic values as:

\[
y_i = \mu + a_1 x_{i1} + a_2 x_{i2} + bx_{i1}x_{i2} + \epsilon_i
\]

(2.1)

where:

\[
x_{ij} = \begin{cases} 
1, & \text{Aa at location } j \\
0, & \text{aa at location } j
\end{cases}
\]

(2.2)

Since we are looking for a two QTL, we have two additive effects, one for each dimension. \(a_1\) is the additive effect for dimension one, and \(a_2\) is the additive effect for dimension two. \(b\) is the epistatic effect between two locations.
Dense genetic maps are available for many species today because of rapid development of off-the-shelf technology and bioinformatics databases. Sometimes, the genetic map is not dense enough to cover all regions along chromosomes, or they are not fully informative. If it is desired to investigate the possibility of having a QTL between two genetic markers, one have to deal with the missing data problem. Haley-Knott regression (HK) model is a cheap way to approximate interval mapping by least-squares. HK interval mapping can handle missing data [16].

Consider a mesh along the chromosome for a backcross. If a point on the mesh is not an informative marker, then we need to calculate the probability of different genotypes on that point, conditional on informative flanking markers before and after it, i.e. $P_{ij}(Aa)$ and $P_{ij}(aa)$.

In the Haley-Knott model we assume $y_i$’s are normally distributed with mean $(P_{ij}(Aa) + P_{ij}(aa))\mu$ and variance $\sigma^2$ [16]. Then, the phenotype is modeled as:

$$y_i = \mu + a_1 w_{i1} + a_2 w_{i2} + bw_{i1,i2} e_i$$  \hfill (2.3)

where $w_{i1}$ and $w_{i2}$ are the conditional expectation of a QTL at position $x_{i1}$ and $x_{i2}$ for individual $i$ given their flanking markers, and $w_{i1,i2}$ is conditional expectation of a QTL at position $(x_{i1},x_{i2})$ for individual $i$ given its flanking markers. One can write this model in a closed form as:

$$y = A(x)b(x) + e(x)$$  \hfill (2.4)

The QTL positions can now be found by minimizing the residual sum of squares over all $x$’s and $b$’s:

$$RSS_{opt} = \min_{x,b}(y - A(x)b(x))^T (y - A(x)b(x)).$$  \hfill (2.5)

The solution to this minimization problem can be separated into two parts [30]:

$$RSS(x) = \min_b(y - A(x)b(x))^T (y - A(x)b(x)),$$  \hfill (2.6)

and:

$$RSS_{opt} = \min_x(RSS(x)).$$  \hfill (2.7)

Thus, we showed that to find an n-dimensional QTL point, one should find the global minimum in an n-dimensional space with the objective function $RSS_{opt}$. One can interpret a QTL point as the position that explains the genetic variance in a phenotype the most among the position. The common way to find the global optimum in equation 2.7 is exhaustive search in an n-dimensional space. The function evaluation for each point is moderately expensive, which makes the whole search reasonably cheap for $n = 1$, but for $n > 1$ the computational cost quickly becomes prohibitive.
2.2 QTL search

As we explained in the last section, one can consider an n-dimensional QTL search to be an n-dimensional global optimization problem. Derivative-free optimization algorithms can be more efficient in comparison to exhaustive search for high dimensional QTL searchers. One of these derivative-free optimization algorithm is DIRECT [21], which was used for QTL searches in [30]. The main problem with DIRECT is the difficulty to define a stopping criteria. Specially, when one searches over a space of permuted data (for significance tests), the space has many shallow local optimums. We can not be sure that the optimum point found by the algorithm is the global one, before all the grid points have been tested, corresponding to an exhaustive search. Since we apply this algorithm many times for different parameters (permuted phenotypes), even a small error can bias the final result [30].

We now introduce a new derivative free optimization algorithm called PruneDIRECT, based on modifying DIRECT.

2.2.1 Lipschitz optimization

There exist many different derivative-free optimization algorithm [38]. One class of derivative free global optimization algorithm is for Lipschitz continuous functions. A function is Lipschitz continuous if there exists a positive constant $K$ where:

$$|f(x) - f(y)| \leq K|x - y| \text{ for all } x, y \in \text{Domain of the function.} \quad (2.8)$$

In other words, a function is Lipschitz continuous if speed of the growth of the function is limited. If a function is Lipschitz continuous and $K$ is known, then we can use Shubert’s algorithm [39].

