Statistical testing and selection
by re-sampling
in genome-wide association studies

Zeyu Zhou

orcid.org/0000-0002-4939-5053

School of Mathematics and Statistics
The University of Melbourne

September 2016

Submitted in total fulfillment of the requirements
of the degree of Doctor of Philosophy
Abstract

A Genome Wide Association Study (GWAS) aims to find genetic variants that are associated with a trait of interest. GWAS is a typical “great \(m\), small \(n\)” in which the number \((m)\) of genetic variants being observed in genome is much larger than the number \((n)\) of individuals in the sample. Performing a successful GWAS is challenging both statistically and computationally. This is due to at least the following complications involved in GWAS: confounding stratification in population structure; strong correlations among genetic variants; and analytical difficulty in model selection under ultra-high dimensionality setting. In this thesis we have developed three statistical methods to tackle these complications.

The first method is rank stability selection (RSS), which is developed in Chapter 2. RSS essentially gives a sub-sampling distribution-free implementation for multiple testing, a method widely used in detecting phenotype-associated genetic variants. Classical multiple testing calculates and tests the association with the phenotype for each genetic variant (covariate), mostly based on the asymptotic distribution of the involved test statistic and ignoring the possible stratification confounding in population structure. This may result in inaccurate family-wise control of both types of I and II errors. The RSS method assesses the association significance of each genetic variant based on its stability in ranking in regard to an association test statistic that is computed for all genetic variants and over a number of sub-samples of the data. It has been shown that RSS under weak regularity assumptions gives accurate and robust (against population stratification) detection of significant genetic variants with respect to their associations with the phenotype. The association test statistic involved can be formed using a variety of association measures.
such as correlation, mutual information and that based on regression models, etc. The sub-sampling involved is similar to data permutation but has better family-wise error control and is computationally more efficient. An important regularity assumption for RSS is the test statistics for non-associated genetic covariates are exchangeable. This is weaker than the i.i.d. one used in classical multiple testing or permutation testing.

The second method we have developed and detailed in Chapter 3 is a dimension reduction method which is used to maximally reduce the number of tests in multiple testing without compromising the test power but having the involved family-wise type I error rate (FWER) well controlled. Note that Bonferroni correction is the main method currently used to adjust P-values in multiple testing, which controls FWER in a more likely conservative manner. Our dimension reduction method takes a very different approach. The method adopts the rationale of Independent Components Analysis (ICA), and assumes that the association effects of all SNPs on the phenotype are realised through those from independent components. Each independent component is statistically determined from a haplotype block, i.e. a set of SNPs deemed to fall into a linkage-disequilibrium (LD) region in genome. In doing this way, we are able to remove large amount of redundant information from the data, so that multiple testing based on independent components can achieve the desired power and FWER control. We applied our dimension reduction method to a subset of the iCOGS breast cancer data, where the individuals all have European ancestors and have observations from 210,935 SNPs. We were able to draw out 57,403 independent components (ICs) from the SNPs. We then applied multiple testing to these ICs, resulting in finding 26 extra loci associated with breast cancer apart from those found by standard GWAS and in the literature.
Our third method, detailed in Chapter 4, is motivated by predicting disease risk by the effecting SNPs. This requires the development and use of an advanced variable or model selection criterion scalable to high-dimensional data. In the presence of millions of predictive genetic variants in GWAS, classical model selection criteria such as AIC and BIC would most likely fail to select the correct model but tend to overfit. We propose a novel model selection criterion, called of Stochastic Complexity Criterion correction (SCCc), which is derived by applying the Minimum Description Length (MDL) principle. Under ultra-high dimensional settings, SCCc uses a sparse formulation for non-zero coefficients of the variables in the model and adds an additional penalty term to the criterion that equals to the code length of the sparse formulation. We have shown that SCCc is model selection consistent and by simulation it has a stable FWER in finite samples.
Declaration

This is to certify that

- The thesis comprises only my original work towards the PhD except where indicated in the Preface

- Due acknowledgment has been made in the text to all other material used,

- The thesis is less than 100,000 words in length, exclusive of tables, figures and the Bibliography.

Zeyu Zhou
Preface

This thesis was written under the supervision of Dr. Guoqi Qian and co-supervision of Dr. Minh Bui and Prof. John Hopper.

During Zeyu’s PhD studentship, he also participated in a research group funded by NHMRC grant: “Complex statistical analyses of genome-wide association studies related to breast and prostate cancers using high performance supercomputing” (APP1033452). The group is lead by Prof. John Hopper from Melbourne School of Population and Global Health. The gene ESR1 related pathway dataset used in Chapter 2 and iCOGS breast cancer GWAS dataset used in Chapter 3 are provided by the group. The iCOGS dataset also comes with first seven principal component scores of the genotypes. All further analysis on these datasets is my own work.
This thesis is dedicated to my dear parents and beloved wife.
Acknowledgements

I would like to express my greatest appreciation to my principal supervisor Guoqi Qian for his guidance. Guoqi led me to this exciting area and kept pushing me forward. Without directions from him I would not be able to finish this thesis. I am also grateful to my co-supervisors Minh Bui and John Hopper for their support.

I would also thank Benjamin Goudey, John Wagner, Mani Abedini, Tom Conway, Justin Bedo and other colleagues at the IBM Research, Australia. It’s been a great experience working with and learning from them during my internship.

I would also show my gratitude to Enes Makalic, Daniel Schmidt and team members from John Hopper’s high performance computing GWAS group, Melbourne School of Population and Global Health for the opportunity to working in a multi-discipline research group.

Thanks also goes to Jason Leung, Yuqing Pan and my other friends in the School of Mathematics and Statistics for the mutual support and sharing of stress.

Finally thanks to my parents and my wife for their constant love and support.
Notation

$A, S, N, M$ Denote sets of indices.

$X, Y$ Denote random variables.

$x$ Denotes a random vector.

$A$ Denotes a real valued matrix.

$x$ Denotes a real valued number.

$x$ Denotes a vector of real valued numbers.

$c$ Denotes a real valued constant.

$n$ Denotes sample size.

$m$ Denotes number of covariates.

$p, q$ Denote real number $\in [0, 1]$
# Contents

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiv</td>
</tr>
</tbody>
</table>

## 1 Preliminaries

1.1 Introduction to GWAS                      | 1    |
1.2 Statistical aspects of GWAS               | 7    |
1.2.1 Logistic regression                     | 7    |
1.2.2 Multiplicative Risk Model               | 9    |
1.2.3 Multiple hypothesis testing             | 10   |
1.2.4 Linkage Disequilibrium                  | 14   |
1.2.5 Hardy-Weinberg equilibrium              | 15   |
1.2.6 Population stratification               | 17   |
1.3 Genetic communication models              | 20   |

## 2 Rank Stability Selection through Sample Splitting

2.1 Introduction                              | 23   |
2.2 Variable selection                        | 28   |
2.2.1 Extension to interactive effects       | 29   |
2.3 Rank Stability Selection                  | 30   |
2.3.1 A rank test for exchangeability        | 31   |
Contents

3.6.1 Simulation on haplotype block under the ECC hypothesis 94
3.6.2 Simulation under the casual SNPs hypothesis . . . . . . 95
3.7 Discussions . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
3.8 Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102

4 Stochastic Complexity Criterion Correction for Ultra High Dimensions 104
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
4.2 Variable selection and model selection criteria . . . . . . . . 108
  4.2.1 Stochastic complexity and two part codes . . . . . . . . 109
  4.2.2 Stochastic complexity corrected for sparsity . . . . . . . 110
  4.2.3 Stochastic Complexity corrected under the null model . . 114
  4.2.4 Asymptotic properties of SCCc . . . . . . . . . . . . . . 116
4.3 Numerical examples . . . . . . . . . . . . . . . . . . . . . . . 118
  4.3.1 Simulation setup . . . . . . . . . . . . . . . . . . . . . . 119
  4.3.2 Computational time . . . . . . . . . . . . . . . . . . . . 120
  4.3.3 Exhaustive search with small $m$ . . . . . . . . . . . . . 120
  4.3.4 LASSO search with large $m$ . . . . . . . . . . . . . . . 123
4.4 Real data example . . . . . . . . . . . . . . . . . . . . . . . . 126
4.5 Conclusion and discussions . . . . . . . . . . . . . . . . . . . 128
4.6 Supplementary . . . . . . . . . . . . . . . . . . . . . . . . . . 130
  4.6.1 A MDL approach of EBIC . . . . . . . . . . . . . . . . . 130
  4.6.2 Extended BIC conditions . . . . . . . . . . . . . . . . . 132
  4.6.3 Rare genetic variants . . . . . . . . . . . . . . . . . . . . 133

5 Conclusion 137

Bibliography 139
List of Figures

1.1 Genome Sequencing Cost[124] ........................................... 3
1.2 A Manhattan plot [43] ...................................................... 7
1.3 Shannon and Weaver’s communication model. .................. 20
1.4 Channel coding communication model. ......................... 22

2.1 FWER of multiple testing methods ................................. 43
2.2 PSR of multiple testing methods .................................. 44
2.3 Multiple testing methods results using Mutual Information .. 45
2.4 LD structure .............................................................. 49
2.5 FDR of multiple testing methods ................................. 56

3.1 Manhattan Plot of iCOGS breast cancer dataset .............. 63
3.2 GWAS communication model. ................................. 67
3.3 Summary of haplotype blocks and independent components . 81
3.4 Population structure of iCOGS breast cancer dataset ....... 82
3.5 QQ-plot of SNP P-values ........................................... 88
3.6 QQ-plot of IC P-values ............................................. 89
3.7 Manhattan plot by chromosomes ................................. 91
3.8 Comparison of findings between methods .................... 92
3.9 Comparison of findings with literature findings .............. 93

4.1 FDR under the null model structure ............................... 123
4.2 Power analysis .................................................. 124
4.3 LASSO regularization path of the breast cancer dataset .......................... 126
4.4 FDR under the null model structure with rare variants .............................. 135
4.5 Power analysis for rare variants ................................................. 136
List of Tables