Here we give an overview of Shubert’s algorithm and DIRECT based on the treatment found in [21]. Shubert’s algorithm works based on iterative piecewise linear approximation of $f(x)$. One step of Shubert’s algorithm is as follows: Approximate $f(x)$ with two lines, one starts from point $b$ with slope $r-K$:

$$f_1(x) = f(a) - K(x - a) \quad (2.9)$$

and the other one ends at point $b$ with slope $K$:

$$f_2(x) = f(b) + K(x - b). \quad (2.10)$$

In order to evaluate $f_1$ and $f_2$, we needed two functions evaluation, $f(a)$ and $f(b)$. $f_1$ and $f_2$ intersect each other in point $(X(a,b),Y(a,b))$, where:

$$X(a,b) = a + b + \frac{f(a) + f(b)}{2K} \quad (2.11)$$

$$Y(a,b) = \frac{f(a) + f(b)}{2} - K(b - a) \quad (2.12)$$
As an example see figure 2.1(a). In the figure, one can see that \( f(x) \) is approximated with \( f_1 \) and \( f_2 \). The approximated minimum based on one step of Shubert’s algorithm is the point \((X(a,b), Y(a,b))\). For the next step, we split the interval \([a,b]\) into \([a, X(a,b)]\) and \([X(a,b), b]\), and then calculate the approximated minimum for each interval. Now we have three intervals, and we choose the interval which has the smaller minimum to split in the next step. This procedure is illustrated in figure 2.1(a-c). We continue with Shubert’s algorithm until the approximated minimum is within some prespecified tolerance of the current best solution.

\[f(x)\]
\[X(a,b)\]
\[f_1\]
\[f_2\]
\[Y(a,b)\]
\[\text{Slope } -K\]
\[\text{Slope } K\]

\( (a) \)

\[f(x)\]
\[\text{Interval 1}\]
\[\text{Interval 2}\]
\[\text{Interval 3}\]
\[\text{Interval 4}\]

\( (b) \)

\[f(x)\]

\( (c) \)

Figure 2.1. Shubert’s algorithm for Lipschitz continuous functions.
Shubert’s algorithm selects an interval for next splitting based on choosing the minimum of the approximated minimums of all the split intervals. In other words, the algorithm find the smallest $Y(a_j, b_j)$ among all $Y(a_i, b_i)$’s. Each $Y(a_i, b_i)$ is a sum of two terms, $[f(a_i) + f(b_i)]/2$ and $-K(b_i - a_i)$. The first term, $[f(a_i) + f(b_i)]/2$ is smaller when the function values at the endpoints $a_i$ and $b_i$ are small. Therefore, the first term acts as a local search since it selects the intervals where previous function evaluations was good. The second term, $-K(b_i - a_i)$, acts as a global search, since it preferences the longer intervals. Here, the Lipschitz constant $K$ acts as weight on global versus local search.

If $K$ is large, there are more global searchers during Shubert’s algorithm. This causes slow convergence for the method. One could improve the convergence by decreasing $K$ when the algorithm is close to the minimum. Unfortunately this is clearly not possible since the minimum is not known. Another challenge is that Shubert’s algorithm do function evaluation on the end points of the divided intervals. For an $n$ dimensional space, this is equal to $2^n$ function evaluations which is very expansive if $n$ is large. These are some of the motives to introduce an algorithm called DIRECT.

2.2.2 DIRECT

DIRECT algorithm (Dividing RECTangles) is based on modified Shubert’s algorithm. In DIRECT, instead of evaluating the function at endpoints, we evaluate it in the center. Thus, one function evaluation is enough for dividing an interval in an $n$-dimensional space, where Shubert’s algorithm needs $2^n$ function evaluations. Here we approximate the function with a different piecewise linear function than Shubert’s algorithm. Consider a 1-dimensional space. The piecewise linear approximation of $f(x)$ in an interval $[a, b]$ is:

$$f_1(x) = f(c) + K(x - c), \text{ if } x \leq c \quad (2.13)$$

$$f_2(x) = f(c) - K(x - c), \text{ if } x \geq c. \quad (2.14)$$

where $c$ is the center of the interval $[a, b]$, i.e. $c = (a + b)/2$. As an example see figure 2.2.

The approximated minimum for $f(x)$ in interval $[a, b]$ occurs at the endpoints $a$ and $b$, with the value $f(c) - K(b - a)/2$. Here $f(c)$ acts as a local search, $K(b - a)$ acts as a global search, and $K$ is a weighting between global and local search.

So far, we explained DIRECT approximates $f(x)$ with a piecewise linear function in a given interval. DIRECT is based on splitting each interval into three intervals. After splitting, the function value is evaluated at the center points of the left and right intervals, $c \pm (b - a)/3$. For each of these intervals, DIRECT approximates $f(x)$ with a piecewise linear function based on its center. Now the question is which interval to choose for the next splitting.
Figure 2.2. Piecewise linear approximation used in DIRECT.

Figure 2.3. Convex hull when $K$ is known.

DIRECT uses a *convex hull* to choose the next interval for dividing. Note that there are fast algorithm for determining the convex hull [10].

Assume that DIRECT has divided the interval $[a, b]$ into $m$ intervals $[a_1, b_1], [a_2, b_2], ..., [a_m, b_m]$. To make the convex hull, we represent each interval with the length of the interval for the horizontal axes, and the coordinate of the function value in the middle of the interval for the vertical axes. In other words, we represent interval $[a_i, b_i]$ with $(b_i - a_i)/2$ and $f(c_i)$, see figure 2.3. In order to choose the interval for next splitting, we pass a line with slope $K$ below all of the points in plot and we shift it upwards until it touches one of the points. The point that it touches is the next interval for splitting.