1.1 Contingency table of multiple testing results. .......................... 12
2.1 5 simulation models .......................................................... 42
2.2 Results on top 20 SNPs among 366 SNPs ............................... 50
2.3 Simulation results for Model 1 ............................................. 57
2.4 Simulation results for Model 2 ............................................. 57
2.5 Simulation results for Model 3 ............................................. 58
2.6 Simulation results for Model 4 ............................................. 58
2.7 Simulation results for Model 5 ............................................. 59
3.1 24 additional haplotype blocks found significant using ecc ...... 85
3.2 Associated loci by methods ............................................... 90
3.3 Experiment on haplotype block with ECC hypothesis(common variant) .................................................. 96
3.4 Experiment on haplotype block under ECC hypothesis(rare variant) .................................................. 97
3.5 Experiment on chromosome 6 under casual SNP hypothesis ...... 98
3.6 Hardy-Weinberg Equilibrium test result ................................. 103
4.1 Exhaustive search result for $S_t = S_0$ ................................. 121
4.2 Exhaustive search result for $S_t = \{1, 2\}$ ............................ 122
4.3 Best model selected in the breast cancer dataset .................... 127
4.4 Exact log likelihood regret $\delta$ simulation ............................ 134
4.5 Type I error comparison with different allele frequency (AF), $S_r = S_0 135$
Chapter 1

Preliminaries

Genome-Wide Association Study (GWAS) is a widely used epidemiology device which has delivered great results in the last decade. In this chapter, we will briefly introduce its background, related technologies, achievements and recent developments. We will then give a short review of the principal statistical methods i.e. logistic regression and multiple testing. In addition, we will also discuss some statistical aspects of GWAS such as Linkage Disequilibrium (LD), Hardy Weinberg Equilibrium and population stratification which will be used to design our statistical methods in the lateral chapters. At last, we will give a brief summary of related work in genetic information theory. DNA sequences are treated as source codes transmitted through generations. Modelling information flowing within the biological systems will help us develop more efficient methods for association analysis.

1.1 Introduction to GWAS

A Genome-Wide Association Study (GWAS) investigates possible associations between genetic variants and a trait in a population. The statistical associations revealed by GWAS may predict the risks of having genetic diseases, therefore lead to biological discoveries of the disease and development of new drugs. GWAS has not only been applied to human but also in agriculture (plants and animals) and bacterium. In this thesis, we will focus on the
1.1. Introduction to GWAS

application of GWAS to human diseases.

In the association analysis, we seek to predict the response variable with the aid of covariates. In human GWAS, the response variable is usually a phenotype, i.e. one characterization of individuals in a population. It can be either numerical such as height and blood pressure, or categorical such as eye colour and blood type. The most widely used and simplest type of phenotype is binary. It is also referred to as case-control study design. In such a design, phenotype denotes whether an individual has a certain trait, for example a genetic disease.

In GWAS, the covariates are the genetic variants. They are also called explanatory variables or features. Genetic variants are defined with respect to a population. The challenge of GWAS lies in the enormous size of the human genome. The human genome has roughly 3,300,000,000 pairs of nucleotide. Fortunately, most of them are identical among individuals. This leaves us 52,126,039 short variants in the human genome including Single-Nucleotide Polymorphisms (SNPs, read “snips”), indels and somatic mutations. The states of these genetic variants are also called genotypes. In short, a phenotype can be regarded as the result of interaction between the genotypes and the environment, and GWAS aims to find the links between the phenotype and the genotypes.

Genotyping, which means obtaining the state of a genetic variant in an individual used to be extremely difficult, slow and expensive. Genotyping first requires a map of the human genome as the reference; then it locates the variants on the map and read the states of those variants. The first challenge was to get a detailed map of the human genome. The Human Genome
1.1. Introduction to GWAS

Project (HGP) is the first attempt to map the whole human genome. It started in 1990 and took 13 years, twenty universities and research institutions around the world and cost over 3 billion US dollars. The second challenge was to identify those genetic variants on the genome. HGP sequenced just a few individuals. It was followed by the Hapmap Project (269 individuals) and 1000 Genome Project (1092 individuals). Hapmap started in 2002, with Phase I finished in 2005 [19], Phase II in 2007[35] and Phase III data released in 2010[20]. 1000 Genome, which launched in 2008 finished sequencing of 1092 genomes in 2012[17] and finalized in 2015[18]. It took less time to finish one project and sequenced more individuals. Actually, the cost of sequencing a genome is going down tremendously from 100 million US dollars in 2001 to around 1000 US dollars in 2015(Figure 1.1) [124] thanks to the next generation sequencing (NGS) technologies.

![Cost per Genome](https://www.genome.gov/11006943)

**Figure 1.1:** Genome Sequencing Cost[124]

---

NGS is a collection of new technologies featuring massive parallel sequencing. The first generation of DNA sequencing is referred to as the capillary-based semi-automated Sanger sequencing [57]. It is capable of generating 1000-basepair-long reads with 99.999% per base accuracy[108]. Comparing to the first generation sequencing, NGS generates shorter reads with higher error rate. Thanks to the new bioinformatics algorithms and higher coverage rate NGS is faster, cheaper and as reliable as the first generation sequencing[83, 108].

Once the loci of genetic variants are known, SNP microarray chips are designed and made to genotype those variants for an individual simultaneously across the genome at a low cost; a typical SNP array chip can genotype up to one million SNPs with > 99% accuracy[66]. This technology makes the genotyping the whole genome of hundreds and thousands of individuals feasible and affordable.

In GWAS, genetic variants are usually specifically referred to as Single-Nucleotide Polymorphisms (SNPs). A SNP is at a specific locus on the genome. The base pair of nucleotides (AT,CG) at this locus varies between individuals in the population. Since human is diploid, at a locus, four possible combination of pairs could occur,i.e. (AT,AT),(AT,CG),(CG,AT),(CG,CG). We assume the order of the pair has no effect and it leaves us with three states. Then each SNP can take value \{0,1,2\}. The value denotes to the number of minor allele(base pair) at the locus with respect to the population.

The statistical aim of GWAS is to identify which SNPs are associated with the phenotype. In a typical GWAS, statistical hypothesis testing is used to examine the association between each SNP and the phenotype. Then an error control procedure called multiple testing evaluates the significance of all SNPs across the genome. Only SNPs believed to have true association with the
phenotype are kept and reported.

The first GWAS was conducted in 2005 [48]. Since then, GWAS as a quantitative analysis tool has been widely accepted which leads to a number of important discoveries in human disease study such as diabetes, height and various cancers etc [54, 118] In recent years, there has been a clear shift in GWAS from Mendelian Disorders to more complex human diseases [74, 78, 112]. Mendelian disorders are diseases related to only one locus on the genome. Complex diseases may be caused not only by interactions between multiple loci on the genome but also interactions between the genotypes and environments.

When dealing with a complex disease, standard GWAS procedure is likely to be inadequate. It often fails to address a large proportion of trait heritability [29, 74, 118]. One explanation of the missing heritability is caused by rare variants. Rare variants referred to as the single nucleotide variations have Minor Allele Frequency (MAF) less than 0.05 within the population. At the time of the Hapmap project was carried out, only few hundreds of genomes were sequenced. Rare variants were not identifiable. Even though technology has progressed, rare variants still face statistical challenges such as low power [6, 30, 44, 104].

Another explanation is that disease heritability may better be accounted for through more complex models, such as models with interaction-effect components [21, 73, 117, 122]. The interaction effects in genetics are also referred to as epistasis. They could explain some higher order interactions between genetic variants associated to the phenotype [75]. This would be undetectable through marginal effect screening (standard multiple testing). However, the computational demand increases exponentially when the level of interaction considered increases. Pairwise interaction is hard and 3-way
interaction is nearly infeasible\cite{46, 97}.

Larger sample size is essential to solve the puzzle, disregarding the causes of missing heritability\cite{5}. Large sample size would reveal weak association signals from both loci with rare alleles and interactions between loci. To archive this, large consortia were set up to collect data globally with the same standards. For example the Collaborative Oncological Gene-environment Study (COGS)\cite{16} has collected more than 250,000 individuals to study hormone-related cancers.

Large datasets bring new challenges. Firstly, they are usually comprised of several small datasets, collected by sub-studies. This would cause multiple sub-populations and lead to a problem called population stratification. Secondly, a large sample size may also lead to more covariates. For example, detecting and genotyping rare variants would be feasible. This brings substantial burden on statistical analysis as the problem becomes more difficult when the number of covariates increases. Novel statistical methods will be needed to cope with this new trend.
1.2 Statistical aspects of GWAS

A standard GWAS procedure basically consists of three steps: quality control, correction for confounding factors and multiple testing. The final result is usually presented in a Manhattan plot (Figure 1.2). The x-axis represents the whole genome by connecting all 23 chromosomes. The y-axis is the $-\log_{10} P$-values for association testing of the corresponding SNPs on the genome. The higher the point, the more significant the association of the SNP is. Manhattan plot gives an intuitive illustration of the positions of the disease related regions on the genome and their signal significances. It is also convenient to compare the results of two GWAS on the phenotype using Manhattan plots. In this section, we will introduce the statistical aspects of the standard GWAS procedure. Those concepts will be used throughout the thesis.

1.2.1 Logistic regression

Logistic regression model is one of the most popular statistical models used in association analysis of case-control studies in GWAS. First of all, it belongs to the generalized linear model families[80]. It uses maximum likelihood
estimator to fit the model and have inference tools such as Wald test for the
coefficient, likelihood ratio test, etc. Secondly, the coefficient of the SNP
in the logistic regression model is the log odds ratio of the cases versus the
controls. This directly shows the degree of impact of the number of alleles to
the phenotype. The corresponding odds ratio is also called “effect size” in the
literature.

Let $Y$ be a discrete response variable taking integer values from either
0 or 1, and $x = (X_1, \cdots, X_m)^\top$ a vector of covariates i.e. SNPs. A logistic
regression model predicts $Y$ given $x$. It has three components.

- A Bernoulli distribution $\text{Bernoulli}(\rho)$ with $\rho \in [0, 1]$. We assume $Y$
  conditioned on $x$ is Bernoulli distributed.

- A linear predictor $\eta = \beta_0 + x^\top \beta$ where $\beta = (\beta_1, \cdots, \beta_m)^\top$
  is an vector of coefficients of interest.

- A function $h(\eta) = \frac{e^\eta}{1+e^\eta}$. $\eta$ transforms the linear predictor $\eta$
in to the expected value of $Y$ i.e. $E(Y)$. Its inverse function, $g(\eta) = \log \left( \frac{\eta}{1-\eta} \right)$, is
called the logistic link function.

Together, we have

$$\logit E(Y|x) = \log \left( \frac{\Pr(Y = 1|x)}{\Pr(Y = 0|x)} \right) = \beta_0 + \beta^\top x \quad (1.1)$$

In GWAS, the most important use of logistic regression is to estimate
the effect size of a genetic variant. Let $\{(y_i, x_i^\top) : i = 1, \ldots, n\}$ be a sequence
of independent observations of $Y$ and $x$. Denote $y = (y_1, \cdots, y_n)^\top$
as the response vector and $X = (x_1, \cdots, x_n)^\top$ as the design matrix. We can use the
maximum likelihood estimator (MLE) $\hat{\beta}$. The log-likelihood function for $\beta$ is
1.2. Statistical aspects of GWAS

\[ \ell(\beta | y, X) = \sum_{i=1}^{n} \{ y_i \log (h(\beta_0 + x_i^T \beta)) + (1 - y_i) \log(1 - (h(\beta_0 + x_i^T \beta))) \} \]

(1.2)

Then we have the maximum likelihood estimator \( \hat{\beta} = \arg \max_{\beta} \ell(\beta^* | y, X) \).

Qian shows the estimation error \( ||\hat{\beta} - \beta|| \) is of order \( O \left( \sqrt{\frac{\log \log n}{n}} \right) \) under the Euclidean norm[94].

1.2.2 Multiplicative Risk Model

SNP is statistically defined as a categorical variable with three states. With no assumption, in order to fit a SNP into a logistic regression model, we will have to introduce two dummy variables to denote the two non-zero states of a SNP. However, SNP can also be present as the count for minor alleles on its corresponding locus. The common practice in GWAS is to treat SNPs as ordinal variables and treat them as integer values variables in the logistic regression models. This is also referred as to the multiplicative risk model.

The risk of having disease \( Y \) within a population can be numerically measured by odds, i.e.

\[ \text{Odds}(Y) = \frac{\Pr(Y = 1)}{\Pr(Y = 0)}. \]

A SNP \( X \) takes value from 0, 1, 2. The baseline risk of \( Y \) can be defined by the odds of \( Y \) when allele “a” is not present, i.e. \( \text{Odds}(Y | X = 0) \). Odds ratio measure how many folds of risk will increase when a risk factor is present compare to the baseline risk, i.e.

\[ \text{OR}(Y; Aa) = \text{OR}(Y; X = 1) = \frac{\text{Odds}(Y | X = 1)}{\text{Odds}(Y | X = 0)}. \]

(1.3)

The multiplicative risk model assumes that the odds ratio of having multiple risk factors is the product of the odds ratio of themselves. If having allele
“a” is a risk factor, then having “aa” at one locus would have the squared odds ratio, i.e.

\[ \text{OR}(Y; aa) = \text{OR}(Y; X = 2) = (\text{OR}(Y; X = 1))^2 \]  \hspace{1cm} (1.4)

If we use a logistic regression to model the odds, and treat \( X \) as a categorical variable, we have

\[ \log(\text{Odds}(Y|X)) = \beta_0 + \beta_1 1(X = 1) + \beta_2 1(X = 2). \]  \hspace{1cm} (1.5)

Then the odds ratio would be

\[ \log(\text{OR}(Y; Aa)) = \log(\text{Odds}(Y|X = 1)) - \log(\text{Odds}(Y|X = 0)) = \beta_1, \]
\[ \log(\text{OR}(Y; aa)) = \log(\text{Odds}(Y|X = 2)) - \log(\text{Odds}(Y|X = 0)) = \beta_2. \]

Assuming Multiplicativ risk and according to (1.4), we have \( \beta_2 = 2\beta_1. \)

Therefore, (1.5) can be rewrited as

\[ \log(\text{Odds}(Y|X)) = \beta_0 + \beta_1 X. \]  \hspace{1cm} (1.6)

This suggests that in a logistic model with multiplicative risk model the log odds scales linearly with respect to \( X \).

### 1.2.3 Multiple hypothesis testing

Multiple testing is the core technique of the association investigation in GWAS. It performs hypothesis test for each genetic variant with the phenotype. This is also called search for marginal association effects. Let \( X_1, X_2, \ldots, X_m \) be random variable corresponding to SNPs and \( Y \) denotes the response variable(the phenotype). To test the association between \( X_i \) and \( Y \) individually, we construct the following null hypotheses for \( i = 1, \ldots, m \),

\[ H_i : X_i \text{ is not associated with } Y, \text{ for } i = 1, \ldots, m. \]  \hspace{1cm} (1.7)
We use \( H_i = 1 \) to denote \( H_i \) is true, and \( H_i = 0 \) otherwise. Since the underlying association can be in any form, it is typical to restrict the study to one specific association test statistic. For example, \( \chi^2 \) test and Wald test for a logistic regression model are two of the most popular tests. \( \chi^2 \) test examines independence between the phenotype and the SNP is based on their contingency table. It is non-parametric and fast in calculation. Wald test estimates the increment of disease risk in the presence of the minor allele of the SNP.

**Hypothesis testing** For each hypothesis \( H_i \), let \( \phi_i \) denotes the corresponding test statistic. We want to reject the null hypothesis based on the value of \( \phi_i \). Here we assume \( \phi_i \) is continuous. Type I error \( \alpha \) is referred to as the chance of rejecting \( H_i \) when \( H_i \) is true. Let \( \Gamma_i(\alpha) \) denote the rejection region of \( H_i \) with respect to type I error rate \( \alpha \) so that

\[
\Pr(\phi_i \in \Gamma_i(\alpha) \mid H_i = 1) = \alpha
\]

The decision of rejection \( H_i \) is based on P-value of \( \phi_i \). Suppose we observe \( \phi_i = t_i \), then we have

\[
P\text{-value}(\phi_i = t_i) = \min_{\{\Gamma_i\}} \Pr(\phi_i \in \Gamma_i \mid H_i = 1, t_i \in \Gamma_i).
\]

In other words, P-value is the minimum type I error rate of all possible reject regions that would reject the null hypothesis given \( \phi_i = t_i \). We reject the null hypothesis \( H_i \) if P-value\( (t_i) < \alpha \). This P-value of \( H_i \) is referred to as raw P-value \( p_i \) in multiple testing.

**Multiple testing** Let there be \( m \) simultaneous hypothesis tests. Table 1.1 summarizes the outcome of multiple testing. And we have

\[
\text{FWER} = \Pr(V > 0)
\]
Family-wise error rate (FWER) is used in GWAS to access the overall error rate in multiple testing. It is the probability of rejecting at least one null hypothesis while they are true [123]. In order to control FWER, raw p-values can no long be used in decision making. Let $\tilde{p}_i$ be the adjusted p-value for FWER. If we reject $H_i$ when $\tilde{p}_i > \alpha$, then we shall have FWER = $\alpha$ for all hypotheses. Bonferroni correction adjusts raw p-values as,

$$
\tilde{p}_i = \min(m \cdot p_i, 1).
$$

(1.11)

It is a conservative lower bound derived from Bool’s inequality and it has no assumption of independence between hypotheses. If the raw p-values are i.i.d of uniform distribution over $[0, 1]$, then Sidak correction gives

$$
\tilde{p}_i = 1 - (1 - p_i)^m.
$$

(1.12)

Another popular error measurement is False Discovery Rate (FDR) [9]. Let $\tau = \frac{v}{v + s}$ when $v + s > 0$ and $\tau = 0$ when $v + s = 0$, then

$$
\text{FDR} = E\tau
$$

(1.13)

FDR is less strict than FWER and has higher power given the same rate [130]. Since it focus on the ratio of the false positives over all positives, it is more flexible within multi-stages variable selection frameworks. Nevertheless, the mainstream research still uses FWER as a hard threshold for publishing GWAS results [56, 85]

<table>
<thead>
<tr>
<th></th>
<th>Failed to reject</th>
<th>Rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>True null hypotheses</td>
<td>$u$</td>
<td>$v$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Non-true null hypotheses</td>
<td>$t$</td>
<td>$s$</td>
<td>$m - m_0$</td>
</tr>
<tr>
<td></td>
<td>$m - r$</td>
<td>$r$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

Table 1.1: Contingency table of multiple testing results.
One issue with multiple testing in GWAS is how to estimate the distribution of test statistics under the null hypothesis. Most hypothesis tests have analytic asymptotic approximation under the null hypothesis. They work fairly well in a single hypothesis test at $\alpha = 0.05$. However, in multiple testing, Bonferroni correction requires us to make decisions on whether the raw p-value is smaller than $\alpha/m$. When $m$ is large, we are looking at the very tail of the distribution. The asymptotic approximation would be less accurate with a finite sample size. Moreover, some tests are based on specific biological assumptions, which is hard to verify or test against. This reduces the accuracy of asymptotic approximation. At last, population stratification is a common issue in GWAS. It is regarded as a confounding factor and may inflate the FWER.

**Permutation test** As an alternative of using the asymptotic distribution, permutation test is usually regarded as a golden standard as it is both solid and expensive. In GWAS, a permutation test permutes the labels of cases and controls to simulate null hypothesis. The permutation p-value $p^*_i$ is calculated by permuting the labels for $b$ times. In one tail test case, we have

$$p^*_i = \frac{\left|\{b : t_{i,b} \geq t_i\}\right|}{b} \text{ for } i = 1, \ldots, m.$$  

$t_i$ is the value of $\phi_i$ given data, $t_{i,b}$ is the value of $\phi_i$ in $b$th permutation. A permutation test is a simple technique and precisely estimates the null hypothesis distribution for the specific data. It is an automatic fix for inaccurate asymptotic estimation and break of biological assumptions. However, it is very computational expensive, $b$ permutations are necessary to reach $1/b$ P-value precision. Usually, $b = 1000$ is recommended for $\alpha = 0.05$. In multiple testings, we need high precision of p-value for adjustment. This would further increase the computational burden. Fortunately, smart techniques such as minP and maxT allow us to calculate the adjusted p-values with the same
number of permutations as a single hypothesis test [123]. Ge gives detailed review on those algorithms[42].