If $K$ is large, then as we discussed before, DIRECT spends a lot of time for global searches which makes the convergence very slow. An alternative approach would be to choose all the potentially optimal intervals, which is equivalent to choosing intervals with different values of $K$. Potentially optimal intervals are presented by the points which are at the lower right part of all other points. See figure 2.4 to see the potentially optimum points.
We can formulate choosing potentially optimal intervals in a mathematical way. Assume interval \([a, b]\) is divided into \(m\) intervals \([a_1, b_1], [a_2, b_2], \ldots, [a_m, b_m]\), and \(f_{\text{min}}\) is the best optimum in the current step. Interval \(j\) is potentially optimal if there exist a constant \(\tilde{K}\) such that:

\[
f(c_j) - \tilde{K}\left(\frac{b_j - a_j}{2}\right) \leq f(c_i) - \tilde{K}\left(\frac{b_i - a_i}{2}\right) \quad \text{for all} \quad i = 1, \ldots, m
\]

\[
f(c_j) - \tilde{K}\left(\frac{b_j - a_j}{2}\right) \leq f_{\text{min}} - \varepsilon |f_{\text{min}}|.
\]

The first condition is to make sure that a potentially optimal point is at the lower right part of the cluster of the points. The second condition is to prevent too many local searches. \(\varepsilon\) can be set between \(10^{-7}\) and \(10^{-3}\). DIRECT is not very sensitive to the value of \(\varepsilon\) [21]. If \(\varepsilon\) is constant, scaling the objective function will influence convergence.

For DIRECT, in contrast to Shubert’s algorithm, it is not necessary to know the value of \(K\). In many application we know that the growth of the objective function is bounded, but we do not know the actual value of the bound. For such cases, DIRECT can be very beneficial. If \(K\) is known, then we can use the stopping criteria from Shubert’s algorithm, otherwise we should define some stopping criteria. Some of the stopping criteria in use are mentioned in [21]. We write the formal DIRECT algorithm for finding the minimum of \(f(x)\) in interval \([a, b]\), in algorithm 1.

One can easily extend DIRECT for n-dimensional spaces, for more details check [21].

It is guaranteed that DIRECT finds the global minimum, since if we do not stop the algorithm, it becomes an exhaustive search. DIRECT is sensitive to transformation of the objective function. Speed of convergence of DIRECT can be affected by linear scaling of the objective function [28]. In [28], the
Algorithm 1: DIRECT Algorithm

1. Set the iteration counter $t = 0$, and the interval counter $m = 1$.
2. Set $[a_1, b_1] = [a, b]$, and calculate the center of interval $c_1 = (a_1 + b_1)/2$, and its function value $f(c_1)$.
3. Find the set of potentially optimal intervals, $S$.
4. Select an interval $[a_j, b_j] \in S$.
5. Set $\delta = (b_j - a_j)/3$, the length of the interval, and $c_{m+1} = c_j - \delta$ and $c_{m+2} = c_j + \delta$. Evaluate $f(c_{m+1})$ and $f(c_{m+2})$, and update $f_{\text{min}}$.
6. Make two new intervals $[a_{m+1}, b_{m+1}] = [a_j, a_j + \delta]$ and $[a_{m+2}, b_{m+2}] = [a_j + 2\delta, b_j]$, with centers $c_{m+1}$ and $c_{m+2}$.
7. Modify interval $[a_j, b_j]$ into $[a_j, b_j] = [a_j + \delta, a_j + 2\delta]$.
8. Set $m = m + 2$.
9. Set $S = S - j$. If $S$ is not empty go to step 4.
10. Set $t = t + 1$, if stopping criteria is not satisfied, go to step 3.

The DIRECT algorithm can be modified for symmetric spaces. [15] have suggested one way of doing this. DIRECT can also be modified to handle noisy function optimization [11]. In [11], it was suggested to use a sampling approach where they replicates multiple function evaluations per point and takes an average to reduce functional uncertainty.

2.2.3 DIRECT and QTL search

Because of the nature of genetic recombination and family structures in QTL experiments, different points in the search space are highly correlated. Thus one can use DIRECT to find the global optimum in a QTL search. However, a critical issue is to know when to stop DIRECT and accept the current minimum as the global minimum. A few termination criteria were evaluated for QTL searches in [21].

2.2.4 PruneDIRECT

We have developed a new derivative-free global optimization algorithm called PruneDIRECT, by combing DIRECT and Shubert’s algorithm. The idea is to
remove (prune) parts of the search space which we are sure cannot contain the global optimum. Within DIRECT, after choosing a potentially optimum interval, we split it into three new intervals. If the Lipschitz constant $K$ is known then we check for each interval, let’s say interval $A$, if there exists another interval where the function value in the center of it is smaller than the approximated optimum for interval $A$; if so, then we prune interval $A$. As an example see figure 2.5. The approximated optimum in interval 3 is larger than $f(c_1)$. This means that no point in this interval can be smaller than the already inspected center in interval 3, therefore we can remove it.