### 1.2.4 Linkage Disequilibrium

As discussed, most of the genome is identical among human population. The rest, genetic variants, are strongly correlated as well. When the alleles at multiple loci are associated with each other, they are in linkage disequilibrium (LD). The underlying mechanism causing LD is yet to be determined. However, so far we know that genetic linkage, natural selection, genetic drifting and population subdivision have significant influences on LD [111].

The above associations can be measured statistically. $D'$, $r^2$ are two widely accepted measurements[2]. Both take values in $[0, 1]$, where 0 is in linkage equilibrium and 1 is in “complete” linkage disequilibrium. They are based on the same statistic measure $D$. Let $p_A$ and $p_B$ be the allele frequency of “A” and “B” instead of “a” and “b” at two loci respectively. A haplotype is a combination of SNPs which tends to be inherited together through generations\(^c\). Denote $p_{AB}$ as the frequency of corresponding haplotype having allele “AB”. Then we have

$$D = p_{AB} - p_A \times p_A \quad (1.15)$$

$$D' = \frac{D}{D_{\text{max}}} \quad (1.16)$$

$$r^2 = \frac{D^2}{p_A(1 - p_A)p_B(1 - p_B)} \quad (1.17)$$

where $D_{\text{max}} = \min(p_A(1 - p_B), p_B(1 - p_A))$.

In human genome, continuous regions of SNPs in strong LD to each other were found throughout the genome. The region size usually lies in the range

from 0 to 50k base pairs. They are referred to as haplotype blocks. The precise definition of haplotype block varies from studies to studies. Yet, there are two main categories. The first is defined as a region with a few haplotypes. The second is defined by some pairwise LD measure, such as D-prime or $r^2$ [119]. In this chapter, we will adapt the definition based on D prime [39].

Haplotype block is important to GWAS for the following reasons. Firstly, assume that a SNP is associated with the phenotype, but this SNP is not necessary observed i.e. the SNP is on the genotyping array. Thanks to LD structures and haplotype blocks, we found another SNP closely related that SNP within in population from other studies. If they are in strong LD i.e. highly related, then we can still capture the association signal through the observed SNP. The observed SNP is also called tag SNP. With tag SNPs, we can still pin point a small region on the genome that is associated with the phenotype. Therefore, as long as we have a genotyping array chip with enough density and coverage over the genome, we can perform GWAS with less than 3 million genetic variants but still be able to find genetic associations. Secondly, when two GWAS are conducted on different genotyping arrays i.e. different set of SNPs are genotyped, then we do imputation with the information of haplotype blocks, and estimate the missing SNPs on both chips with certain confidence [76]. Furthermore, in this thesis we perform dimension reduction on the genotypes based on haplotypes (Section 3.1). This allows us to increase the multiple testing power.

1.2.5 Hardy-Weinberg equilibrium

Hardy-Weinberg Equilibrium(HWE) is an important component in the GWAS quality control and quality assurance(QC/QA). QC includes checking the accuracy and completeness of SNP genotyping such as checking the allelic
probe intensities and statistical tests for genetic assumptions[68]. HWE is derived from classic population genetic models and breaking that equilibrium usually suggests genotyping artifacts[68, 115]. HWE states that the allele frequency remains constant from generations if there is free mating[49]. Let “A” and “a” denote two possible alleles at a locus on one chromosome. Let the frequency of A be \( p \), the frequency of a be \( q = 1 - p \). Without loss of generality, we assume a is the minor allele i.e. \( p > q \), and \( q \) is called the minor allele frequency (MAF). Since human is diploid, there are three possible alleles, “AA”, “Aa”, and “aa”. Under random mating, the distribution of the genotype at the next generation would be

\[
\Pr(AA) = p^2, \quad \Pr(Aa) = 2pq, \quad \Pr(aa) = q^2.
\]  

(1.18)

And the allele frequency of “A” in the next generation is \( \Pr(A) = \frac{1}{2}(p^2 + 2pq) = \frac{1}{2}(2p^2 + 2p(1 - p)) = p. \) Therefore, HWE holds. A HWE test tests whether mating process is independent or not, popular tests are Pearson’s \( \chi^2 \) test and Fisher’s exact test. [34]

Another benefit of testing for HWE is to reduce the degrees of freedom in modelling SNP values to 1. Recall a SNP takes value from \( \{0, 1, 2\} \). Under HWE, a SNP can be modeled by a binomial distribution with 2 trials and the successful probability as the minor allele frequency \( q \). On the contrary, if HWE does not hold, we would need two dummies variables to model the distribution of the SNP. Checking HWE not only detects artefacts but also simplifies the calculation and interpretation of the subsequent association analysis.
1.2.6 Population stratification

In association analysis, we test for the difference in allele frequencies of a SNP between cases and controls. However, such difference would be influenced by systematic differences in ancestries. This possible confounding factor is referred to as population stratification[13]. This problem is severe and persistent in large GWAS studies consisting of several sub-studies. When population structure presents, the allele frequency of a SNP would vary between sub-populations. If the population structure is correlated to phenotype, then all SNPs with different allele frequencies between sub-population would appear to be correlated to the phenotype as well despite their true relationship with it. When doing multiple testing on all the SNPs, the tail of the P-values distribution would inflate if population structure is a confounding factor. This would cause substantial amount of unexpected false positives[36]. Detecting and correcting population stratification is important and necessary.

Genomic Control  Genomic Control(GC)[26] is a simple technique that evaluates and corrects P-values for population stratification. It is originally designed for Armitage’s trend test[3]. Under the null hypothesis, Armitage’s trend test statistic follows $\chi^2_{df=1}$. GC calculates an inflation parameter $\lambda_{gc}$. Let $\xi_1, \ldots, \xi_p$ be the test statistic of $p$ SNPs, and $Q_\gamma$ denote the $\gamma$ quantile. Then we have

$$\lambda_{gc} = \frac{Q_{0.5}\xi_{df=1}^2}{\text{Median}\{\xi_1, \ldots, \xi_p\}}.$$  \hspace{1cm} (1.19)

when the population stratification present. The values of $\xi$ would be inflated due to population structure as confounding factors. So does the median of $\xi$. $\lambda_{gc}$ is simply the inflation ratio of the observed median of the $\chi^2$ statistic. If $\lambda_{gc} = 1$, then there is no population stratification. Usually $\lambda_{gc} > 1.04$ is
considered the evidence of possible population stratification[90]. To adjust
test scores for population stratification, we can just divide the raw test statistic
values by $\lambda_{gc}$ and recalculate their P-values according to $\frac{\chi^2}{df} = 1$.

Correcting population structure would either compromise the power of the
analysis or incur an expensive calculation of a kinship matrix of the sample
[25, 64, 88, 90]. And these correction methods only fits classical techniques
such as Pearson’s $\chi^2$ test and logistic regression. They are not suitable for
many novel machine learning techniques.

**Principal Component Analysis**  Principal Component Analysis (PCA) is a
statistical procedure that decorrelates a set of variables. Basically, it finds
an orthogonal projection that maps the variables to a new linear space so
that the projected variables are uncorrelated. To perform PCA, we first need
the covariance matrix and then do a singular value decomposition on it. Let
vector $x = (X_1, \cdots, X_P)$ denotes all the SNPs on the genome. Then, the
covariance matrix of $x$ is

$$Q_x = E\left((x - E(x))(x - E(x))^\top\right).$$ (1.20)

Note that $Q_x$ is positive definite. PCA performs singular value decomposition
on it, i.e.

$$Q_x = W\Lambda W^\top,$$ (1.21)

where $W$ is a orthogonal matrix with $W^\top W = I$, and $\Lambda$ is a diagonal
matrix filled with the eigenvalues of $Q_x$, $\lambda_1, \ldots, \lambda_m$. $\Lambda$ is ordered so that
$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. The $k$th column of $W$ is the $k$th eigenvector of
$Q_x$ corresponding to eigenvalue $\lambda_k$. The columns of $W$ are referred to
as principal components in PCA. The columns of $Wx$ are called principal
component scores of $x$. 

18
1.2. Statistical aspects of GWAS

To adjust for population stratification, we use PCA for cluster analysis. Assume that we have some observations from a few sub-populations, then we also have an underlying cluster of those observations indicating their corresponding sub-populations. If the between-cluster variation is much greater than the within cluster variation, then the first few principal components would capture the direction of the between-cluster’s difference ([61] Section 9, Page 202). In other words, the first few principal component scores can be regarded as the confounding covariates linked to the population structure.

In GWAS, PCA is performed on all the SNPs. Denote $w_i$ as the $i$th principal components and we have $W = (w_1, \cdots, w_m)$. The $i$th principal component scores $s_i = w_i^T x$. Then we include top $m$ principal component scores as addition covariates into the marginal logistic regression model. For the $j$th SNP $X_j$, we have

$$
\logit \mathbb{E}(Y|X_j) = \beta_0 + \beta_1 X_j + \sum_{i=1}^{m} \beta_{i+1} s_i
$$

(1.22)

$m$ is selected as the smallest integer that the $\lambda_{gc}$ for the PCA adjusted P-values are close to 1.
1.3 Genetic communication models

From the perspective of communication theory, DNA sequences can be regarded as are biological messages passing through generations. Communication theory investigates how information is transferred from information source through an information channel to the destination. In the classical Shannon and Weaver’s communication paradigm, a message (from information source) is encoded into the source code and transmitted through some noisy channel; then the received source code (including noise) is decoded back to a message (Figure 1.3) [106]. Here, DNA sequences serves as the source code transmitted through generations.

We are particularly interested in the receiver, i.e. how the source code is decoded into biological messages. DNA sequences are passed on through generations. Each individual already possesses its DNA sequences since its existence. How DNA sequences are formed originally is still puzzle. However, we have seen how organs are developed from an embryo in everyday life. Studying the information flowing from DNA sequences to individuals would greatly help us to understand the underlying biological meanings of DNA sequences.

One well known function of DNA sequences is to produce proteins. Genes are segments of DNA sequences that can be converted into mRNA and further into proteins. Communication models have been built to describe this DNA-
1.3. Genetic communication models

mRNA-protein information flow [41, 131]. They attempt to build frameworks to track how nucleotide pairs end up in protein with amino acids. May et al. did a survey on these communication models[77]. One interesting aspect of these communication models is how to handle communication noise. They also propose their own model emphasizing this [77]. In general, during the DNA sequences transmission, errors may be expected from multiple sources. They could be from replication during reproduction, error in the DNA-mRNA-protein processes and mutations caused by radiation, toxic substance, etc. In May’s model, un-replicated DNA sequences are transmitted(replicated) by a genetic channel and then is translated into protein by a genetic decoder. Furthermore, their genetic channel is expected to be noisy and error would be injected into the source code. Their model assumes that redundancies are in the DNA sequences such that the correct biological messages can be recovered despite of the errors.

Channel coding is a sub-discipline under the communication theory. It studies how to determine and even correct the noise in the communication channel. A basic channel coding model extends the Shannon and Weaver’s model by introducing channel encoder and decoder. Source coding and channel coding are two components of transmitter and receiver(Figure 1.4). A source encoder compresses messages into code. A channel encoder adds redundant parity bits and either detects or corrects transmitting errors through those parity bits. Automatic Repeat Request(ARQ) policy sends requests for re-transmission of the error code detected. Forward Error Correction(FEC) policy insert code with additional parity bits so that the original code can be induced from noisy channel without further communication. Code for these two purposes are called error detecting code(EDC) and error correcting code(ECC)[102]. In theory, if the transmission rate is less than the channel
capacity, the transmission error can be arbitrarily small with proper channel coding [107]. Suppose error correcting mechanism do exist in DNA sequences channel coding model would greatly reduce the complexity of DNA sequences through removing the redundancies.

![Channel coding communication model](image)

**Figure 1.4:** Channel coding communication model.

To prove the error correction code existence is hard. The genetic communication is one way and is only transmitted once per cycle. Nature selection is a brutal force terminating unsuitable individuals. Basically, it can be regarded as an error detection and termination policy. It stops some “error” messages being transmitted to the next generation. Apart from natural selection, suspicions of ECC have been raised in the literature. Battail has argued ECC is necessary for genetic information to be transmitted in a noise channel through generations[8]. Introns, the non-protein-coding regions, have been suspected as ECC for exon, the protein-coding region[32, 33]. Furthermore, in the translation of mRNA to protein, 64 nucleotide triplets only maps 20 amino acids. This has been regarded as evidence of redundancy so that ECC can correct communication noisy with[114]. Recently, Faria[31] found TRAV7 gene on the plasmid Lactococcus lactis genome could be identified as error correcting Hamming code.
Chapter 2

*Rank Stability Selection through Sample Splitting*

In GWAS, asymptotic multiple testing is the most widely used method in detecting genetic variants associated with the phenotype. Multiple testing calculates and tests the association statistic for every covariate. In human genomics, it suffers from issues such as inaccurate probabilistic models, population stratification, etc. We propose a novel variable selection method, rank stability selection (RSS). It is a non-parametric method for multiple testing, which is suitable for varieties of association measures or tests, such as correlation, mutual information, regression models, etc. RSS selects variable with persistent high rank of association test statistic in subsamples. It is an alternative of permutation tests with family-wise error control but computationally more efficient. One important assumption of RSS is that the statistics for non-associated covariates are exchangeable. It is weaker than i.i.d assumption required by multiple testing nor permutation tests. This gives RSS robustness in the presence of population stratification.

2.1 Introduction

Multiple testing is the core instrument of GWAS in searching for loci associated with the phenotype. For each locus, multiple testing examines the null hypothesis that not associated with the phenotype against some alternative hypothesis. After loci is tested, the rejection decision will be made based
on the overall error rate. Asymptotic hypothesis testing and permutation tests are two realizations of multiple testing. They both have their own strengths against each other and are widely used in GWAS for different purposes. However, when it comes to population stratification, without special treatment, they will both produce unexpected false positives. In this chapter, we present a novel non-parameter procedure for multiple testing to deal with population stratification.

Standard multiple testing uses the asymptotic distribution of the test statistic under the null hypothesis for decision making. For example, $\chi^2$ distribution is for Pearson’s Chi-squared test and normal distribution is for Cochran-Armitage trend test. Permutation technique is an alternative approach of finding the asymptotic distribution of the test statistics. The multiple testing with permutation method is also called a permutation test. A permutation test is non-parametric and exact[42]. Therefore, it can capture any deviations of the distribution of the test statistic from its asymptotic version under the null hypothesis. In GWAS, a permutation test generates the distribution of the test statistic under the null hypothesis by randomly permuting the labels of cases and controls. By doing so, any covariates should not be associated with the case and control labels. However, a permutation test is highly computational demanding which makes it less popular to the asymptotic multiple testing methods. Furthermore, it assumes random sampling. Non-trivial population structure would greatly bias the permutation test $P$-values. In this chapter, we seek to adopt non-parametric resampling method that would tackle this computing challenges and be robust to population structure as a confounding factor.

The key underlying idea of this chapter is stability. Stability of a statistic is its sensitivity to the subsampling of the original dataset. It is originally
used to examine the robustness to outliers of the statistical results. Jackknife randomly removes one sample point and monitors the distribution of the statistic. It is the earliest application of stability in statistical inference[28].

Only recently, stability has become a variable selection methodology in high-dimensional spaces problems. Unlike the significance in hypothesis testing, goodness of fit in model selection, stability is regarded as some robustness to perturbations [53]. The basic idea is to monitor the change in results of the method given different subsamples. It either selects a subset of covariates or gives a ranking of covariates based on some stability measure. Kalousis et al. studied three such measures for different types of variable selection problems[62]. Spearman’s rank correlation coefficient can be used for rankings. And Tanimoto distance metric is suited for subset selection. Since then, varieties of new stability measures were proposed [11, 24, 65, 72]. In gene differential expression problems, stability measure become quite important and necessary because of inconsistent experiment results [11]. However, the choice of cutoff threshold of the stability measurement remains a problem. Usually the choice is quite subjective, or computational intensive methods such as permutation tests are used.

Meinshausen et al. developed a new stability selection procedure with error control[82]. It is based on subsampling and randomized LASSO. It is the first time that error control is introduced to stability selection. By monitoring the frequencies of a covariate inside subsets selected by randomized LASSO, an upper bound of $E(V)$, the expected number of false positives, is established. The core assumption of their method is that the chances of non-associated covariates be selected by LASSO are exchangeable. Later on, Shah et al. improved it by introducing complementary pairs stability selection[105]. They manage to lower the error bound by using sample splitting. Baranowski
et al. applied the stability to rankings [7]. Their method dropped the error bound but remains consistency in variable selection. However, questions were raised on the power of stability selection. The letters to [82] and [1] find stability selection have relatively weak power in some cases including a GWAS simulation.

In this chapter, we present a novel variable selection method, Rank Stability Selection (RSS). It has two components. The first component is a rank based hypothesis test for exchangeability. Given there are two observations of a vector of random variables, if variables are exchangeable with each other, then the rankings of the observed values in one observation should be a random permutation of the rankings of the other observation. By comparing the rankings the test determines whether those variables are exchangeable with each other. Furthermore, if some important variables have a higher mean than the noise variables, and noise variables are exchangeable, then we can find the important variables by finding the largest set of exchangeable variables. The second component is random sample splitting. By splitting the sample into two equal-size halves, we can obtain two rankings of variables, and therefore apply the rank test to search for important variables. RSS repeatedly split the sample into halves to boost its variable selection strength.

Like permutation tests, RSS is non-parametric and is able to control Family-wise Error Rate (FWER). RSS requires the following assumptions:

- Test statistics of loci not associated with the phenotype are exchangeable.
- Loci associated with the phenotype are sparse among all loci.
- The test statistic ranks loci associated higher than loci not associated when the sample size is large enough.
2.1. Introduction

The exchangeable assumption is weaker than identically independent distributed (i.i.d.). It allows variables to be correlated. De Finetti’s theorem states that an infinite sequence of exchangeable random variables is a mixture of i.i.d. random variables [27]. In GWAS, exchangeable assumption in default assumes the population is a mixture of some sub-population. This population structure is also referred to as population stratification. It is potential confounder that would lead to false discoveries of associated loci. For asymptotic tests and permutation tests, population stratification needs to be corrected before multiple testing. RSS doesn’t require correction due to its exchangeable assumption.

We also derive explicit asymptotic conditions for RSS to be consistent under the high-dimensional problem setting. A variable selection method is consistent if it is able to select all loci that are associated with the phenotype given the sample size is large enough. We will show that RSS can well adapt the $m \gg n$ high-dimensional variable selection problem. It is also more efficient in terms of computation times.

In the end, we will give two numerical examples. In the simulation example, we construct different high dimensional variable selection problems and different scenario of population stratification. RSS is found that it is competitive to asymptotic methods and permutation methods. Most importantly, RSS is able to control FWER under strong population stratification without any alteration while asymptotic methods and permutation tests fail to do so. We also compare three methods on a dataset from a breast cancer GWAS. After adjusting for linkage-disequilibrium. RSS finds out the same SNP identified by Cochran-Armitage trend test and permutation test but with much small P-value reaching genome-wide significance.
2.2 Variable selection

Variable selection aims to find explanatory variables among a candidate set that are associated with the response variable. Let $Y$ be the response variable, and $\{X_1, \cdots, X_m\}$ be the explanatory variables. We use $A$, $S$, $N$ to denote the sets of indices of all explanatory variables, the explanatory variables associated with $Y$ and the explanatory variables not associated with $Y$ respectively. We further assume that $S$ and $N$ are disjoint i.e. $S \cap N = \emptyset$ and $S \cup N = A$. A variable selection problem then becomes estimating the set $S$ given some observation of $Y$ and $X_1, \cdots, X_m$.