As we described, if the Lipschitz constant is known, we can cut parts of the search space in DIRECT, which we are sure which do not contain the optimum point. Again, the problem with DIRECT is that, the algorithm can get close to the basin of an optimum very fast, but it takes longer for it to definitely find the global optimum. With pruning, when DIRECT gets close to the global optimum, many parts of the search space will be pruned, which increases the speed of convergence. More importantly, when we prune parts of the search space, we do not explicitly need any stopping criteria. We can continue with DIRECT until no boxed remain to split. We call this algorithm PruneDIRECT [35].

One problem with Shubert’s algorithm is that if $K$ is large, we have slow convergence. The solution would be to change $K$ but this is not easy. That was one of the reasons to introduce DIRECT. DIRECT does not directly assume any for $K$. We are suggesting to use the value of $K$ in DIRECT to prune some of the intervals.

DIRECT has been used for applications such as QTL mapping when $K$ is unknown [30]. In [35] we explained how to calculate the Lipschitz bound $K$, with and without infinite size population assumption, and use it to prune the search space.
In [35], we introduce a new objective function which we believe it speeds up the convergence of DIRECT algorithm. This new objective function is:

\[ f(x) = -\log(\text{Total variance} - \text{Residual variance at } x) \]  

(2.17)

The reason to choose \( \log \) of the residual variance instead of the residual variance is that the relationship between two points in a chromosome is based on the genetics map which is an exponential function. Based on Haldane’s mapping function, the probability of recombination for a location \( x \) centimorgan further than \( x_0 \) is:

\[ p(x + x_0) = 0.5 + 0.5e^{-2x/100} \]  

(2.18)

By taking the logarithm of the residual variance, we get a linear relation between two points. In article I we explain how one can approximate a bound for \( K \) based on infinite size population theory. For one dimensional space \( K \) is equal to 0.04. Although we know that for small population, the assumption of an infinite size population not valid, we can still use \( K = 0.04 \).

In article I we calculate an approximation of \( K \) for real populations without infinite size population assumption. If \( x \) is a putative QTL, for real populations, we cannot write a deterministic equation explaining how the value of residual variance at \( x + x_0 \) changes in comparison to point \( x \), since these two points are related together via recombinations, which are stochastic events controlled by the linkage map. But we can calculate the distribution of the residual variance at \( x + x_0 \). From this distribution, we can understand how the objective function is behaving at this point. In other words, for some problems with stochastic nature, we cannot calculate \( K \) deterministically, but we can calculate a stochastic bound for it. Mathematically speaking, if we have the residual variance at \( x \), what is the distribution of residual variance at \( x + x_0 \). We showed in [35] that the distribution of residual variance at \( x + x_0 \) is a two-level normal binomially weighted sum of mixture normals. Using this distribution, we can calculate the probability that an approximated optimum in a newly split interval is larger than the center of some other intervals. If this probability is large than \( 1 - \epsilon \), with \( \epsilon \) in the range of \( 1e^{-9} \), then we prune the interval.

It is important to mention that in human population one cannot use DIRECT because of the lack of structure in comparison to experimental crosses.

We have released an R package for PruneDIRECT. The details of this packages, including the implementations, are provided in article II.

### 2.3 Significance test

In order to assess the significance level for a QTL point, it is common practice to permute the phenotype vector, keeping the genotypes fixed [6], [14]. The point of this procedure is to replicate the overall properties of the population,
while removing all genetically correlated contributions to the phenotypes. In article III we explain that around $10^5$ to $10^6$ permutations are needed to get accurate significance level for a QTL point.

2.3.1 PruneDIRECT for permutation testing

PruneDIRECT can be very beneficial for permutation testing. When we calculate the minimum of the objective function for a QTL point, $f_{\text{min}}$, then we can use this $f_{\text{min}}$ to prune the search space for each permutation. If we do this, a majority of runs terminate at the very beginning, since the bound shows that the given space cannot contain a point smaller than the global minimum for the true QTL. If this information is not provided, PruneDIRECT needs more function evaluations, becoming more similar to unmodified DIRECT, even for QTL with large effect.

2.3.2 Parallel frameworks for permutation testing

We have used cloud computing technologies and concepts to develop a new framework for massively parallel permutation testing. The computational demands of the QTL search algorithm grows polynomially with the number of search dimensions. Different efforts have been made both at the level of algorithm as well as adopting different computational models. To address this demand, strategies like static partitioning of the search domain were used to divide the problem into subspaces. Whereas MPI and OpenMP based solutions have also been capitalize for higher dimensional search space [17]. Computational grid resources have also been employed [18, 19] in order to fulfill the computing demands.

Computational models based on parallel and distributed computing significantly improve the performance, but most of the time they require a certain level of expertise in order to run the application. This is lacking in much of the scientific community and makes it difficult for these users to gain maximum benefit out of the proposed algorithms for QTL analysis. Thus, together with meeting the computational demands, it is highly important to computational environment that is efficient, scalable and easily adaptable, within a familiar computing environment.

Based on application characteristics, and considering the targeted community, we focused on maximizing the flexibility while keeping the user working environment intact. For this purpose we have chosen R, a well-known and wide spread working environment for biologists and statisticians.

In permutation testing, each permutation is independent of the others, which makes a case of trivial parallelization. Map-Reduce programming framework is a high level programming framework which can be used for massive parallelization of independent tasks. Apache Hadoop is a software framework
for distributed storage and processing of large data sets using a map-reduce model. Although we are not always dealing with large data sets in QTL analysis, but in close future there will be more large data sets available for such analysis. Also, the users can benefit from resiliency of the ResourceManager in Hadoop. Therefore, it is a good idea to use a framework which supports working with large data sets [1].

RHadoop provides a set of R packages to allow map-reduce programming model within R framework [2]. It is a specific set of APIs based on Apache Hadoop’s version of map-reduce model. RHadoop claims to reduce the code length by one-two orders of magnitude compared to pure Hadoop based Java programs.

RHadoop may not be the most efficient way of running map-reduce programs, yet its simplistic approach to use basic functionality of map-reduce within a familiar environment greatly influence the application’s usability within the scientific community. The level of abstraction provided by RHadoop encourages biologists and statisticians to adopt modern programming paradigms rather than continuing with the monolithic approach of R package developments. In article [32] we explain how to use Hadoop map-reduce framework for permutation testing. We show that our framework scales out almost linearly as the number of virtual CPUs increases.

Another approach is to use map-reduce implemented in Apache Spark. Spark runs the calculation in the memory if the data size is not too big which makes it 10 to 100 times faster than the Hadoop map-reduce framework. In article [32], we explain how to use map-reduce in Spark for permutation testing.

2.4 QTL as a Service: Cloud computing platform for QTL searches

The offering of cloud based infrastructure tremendously influence the computing environment both in industry and academia. Cloud setups can either be private or public. Some of the major public cloud providers are Amazon, Google and Microsoft. Software stacks like OpenStack, CloudStack and OpenNebula can be used for structuring private or community based cloud setups. Apart from the resilience and flexibility in infrastructure management, clouds enable applications to provide their functionality as a service. These services can further be consumed either by the end-users or other services. Together with the elasticity provided by the Cloud infrastructure, applications was never been so powerful before.

All these features provide added value to the application and enhance its usability and extensibility. We have also ported the QTL application on a public cloud. In article IV and V we present cloud-based PruneDIRECT as a service for the analysis of complex traits.
2.5 Outlook

In the previous sections we have explained the idea of pruning a search space and how to calculate an approximation of the Lipschitz constant $K$ for QTL applications. It should be noted that PruneDIRECT can also be used for other applications if $K$ is known or if one can calculate an approximation of $K$ based on understanding the sources of randomness in the problem. This is done by finding the distribution of the objective function at point $x + x_0$ when value of the objective function at point $x$ is known. It is not always easy to analytically calculate this distribution, but it might be possible to define a lower bound for it that is sharp enough for pruning.

One possible approach to calculate this distribution is to use a Monte Carlo approach. However, here it is an issue that we need a very sharp estimate for the pruning. This is equivalent to calculating the extreme tail of the distribution. Many Monte Carlo samples are needed to calculate the tail distribution at point $x + x_0$, which makes it computational impractical. A solution to this problem might be to use importance sampling or to approximate the tail of the distribution [26]. It would be possible to develop a generic method to calculate this distribution, the PruneDIRECT can be used for more applications.

Another potential path for future work is to investigate the possibility of using PruneDIRECT for linear mixed models, since likelihood optimization for linear mixed models are computational very expensive.
3. Sparse matrix techniques for linear mixed models

Linear mixed models are linear models which they contain both fixed effects and random effects. These models are very useful when we have repeated measurements. Fixed effects are the effects which are fixed and often reputable, such as treatment level for a clinical trial. Random effects are the subject-specific effects and they estimate the variability, such as an individual drawn randomly from a population for a clinical trial.

Linear mixed models can be applied for QTL mapping, especially for the analysis of advanced intercross lines (AIL) [36].

Article VI is the supporting material for this chapter.

3.1 Linear mixed models

Consider the linear mixed model:

\[ y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + Z_{n \times q} u_{q \times 1} + e_{n \times 1}, \]  
\[ \text{where } u \sim N(0, A \sigma_u^2), \text{ and } e \sim N(0, I_n \sigma_e^2). \]

In this model, \( y \) is the known vector of observations, \( X \) is the design matrix for fixed effects, \( \beta \) is an unknown vector of fixed effects, \( Z \) is the design matrix for random effects, \( u \) is an unknown vector of random effects, and \( e \) is an unknown vector of random errors.

The estimation of variance components for models with large correlation matrices is computationally expensive. In some applications, such as animal breeding, the correlation matrix is known to be large and its inverse is sparse. In article VI we use the sparsity pattern of the inverse of the correlation matrix to develop a more efficient algorithm for estimating variance components.