The simplest way to estimate $S$ is to examine $Y$ with one $X_i$ at a time. The association can be measured using a test statistic, $\phi(\cdot, \cdot)$. Let $D = (d_1, \cdots, d_{2n})^\top$ denotes a dataset of $2n$ observations. $d_i = (Y_i, X_{i,1}, \cdots, X_{i,m})$ is one independent identical observation of $(Y, X_1, \cdots, X_m)$. $D$ can also been decomposed into column vectors, i.e. $D = (y, x_1, \cdots, x_m)$, where $y = (Y_1, \cdots, Y_{2n})$ and $x_i = (X_{1,i}, \cdots, X_{2n,i})$ for $i = 1, \cdots, m$. We denote $\phi_i = \phi(y, x_i)$ with respect to $D$. Without losing generality, we expect $\phi_i > \phi_j$ for $i \in S, j \in N$.

Note that $\phi_i$ is also a random variable. We can test the hypothesis of $\phi_i$ on whether $X_i$ is associated with $Y$. For example, in GWAS, $Y$ would be a binary case-control variable denoting some disease, and $X_i$ would be some genotyped SNPs. Let $\beta_i$ be the log odds ratio of being the case given $X_i$ having one minor allele. Then we can construct the null hypothesis as $J_i : \beta_i = 0$, and the alternative hypothesis as $J_i^* : \beta_i \neq 0$. For example, Pearson’s $\chi^2$ can be used as an association statistic. When Pearson’s $\chi^2$ is greater than some constant, we reject the null hypothesis.

To find out which explanatory variables are associated with $Y$, we can test $J_i$ for $i = 1, \cdots, m$ simultaneously and estimate $S$ using the hypotheses.
2.2. Variable selection

rejected. This technique is referred to as a multiple testing procedure. Let \( \hat{S} \) be the set of indices of the selected associated variables. Multiple testing aims to control the Family-Wise Error Rate (FWER), i.e.

\[
\text{FWER} : \Pr \left( \hat{S} \cap N \neq \emptyset \right).
\]

Bonferroni correction is a conservative and widely used technique to bound FWER. Basically, if the target FWER is \( \alpha \), it sets the type I error threshold for each hypothesis \( J_i \) as \( \frac{\alpha}{m} \), \( i = 1, \ldots, m \).

When the distribution of \( \phi_i \) under the null hypothesis is unknown, permutation technique is usually used to calculate its exact distribution. However, when population stratification is present, permutation technique is no longer suitable. Therefore, we introduce our Rank Stability Selection as an alternative non-parametric method.

2.2.1 Extension to interactive effects

\( \phi \) only captures marginal effects between an explanatory variable and the response variable. In case that both \( X_i \) and \( X_j \) are insignificantly associated with \( Y \) but \( X_i \cdot X_j \) are strongly associated with \( Y \). \( X_i \cdot X_j \) is one of the interactive effects called multiplicative effect. The variable selection framework can be easily extended to include interactive effects.

Let \( \tau_i \) be a non-empty subset of indices of all explanatory variables denoting their interactive effect. If there is only one element in \( \tau_i \), then \( \tau_i \) simply denotes the marginal effect of that variable. We can use \( \{ \tau_i | i = 1, \ldots, 2^m \} \) to denote all the marginal and interactive effects of explanatory variables.

In stead of using \( A, S \) and \( N \), we use \( A^*_i, S^*_i \) and \( N^*_i \) to denote the sets of \( \tau_i \) of all explanatory effects. However, in this case, the size of \( A^*_i \) would be \( 2^{2^m} \).
Despite of the scale of the problem, the rest of variable selection framework would still apply on \( A^*, S^* \) and \( N^* \). To test for the explanatory variable effects, we have to use a different test statistic \( \phi^*(\cdot, \cdot) \).

### 2.3 Rank Stability Selection

Rank Stability Selection (RSS) selects variables by finding the compliment set \( N \). Assuming \( S \) is sparse i.e. \(|S| < |N|\) and \( \{\phi_i|i \in N\} \) are exchangeable, we can estimate \( N \) using \( \hat{N} \) where \( \hat{N} \) is estimation of the largest set so that \( \{\phi_i|i \in \hat{N}\} \) are exchangeable.

To estimate \( N \), we first develop a hypothesis test to determine whether a set of variables are exchangeable based on two independent observations. Then, the RSS algorithm random, which will be introduced after the rank test, splits the original sample into two equal sizes. It artificially creates two mutually independent datasets which allows us to test the association statistics for exchangeability. To find the largest exchangeable subset, we perform the backward sequential test. It starts from the set of all explanatory variables, reduces the set by dropping variables one by one and stops when we fail to reject the null hypothesis that the test statistics of corresponding variables are exchangeable.

Exchangeability is a weaker assumption than the identically independently distributed (i.i.d.) assumption. It is able to naturally handle the population structure problem. By De Finetti’s theorem, an infinite sequence of exchangeable variables are a sequence of mixture of i.i.d. random variables. In this sense, \( \phi_1, \ldots, \phi_m \) can be regarded as i.i.d. conditioned on some unknown random variable \( \Theta \) and \( \Theta \) represents the population structure. In GWAS, mixture of sub-populations can be a significant confounding factor leading to
false discoveries of disease associated loci. Therefore, estimating the mixture is necessary to correct such biases in standard GWAS procedures, but this is not necessary under the exchangeable assumption.

2.3.1 A rank test for exchangeability

Before we embark on the full detail of the RSS algorithm, we first construct a hypothesis test on whether \( m \) variables are exchangeable. Let there be a random vector \( \phi^{(1)} = (\phi_1^{(1)}, \phi_2^{(1)}, \cdots, \phi_m^{(1)}) \). Under the sparsity assumption and exchangeable assumption, we expect that most components in \( \phi^{(1)} \) are exchangeable with each other but with a few exceptions. Here we reuse the set notations for simplicity. Let \( A, N, S \) denote the sets of indices of all variables, the largest subset of exchangeable variables and the rest in \( \phi^{(1)} \) respectively. We want to test whether \( N = A \).

Suppose that we observe another i.i.d. random vector of \( \phi^{(1)}, \phi^{(2)} = (\phi_1^{(2)}, \ldots, \phi_m^{(2)}) \). Define the rank of \( \phi_i^{(j)} \) with respect to \( \phi^{(j)} \) as

\[
R_i^{(j)} = \text{rank}(\phi_i^{(j)}) = \sum_{k=1}^{m} 1(\phi_k^{(j)} \geq \phi_i^{(j)}).
\] (2.2)

\( r^{(j)} = (R_1^{(j)}, \ldots, R_m^{(j)}) \) is the rank list of \( \phi^{(j)} \); and \( v^{(j)} = (V_1^{(j)}, \ldots, V_m^{(j)}) \) is an order list of \( \phi^{(j)} \) if \( R_{V_i^{(j)}} = \text{rank}(\phi_{V_i^{(j)}}^{(j)}) = i \) for \( i = 1, \ldots, m \). We call rank with value 1 the highest rank and rank with value \( m \) the lowest rank, and we assume there is no tie in the rank list, i.e. the rank list and the order list of \( \phi^{(j)} \) are permutations of \( (1, 2, \cdots, m) \).

The null hypothesis is that all components in \( \phi^{(1)} \) are exchangeable. The alternative hypothesis is \( S \neq \emptyset \) and \( E(\phi_i) > E(\phi_k) \) for \( i \in S \) and \( k \in N \). Under the null hypothesis, two rank lists should be a uniform random permutation of each other. We use a statistic called the highest lower rank, \( \xi \), to describe
2.3. Rank Stability Selection

the uniformity of the two rank lists,
\[
\xi = \min_{i=1}^{m} \left\{ \max\{R_i^{(1)}, R_i^{(2)}\} \right\}.
\] (2.3)

For example, when \( m = 5 \), we observe \( r^{(1)} = (4, 2, 1, 3, 5) \) and \( r^{(2)} = (3, 1, 5, 2, 4) \).
The lower ranks are \( (4, 2, 5, 3, 5) \), and the highest lower rank is 2. Lemma 2.1 states the exact distribution of highest lower rank under the null hypothesis given \( m \).

**Lemma 2.1.** Let there be two random vectors consisting of \( m \) exchangeable random variables as components. The highest lower rank \( \xi \) of two rank lists of those sets of variables has following probability mass function.
\[
\Pr(\xi = i) = (1 - A(i)) \cdot \prod_{k=1}^{i-1} A(k) \text{ for } 1 \leq i \leq \left\lfloor \frac{m}{2} \right\rfloor, \ i \in \mathbb{N}
\] (2.4)
where \( A(k) = \left(1 - \frac{k}{m-k}\right)^2 \), and \( \lfloor x \rfloor \) is the biggest integer smaller than \( x \).

**Proof.** The highest lower rank is \( \xi \) means there is an overlap in both top \( k \) elements of \( v^{(1)} \) and \( v^{(2)} \), but no overlap between both top \( k - 1 \) order lists. \( A(k) \) describes the probability of no overlap between both top \( k \) order lists conditioning on there is no overlap between both top \( k - 1 \) order lists. (2.4) is the product of conditional probabilities that there is no overlap for top \( i \) elements, \( i = 1, \ldots, k - 1 \) and times the probability there is an overlap in top \( k \) elements.

As a decision rule, we reject the null hypothesis if \( \xi \) is smaller than some integer. Under the alternative hypothesis, we would observe that some variable, \( \phi_i, i \in S \), has high ranks in both rank lists due to its greater mean. Consequently, \( \xi \) is expected to have a distribution left to its distribution under the null hypothesis. Therefore, we can construct a one-tail hypothesis test on \( \xi \). According to Lemma 2.1, the P-value of \( \xi \) with observed value \( q \) is
2.3. Rank Stability Selection

\[ \Pr(\xi \leq i) = 1 - \prod_{j=1}^{i} A(j) \]  
where \( 1 \leq i \leq \left\lfloor \frac{m}{2} \right\rfloor \), \( A(j) = \frac{(M-2j+2)(M-2j+1)}{(M-j+1)^2} \).

2.3.2 A resampling sequential selection algorithm

Rank Stability Selection (RSS) algorithm selects variables by examining how the rank list of \( \phi^{(s)} \) varies with respect to a subsample \( D^{(s)} \). Let \( D^{(s)} = (d_{i1}, \ldots, d_{in})^T \) be a subsample of \( D \), where \( \{i_1, \ldots, i_n\} \subset \{1, \ldots, 2n\} \). We may also rewrite \( D^{(s)} = (y^{(s)}, x_1^{(s)}, \ldots, x_m^{(s)}) \) and denote \( \phi^{(s)}_i = \phi(y^{(s)}, x_i^{(s)}) \).