One can re-write the linear mixed model in 4.6 as a weighted regression:

\[ y_a = X_a \beta_a + e_a \]

where:

\[ \beta_a = \begin{pmatrix} \beta \\ u \end{pmatrix}, y_a = \begin{pmatrix} y \\ 0 \end{pmatrix}, X_a = \begin{pmatrix} X & Z \\ 0 & A^{-\frac{1}{2}} \end{pmatrix}, \]

\[ W = \begin{pmatrix} \frac{1}{\sigma_e^2} I_n & 0 \\ 0 & \frac{1}{\sigma_u^2} I_n \end{pmatrix} \]
Then one should solve:

\[(X'_aWX_a)\beta_a = X'_ay_a\]  \hspace{1cm} (3.7)

which is equivalent to solving the least square problem:

\[
\min_{\beta_a} ||W^{\frac{1}{2}}(X_a\beta_a - y_a)||_2
\]  \hspace{1cm} (3.8)

The solution of this least square problem is

\[
\beta_a = (X'_aWX_a)^{-1}X'_ay_a,
\]  \hspace{1cm} (3.9)

Estimation of the variance components is based on Lee & Nelder’s iterative algorithm. In this algorithm we need to determine the hat matrix which is:

\[
H_a = X_a(X'_aWX_a)^{-1}X'_aW.
\]  \hspace{1cm} (3.10)

In article VI we use Lanczos algorithm to solve the least squared problem (3.8), and then use the solution to calculate the hat matrix. We show that our method is about 30 times faster than using the direct solver for a population of size 6438. We should emphasise that a faster algorithm is not the only achievement. Working with sparse matrices requires significantly less memory.
4. Computational challenges in modeling maternal effects in psychiatric disorders

4.1 Introduction psychiatric genetics

Psychiatric genetics tries to answer the question of how behavioral and psychological conditions and deviations are inherited [20]. It is known that psychiatric disorders are highly heritable. The heritability is much higher for mental disorders than for somatic diseases such as breast cancer [20].

Today, diagnosing a patient with mental disorder is mainly based on interviewing the patient together with observing the patient and excluding physical disorders as the main reason [3]. With deeper knowledge of psychiatric genetics, we might be able to develop new accurate and time-saving diagnostic procedures in the future, such as diagnosis of mental disorders based on blood samples. Through blood samples, we might then examine arrangements of multiple gene variants and biological pathways that would be linked to different mental disorders. Another benefit would be that it might make it easier to separate different diagnoses which have a common manifestation (syndrome), or to separate between diagnoses within an individual with comorbidity.

In epidemiology we study the causes of health outcomes and diseases in populations. In genetic epidemiology, we focus on how genetic factors and their interactions with other factors increase vulnerability to a disease, or protection against a disease. There are different study designs in genetic epidemiology, including twin studies, family studies and adoption studies.

Family studies are based on the closer you are related to someone, the larger proportion of your genes you share with that person. On average, first degree family members (parents to children) share 50 percent of their genes, second degree 25 percent, and third degree 12.5 percent. Studies show that if a mother has a mental disorder, such as schizophrenia, her offspring has a significantly higher risk of getting the disease compared to offspring from parents with no mental disorder [20]. Adoption studies show that the genetic inheritance from your biological parents significantly increases the risk of getting a mental disorder, regardless of environmental factors [20].

There are different sources of complexity in psychiatric genetic studies. One the main sources of these complexities is the lack of validity of the classification of psychiatric disorders and their diagnosis. As an example, there are problems with validity of the structured interviews used for diagnosing, which might make the data inaccurate. Clinical diagnosis can be very complex, since many mental disorders have common symptoms. For example, diagnosing a
person with bipolar disorder could take many years, since it is the pattern of the mood periods, ups or downs, which decides the type of bipolar disorder and also separates it from a unipolar depression.

It has recently been suggested that there are chromosomal risk regions that might contribute to the development of obsessive compulsive disorder (OCD) [42, 34, 33, 7]. The aim of article VII is to investigate existence of maternal effects in OCD, based on family studies in Swedish registry population. Maternal effects are the effects where the phenotype of an organism is determined not only by the environment and its genotype, but also by the environment and the genotype of its mother. Another definition of maternal effects is that maternal effects are the causal influence of the maternal phenotype on the offspring’s phenotype [44]. Theoretical work has shown that maternal effects can have important and even counterintuitive effects on the response to selection and possibly facilitate the maintenance of additive genetic variation [25, 5]. Article VII is the supporting materials for this chapter. For details of results, we refer to the article.

4.2 Linear mixed model with maternal effects
Maternal effects can be added to a linear mixed model with, for example the following structure:

\[ Y = \text{Covariates(individual)} + \text{Covariates(maternal)} + \text{SumAveE(individual)} + \text{SumAveE(maternal)} + \text{Permanent(mother)} + \text{Residual(individual)} \]  

(4.1)

where:
- \( Y \): Phenotype values.
- \( \text{Covariates(individual)} \): Covariates attribution directly related with the individual (e.g. gender).
- \( \text{Covariates(maternal)} \): Covariates attribution from the mother. It affects the observation on the child but can be attributed to the mother (e.g. smoking).
- \( \text{SumAveE(individual)} \): Sum of the average effects of the individual alleles. Direct additive genetic effects.
- \( \text{SumAveE(maternal)} \): Sum of the average effects from the mother. Additive genetic expressed through the mother.
- \( \text{Permanent(mother)} \): Permanent environment effect, effect of mother on all offsprings, but it is not inherited from parents to their offspring.
We formulate maternal effects mathematically in this way:

\[ y = X\beta + Z_a a + Z_m m + Z_p p + e \] (4.2)

where:

\[ \text{mean}(y) = X\beta, \] (4.3)

\[ \text{var}(y) = (Z_a | Z_m) G (Z_a | Z_m)' + Z_p P Z_p' + R, \] (4.4)

\[ R = I \sigma^2_e, P = I \sigma^2_p \] (4.5)

\[ G = \begin{pmatrix} A \sigma^2_a & A \sigma^2_{am} \\ A \sigma^2_{am} & A \sigma^2_m \end{pmatrix} \] (4.6)

Here \( X \) is the incidence matrix for fixed effects, \( Z_a \) is the incidence matrix for random effects for direct additive genetic, \( Z_m \) is the incidence matrix for random effects for maternal additive genetic, and \( Z_p \) is the incidence matrix for maternal additive permanent effect. \( A \) is the matrix of covariance among the individuals and can be derived from pedigree information. \( I \) is the identity matrix.

Psychiatric disorders tend to be seen as binary traits. Estimation of variance components with binary outcomes, when we do not have repeated measurements, can be problematic [13]. In article VII, we explore a few methods for estimation of variance components, dealing with binary outcomes. One might be able to use sparse matrix techniques used in article VI here.
5. Summary of Attached Papers

The goal of the thesis is to explore and improve some of the advanced modern computational methods in statistics, focused in applications in genetics. This thesis contains three lines of work. The first one is model development for QTL analysis of experimental crosses based on looking at maximum likelihood as a global optimization problem; the supporting materials for this line of work are articles: I, II, III, V and IV.

The second line of work is considering using sparse matrix methods for solving linear mixed models, which can also be used for QTL mapping for advanced intercross; the supporting materials for this is article: VI.

The third line of work is focused on computational challenges of adding maternal effects to linear mixed models; the supporting materials for this is article: VII.

List of Articles:


  **Abstract:** We introduce a new algorithm, PruneDIRECT, for multi-dimensional QTL searches. The idea is to consider maximum likelihood as a global optimization problem and to use application-specific features to improve efficiency and accuracy.

  **Contribution:** The theory for correcting for finite-size populations was developed jointly by this author and Dr. Nettelblad. The transform and the applications underlying this article was suggested by Dr. Nettelblad.


  **Abstract:** We describe how the PruneDIRECT package has been implemented in R and how it can be used. The package is implemented
into different building blocks which gives the users the possibility to re-arrange the blocks and adopt both the search algorithm and the parallelization steps to their needs.

**Contribution:**
PruneDIRECT is re-implemented in R and partially in C by the author.

- **Article III.** A flexible computational framework using R and MapReduce for permutation tests of massive genetic analysis of complex traits. Behrang Mahjani, Salman Toor, Carl Nettelblad, Sverker Holmgren, In IEEE/ACM Transactions on Computational Biology and Bioinformatics, Accepted.

  **Abstract:** We analysis PruneDIRECT more in detail and discuss how to use it for demanding permutation testing using distributed computing and Hadoop.

  **Contribution:**
The parallel framework introduced in this article is developed by the author together with Dr. Toor. The cloud settings are done by Dr. Toor. Design and analysing of the experiments are done by the author.


  **Abstract:** This is an application note which introduces the PruneDIRECT software as a service.

  **Contribution:**
Design and implementation of QTL as a Service is developed by the author together with Dr. Toor. The cloud settings are done by Dr. Toor.

- **Article V.** Software as a service in analysis of quantitative trait loci, Behrang Mahjani, Salman Toor, Supporting material for article IV, 2016

  **Abstract:** We describe the concepts behinds QTL as s Service, how the service is implemented, and how to use it.

  **Contribution:**
Design and implementation of QTL as a Service is developed by the author together with Dr. Toor. The cloud setting are done by Dr. Toor.
• **Article VI.** Fitting Linear Mixed Models using Sparse Matrix Methods and Lanczos factorization, with applications in Genetics, Behrang Mahjani, Lars Lars Rönnegård, Lars Eldén, Submitted.

**Abstract:** We explain how to use sparse matrix techniques to solve Henderson equation and estimate the variance components for large data sets.

**Contribution:** Using Lanczos method for solving linear mixed model is done by the author with help and guidance of Prof. Eldén. Statistical analysis of the model is done by the author together with Prof. Rönnegård.

• **Article VII:** Computational challenges in modeling maternal effects in psychiatric disorders. Behrang Mahjani, Yudi Pawitan, Bert Klei Lambertus, Bernie Devlin, Joseph Buxbaum, Dorothy Grice, Avraham Reichenberg, Sven Sandin, Manuscript, 2016.