Let \( D^{(1)}, D^{(2)} \) denote two equal size subsamples so that \( D^{(1)} \cap D^{(2)} = \emptyset \) and \( D^{(1)} \cup D^{(2)} = D \).

RSS has three stages. Firstly, we order \( \phi_i \) for \( i = 1, \ldots, m \) over the full sample \( D \) by the strength of association, and denote the order as \( v = (V_1, \ldots, V_m) \), where \( \text{rank}(\phi_{V_i}) = i \).

In the second stage, we randomly split the full sample into two equal-size subsamples for \( b \) times. \( D^{(k,1)}, D^{(k,2)} \) denote the subsamples at the \( k \)th split. And we have \( \phi_1^{(k,1)}, \ldots, \phi_m^{(k,1)} \) with respect to \( D^{(k,1)} \) and \( \phi_1^{(k,2)}, \ldots, \phi_m^{(k,2)} \) with respect to \( D^{(k,2)} \). Since all sample points are exchangeable and \( D^{(k,1)}, D^{(k,2)} \) have equal sample sizes, it is easy to show that \( \{\phi_i^{(j)} | i \in N\} \) are exchangeable for \( j = 1, 2 \) respectively. This means two rank-lists of \( \phi \) over two subsamples should be random permutation of each other if \( N = A \). Then construct \( m \) null and alternative hypotheses. For any \( k \in \{1, \ldots, m\} \) and \( i = 0, \ldots, m - 1 \), we have

\[ H_i : \{\phi_{V_{i+1}}^{(k,j)}, \ldots, \phi_{V_m}^{(k,j)}\} \text{ are exchangeable to each other for } j = 1, 2. \]  
(2.6)

\[ H_i^* : \text{There exists some } \phi_i^{(k,j)} \text{ so that } E(\phi_i^{(k,j)}) \gg E(\phi_k), \]  
(2.7)

where \( i, k \in \{V_{i+j}, \ldots, V_m\} \) and \( j = 1, 2 \).  
(2.8)
2.3. Rank Stability Selection

We perform two rank-list tests on $H_i$ against $H_i^*$ and record the P-value as $U_{i+1,k}$, $i = 0, \ldots, m - 1$. At the end of the second stage, we obtain an $m \times K$ matrix of P-values. The $i$th column of the matrix, $(U_{1,i}, \ldots, U_{m,i})$, is the vector of P-values for the same hypotheses $H_i$.

In the last stage, RSS aggregates the columns of the P-value matrix and then determines which explanatory variables are significant. We can use Brown’s method to combine those P-values under the same hypothesis (Details are in Section 2.7.2). Let $U_{i}^*$ be combined P-value of $(U_{1,i}, \ldots, U_{m,i})$ for $i = 1, \ldots, m$. We sequentially test hypothesis $H_i$ against $H_i^*$ for $i = 0, \ldots, m - 1$ and stop once we fail to reject $H_i$. Obviously, we have $H_1 \Rightarrow H_2 \Rightarrow \cdots \Rightarrow H_{m-1}$. Since all $H_{j:j>i}$ are nested in $H_i$, we do not have to test them if we believe $H_i$ is true. If we reject $H_{i-1}$ and fail to reject $H_i$, then we accept the claim that $\{V_{i+1}, \ldots, V_m\} \subseteq N$ and believe that $V_i \in S$. By assumption A2, we draw to the conclusion that $S = \{V_1, \ldots, V_i\}$. To restrict FWER from accumulating in sequential testing, we use Bonferroni correction to adjust P-values. We set P-value rejection threshold for each hypothesis test to be targeted FWER, $\alpha$, divided by the total number of tests performed in the procedure. Let $\hat{s} = \max\{k \mid U_j^* \leq \frac{\alpha}{k} \text{ for all } j = 1, \ldots, k\}$. We have $\hat{S} = \{V_1, \ldots, V_{\hat{s}}\}$ as our RSS estimation of $S$.

The RSS algorithm can be summarized as

1. Calculate order list $V = (V_1, \ldots, V_m)$ of $\phi_1 \ldots \phi_m$ with respect to $D$ in decreasing order.

2. Construct hypotheses $H_0, \ldots, H_{m-1}$, where

$$H_k : \{\phi_{V_{k+1}}, \ldots, \phi_{V_m} \text{ are exchangeable with each other.}\}$$

$k = 0, \ldots, m - 1$.

3. For $k = 1, \ldots, b$: 

---

34
2.4 RSS asymptotic properties

a) Randomly split $D$ into two equal datasets $D^{(k,1)}$ and $D^{(k,2)}$.

b) Calculate $\phi_1^{(k,1)}, \ldots, \phi_m^{(k,1)}$ with respect to $D^{(k,1)}$.

c) Calculate $\phi_1^{(k,2)}, \ldots, \phi_m^{(k,2)}$ with respect to $D^{(k,2)}$.

d) For $j = 1, \ldots, K$:
   
   • Test the $H_{j-1}$ against $H_j$ and record the P-value as $u_{i,j}$.

4. Let $u^*_j$ be combined P-value of $u_{1,j}, \ldots, u_{m,j}$.

5. Let $\hat{s} = \max\{k \mid u^*_j \leq \frac{\alpha}{k} \text{ for all } j = 1, \ldots, k\}$. Report explanatory variables with indices $\{t_1, \ldots, t_{\hat{s}}\}$ as significant.

2.4 RSS asymptotic properties

In this section we will investigate the asymptotic behaviour of RSS when sample size $2n$, number of covariates $m$, and resampling times in RSS all go to infinity. Note that we assume $|S|$ is fixed. Conceptually, a different $|S|$ would be a different variable selection problem.

Before state the asymptotic result, we first introduce three assumptions of RSS and their related conditions.

**Assumption 2.1.** $\{\phi_j : j \in N\}$ are exchangeable to each other.

**Assumption 2.2.** For any $i \in S$, $\lim_{n \to \infty} \Pr(\phi_i > \max\{\phi_j | j \in N\}) = 1$.

**Assumption 2.3.** Only a small proportion of covariates could be associated to the response variable, i.e. $|S| \ll m$.

**Remark 2.1.** A stronger version of assumption 2.1 would be $\{\phi_i | i \in N\}$ are i.i.d. The major difference is that if $\{\phi_i | i \in N\}$ are correlated to each other, then they could still be exchangeable but not i.i.d. In an association analysis,
when a confounding factor presents, \( \{\phi_i | i \in N\} \) are likely to be correlated to each other and appears to be associated to the response variable. In this case, i.i.d. assumption would be violated but it would bot necessarily be the case for the exchangeable assumption. Therefore, the exchangeable assumption is more robust than the confounding factors.

Remark 2.2. Assumption 2.2 simply requires that when we rank the variables in \( A \), all variables in \( S \) have higher rank than variables in \( N \) with probability one when sample size is large enough. Since RSS is a forward selection algorithm, this is essential for RSS to control FWER.

Remark 2.3. In cases of small sample sizes, we can use a working model with \( S^* \subset S \) and \( N^* = A \setminus S^* \) such that Assumption 2.2 is satisfied. Although the working model would be in conflict with Assumption 2.1, the exchangeable rank test may still fail to reject its null hypothesis. Small sample sizes would make weakly associated variable indistinguishable with the noise variables.

We then discuss the necessary conditions of the assumptions.

Definition 2.1. Let \( \phi(\cdot, \cdot) \) be a test statistic for the association between the response variable and a covariate, and \( \phi_i = \phi(Y, X_i) \) be a random variable. Say \( \phi \) is consistent at rate \( \kappa \) if \( \Pr(\phi_i \leq \max_{j \in N} \phi_j) \leq O(n^{-\kappa}) \) for any \( i \in S \).

Condition 2.1. \( m = O(n^{\kappa^*}) \) for some \( \kappa^* < \kappa \), where \( \phi \) is consistent at rate \( \kappa \).

Condition 2.2. \( |S| < 2 \sqrt{-\log(1 - \exp(-2.5))m} < 0.6 \sqrt{m} \)

Condition 2.3. \( b = c \log(m) \), for some \( c > 1 \).

Remark 2.4. Condition 2.1 indicates that \( m \) should not increase faster than the consistent rate of \( \phi \). It is a necessary condition for Assumption 2.2 (Lemma 2.2). Condition 2.1 is a weak condition for most of the association statistics. Take MLE as an example, its consistency rate is \( \kappa = \infty \). By Cramer-Rao bound,
we know that the difference between the estimator $\hat{\beta}$ and its true value $\beta$ is asymptotically normal distribution, i.e. $\sqrt{2n}(\hat{\beta} - \beta) \to N(0, \sigma^2_{\beta})$, where $\sigma^2_{\beta}$ is the inverse of Fisher information of $\beta$. Then since the estimator $\hat{\beta}_1 \to \beta_1$ and $\hat{\beta}_2 \to \beta_2$ with $\beta_2 > \beta_1$ as $n \to \infty$, $\Pr(\hat{\beta}_1 > \hat{\beta}_2)$ would vanish to 0 faster than any polynomial rate of $n$. The condition: “$m = O(n^c)$, $c > 0$” means $m$ could increase at polynomial rate of $n$. This is a common condition in high dimensional problem literature, i.e. extending BIC for Generalized Linear Regression Models(GLM)[15] and the modelling variability of rankings[50].

Remark 2.5. Condition 2.2 is the sparsity condition for RSS. The original sparsity assumption is that most of the coefficients of covariates in a linear model are zero. It is regarded as an essential assumption for $m \gg n$ variable selection problems. Basically it limits the size of the model space to search for the best model and prevents us from over fitting the model with only limited sample size. It was until recently that, for LASSO[113], the sparsity condition is found as $|S| = O(n^c)$ where $0 \leq c \leq 1$ [81, 133]. In contrast, Condition 2.2 gives explicit bound for $|S|$ in terms of $m$.