**Abstract:** We describe the challenges behind adding maternal effects to linear mixed models with binary traits, when dealing with large data sets.

**Contribution:** Simulation and analysis of data is done by the author with help and guidance of Dr. Sandin, Prof. Pawitan, Dr. Lambertus and Prof. Devlin.
6. Svensk Sammanfattning

De flesta viktiga egenskaper hos människor, djur och växter är kvantitativa, vilket betyder att de är egenskaper som uppvisar en kontinuerlig fenotypfördelning. Positionerna i genomet som beskriver den genetiska uppbyggnaden i en kvantitativ egenskap kallas Quantitative Trait Loci (QTL). Det är känt att både den genetiska sammansättningen och miljöfaktorer påverkar kvantitativa egenskaper. Detta gör det betydligt svårare att upptäcka de genetiska faktornerna bakom sådana egenskaper. Man behöver utveckla och implementera bra statistiska modeller för att fånga upp effekten av både genetiska och miljömässiga faktorer för att hitta QTL. Olika statistiska metoder har utvecklats för kartläggning av QTL [8, 23, 24, 41, 40, 37].

Målet med denna avhandling är att utforska, förbättra och implementera vissa avancerade moderna beräkningsmetoder i statistik, med inriktning på tillämpningar inom genetik. Avhandlingen har tre huvudlinjer.


Vi har använt cloud computing-tekniker och nya koncept för att utveckla ett nytt ramverk för massiva parallella permutationstester [32]. I permuta-
tionstester är varje permutation oberoende av varandra, vilket gör parallelliseringen relativt enkel. I detta fall använder vi Map-Reduce-verktyget vilket är ett ramverk på hög nivå som kan användas för massiv parallellisering av oberoende uppgifter.

Förutom Map-Reduce-programmering finns det många andra viktiga förde-

Den andra delen av avhandlingen fokuserar på att använda glesa matris-
metoder för att lösa linjära blandade modeller med stora korrelationsmatriser. Linjära blandade modeller är linjära modeller som innehåller både fasta effek-

För populationer med kända släktrelationer visar vi att inversen av kovari-
ansmatrisen är gles. Vi beskriver hur man använder den här glesheten för att utveckla en ny metod för att maximera sannolikheten för och beräkna varian-

I den sista delen av avhandlingen studerar vi beräkningsutmaningar inom psykiatrisk genetik utifrån enbart härstamningsinformation. Syftet är att undersöka förekomsten av maternella effekter i tvångssyndrom. Vi lägger till moderns effekter i den linjära blandade modellen som används i den andra delen av denna uppsats, och vi beskriver beräkningsutmaningarna med att arbeta med binära egenskaper.
7. Acknowledgments

I would like to thank my parents who have always supported me and made it possible for me to study. I am truly lucky to have such a caring mom, and a father who has deep passion for science. Thank you my darling love Christina. You stood at my side all these years and patiently supported me through all difficulties. You have also been incredibly helpful to me in editing this thesis.

I would like to express my deepest gratitude to Professor Sverker Holmgren for his supervision and guidance in this thesis. His invaluable support of my ideas and understanding was of enormous help during this research work. I want to acknowledge Dr. Carl Nettelblad for helping me through this work and always being available and supportive. I would also like to thank my other advisor, Professor Lars Rönnergård for his valuable supervision. He has always been supportive, motivating and available to help me.

Many thanks to my good friend and colleague Dr. Salman Toor for teaching me about cloud computing and supporting me until the last minutes of writing this thesis.

At TDB, I would like to thank Dr. Andreas Helander for engaging me in scientific discussions towards cloud computing, Mr. Stefan Pålsson for supporting me through my teaching and Ms. Carina Lindgren who has always been extremely kind and helpful. I want to acknowledge Dr. Tom Smedsaas for being very supportive and welcoming me to the department when I started my PhD. I thank Dr. Emanuel Rubensson for all the interesting discussions that we had these years.

I thank my new colleague Dr. Sven Sandin for introducing me to the amazing world of psychiatric genetics and being always supportive, helpful and positive. It is a true pleasure to work with you. I also thank Professor Yudi Pawitan for supporting me and trying to teach me how to think like a statistician. I would like to acknowledge Dr. Silvelyn Zwanzig. She supported me tremendously with teaching statistics and developing my pedagogical skills all these years. I would like to thank Professor Lars Elden for teaching me numerical methods.

Many thanks to my friend Dr. Markus Kowalewski for the good time we had and many interesting scientific discussions. Thanks to my dear friend Douglas Potter for his friendship, support and encouraging me in sports. My special thanks to all my friends at TDB and my roommate Lina Meinecke for being a nice roommate.

I feel very lucky to know you all. This work would never be possible without all your helps. I am looking forward to learn more from all of you in future.

39
References


[33] Sergi Mas, Patricia Gassó, Astrid Morer, Anna Calvo, Nuria Bargalló, Amalia Lafuente, and Luisa Lázaro. Integrating Genetic, Neuropsychological and


A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)