Remark 2.6. Condition 2.3 describes how many times of sample splitting is required for RSS. First of all, RSS retains FWER control even if $b = 1$. Doing multiple splitting increases its variable selection power. Condition 2.3 suggests that for RSS to be consistent at any FWER $\alpha$, we only need a very small number of sample splitting regarding to $m$. In contrast, permutation test will fail to control FWER when $m$ is small, and usually requires $b = 1000$. For bootstrap methods which are not reliable for non-parametric hypothesis testing, $b = 100$ is necessary. For RSS, when $m = 2000$, we found $b = 50$ is sufficient in Section 2.5.1.

Lemma 2.2. Given condition 2.1, assumption 2.2 is satisfied with probability 1
2.4. RSS asymptotic properties

when \( n \to \infty \).

Proof. Under condition 2.1, the probability of \( \phi_i > \phi_j \) for any \( i \in S \) and all \( j \in N \) is

\[
\Pr \left( \bigcup_{j \in N} \{ \phi_j > \phi_i \} \right) \leq \sum_{j \in N} \Pr(\phi_i > \phi_j) \\
\leq mO(n^{-\kappa}) = O(n^{\kappa - \kappa}) = o(1) \tag{2.9}
\]

\[\square\]

Lemma 2.3. Under the null hypothesis, when \( m \to \infty \), the CDF of the highest lower rank \( \xi \) is

\[
\Pr(\xi \leq k) = 1 - \exp \left( -\frac{k^2}{m} \right) + O \left( \frac{k^3}{m^2} \right). \tag{2.10}
\]

Furthermore, if \( k = o(m^{\frac{2}{3}}) \) i.e. \( k = O(m^{\frac{1}{2}}) \), then the quantile of \( \xi \) at \( \alpha \) is \( \sqrt{-\log(1-\alpha)m} + o(1) \).

Proof.

\[
\Pr(\xi \leq k) = \sum_{i=1}^{k} \left\{ (1 - A(k_i) \prod_{j=1}^{i-1} A(j)) \right\} = 1 - \prod_{i=1}^{k} A(i).
\]

Then we have,

\[
\log (1 - \Pr(\xi \leq k)) = \sum_{i=1}^{k} \log A(i)
= 2 \sum_{i=1}^{k} \log \left( 1 - \frac{i}{m-i} \right)
= 2 \sum_{i=1}^{k} \left( -\frac{i}{m-i} - O \left( \left( \frac{i}{m-i} \right)^2 \right) \right)
= -\left( \frac{k^2}{m} \right) + O \left( \frac{k^3}{m^2} \right)
\]

\[\square\]
Theorem 2.4. Under Condition 2.1, 2.2 and 2.3, RSS is consistent for any FWER $\alpha > 0$.

Proof. Fisher’s method is used to combine the P-values across $b$ sample splits. We first show that under the null hypothesis $H_j$, the aggregated value follows $\xi^2$ distribution, i.e.

$$-2 \sum_{i=1}^{b} \log u_{i,j} \sim \chi^2_{df=2b}. \quad (2.11)$$

Then we find a bound of the quantile of the $\chi^2$ distribution at the significant level. Lastly, we show that with probability 1, under the alternative hypothesis, the test statistics would be larger than that quantile.

Let there be $b$ sample splits, and the P-value $u_{i,j}$ of the $H_j$ in the $i$th sample split is essentially a statistic of two independent subsamples. When $n \to \infty$, ${u_{i,}}$ would converge \cite{10} to the uniform distribution on $[0, 1]$ under $H_j$.

(2.12) is an upper bound of quantile of $\chi^2$ distribution at $\alpha$ with degrees of freedom $2b$, $\chi^2(\alpha^*, 2b)$, where $\alpha^* = \frac{\alpha}{|S|}$ \cite{67}.

$$\chi^2(\alpha^*, 2b) \leq 2l - 2 \log(\alpha^*) + \sqrt{-8l \log(\alpha^*)} \quad (2.12)$$

Since the number of sample split $b$ is independent from the number of covariates $m$, condition 2.2 gives us $\log(|S|) < \log(1.36\sqrt{m}) < 0.5 \log(m) + 1.17$, and with condition 2.3 we have $l = c \log(m)$, $c > 1$ and

$$\limsup_{l \to \infty} \frac{\chi^2(\alpha^*, 2b)}{2b} \leq \limsup_{l \to \infty} \left(1 - \frac{\log(\alpha^*)}{b} + \sqrt{-\frac{2 \log(\alpha^*)}{b}}\right) \leq \limsup_{l \to \infty} \left(1 + \left(\frac{\log(|S|) - \log(\alpha)}{b}\right) + \left(\frac{2 \log(|S|) - 2 \log(\alpha)}{b}\right)\right) \leq 1 + \frac{1}{2c} + \frac{1}{\sqrt{c}} < 2.5$$

Then we show that RSS can reject $H_j$ with probability 1 when $n \to \infty$. If $j \in S$, then $\Pr(R_j > |S|) = O(n^{\kappa^*-\kappa})$. We also have $\Pr(\xi_j > \frac{|S|}{2}) = O(n^{\kappa^*-\kappa})$. The log
2.4. RSS asymptotic properties

P-value of $\xi_j$ would be bounded as well, i.e. $\sup_{j \in S} \log u_{i,j} \to 1 - \exp(-\frac{|S|^2}{m})$ with probability $1 - O(n^{\kappa^*-\kappa})$.

$$\lim_{n \to \infty} \Pr\left(-\frac{2}{2b} \sum_{i=1}^{b} \log u_{i,j} \geq 2.5\right)$$

$$\leq \lim_{n \to \infty} \Pr\left(-\log \left(1 - \exp\left(-\frac{|S|^2}{4m}\right) + O(m^{-0.5})\right) \geq 2.5\right) \cdot \lim_{n \to \infty} \Pr\left(\xi_j \leq \frac{|S|}{2}\right)$$

$$\leq \lim_{n \to \infty} \Pr\left(-\log \left(\exp(-2.5) + O(m^{-0.5})\right) \geq 2.5\right) \cdot \lim_{n \to \infty} \Pr\left(\xi_j \leq \frac{|S|}{2}\right)$$

$$\leq \lim_{n \to \infty} (1 - O(n^{\kappa^*-\kappa})) = 1$$
2.5 Numerical examples

2.5.1 Simulation studies

We construct a GWAS case-control simulation study to evaluate the performance of RSS. The response variable \( Y \) is binary denoting some trait in interest. The covariates \( \{X_1, \ldots, X_m\} \) are single nucleotides polymorphisms (SNPs) taking values from \( \{0, 1, 2\} \). The collection of covariates are also called genotype data in GWAS. We simulate genotype data in two scenarios. In the first case, all covariates are i.i.d. For each SNP \( X_i \), we denote its allele frequency as \( \theta_i \). \( \theta_1, \ldots, \theta_m \) are i.i.d. with Beta distribution, \( \text{Beta}(\alpha = 2, \beta = 2) \). The values of \( X_i \) are generated from a binomial distribution, \( \text{Binom}(2, \theta_i) \). In the second case, we assume there is a non-trivial mixture of sub-populations. We assume SNPs within in each sub-population are independent. Let \( \theta^j_i \) be the allele frequency of \( i \)th SNP in the \( j \)th sub-population. We have \( \logit(\theta^j_i) \sim N(\logit(\theta_i), \sigma^2_{\theta_i}) \) with \( \sigma_{\theta_i} = 0.6 \) and some \( \theta_j \sim \text{Beta}(\alpha = 2, \beta = 2) \).

Let \( S \) denotes the index set of SNPs truly associated with \( Y \). We simulate values of \( y \) through a logistic regression model of \( Y \), i.e. \( Y \sim \text{Bernoulli}(\gamma) \), where

\[
\logit(\gamma) = (1, x_S)\beta_S, \tag{2.13}
\]

where \( x_S = (X_{j_1}, \ldots, X_{j_k}) \) for \( j_i \in S, k = |S| \) and \( \beta_S = (\beta_0, \beta_1, \ldots, \beta_k)^\top \) is a vector of coefficients in a logistic regression model.

We construct five models with coefficients listed in Table 2.1. They covered different sizes of \( S \), strength of the signal and population structures. We also set up three scenarios with the number of covariates \( m = 100, 500, 2500 \). In total, we simulated 15 datasets with sample size \( n = 400 \).
2.5. Numerical examples

<table>
<thead>
<tr>
<th>Model #</th>
<th>Non-zero Coefficients</th>
<th># sub-pop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8, -0.7, 0.6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2.0, -1.9, 1.8, -1.7, 1.6, -1.5, 1.4, -1.3, 1.2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.8, -0.7, 0.6</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2.0, -1.9, 1.8, -1.7, 1.6, -1.5, 1.4, -1.3, 1.2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0.8, -0.7, 0.6</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 2.1: 5 simulation models

To study the performance of RSS, we choose two association statistics, Wald z-score of logistic regression coefficient and mutual information. The asymptotic distribution of z-score is standard normal under the null hypothesis when two variables are not associated. Mutual information takes values in $[0, \infty]$, where 0 means two random variables are independent. There is no asymptotic distribution under the null hypothesis available. Four methods of variable selection are included. They are asymptotic tests with Bonferroni correction (ASYM), asymptotic test with Bonferroni correction and genomic control (GC), MaxT permutation test (PERM)[42] and RSS. To access the performance, we use positive selection rate (PSR) to measure the variable selection power and Family-Wised Error Rate (FWER) to the error. All methods aim to control the FWER at $\alpha = 0.05$. Let $\hat{S}^{(j)}$ be the indices of variables selected in the $j$th replication $j = 1, \cdots, R$, then

$$PSR = \frac{1}{R} \sum_{j=1}^{R} \frac{|\hat{S}^{(j)} \cap S|}{|S|}, \quad FWER = \frac{1}{R} \sum_{j=1}^{R} 1(\hat{S}^{(j)} \setminus S \neq \emptyset)$$ (2.14)

Figure 2.2 and 2.3 shows the results on the logistic regression. In Figure 2.1, RSS and GC controls FWER well in all cases. PERM and ASYM are able to control FWER in Model 1 and 2. When the mixture of sub-populations presents, they fail to do so in Model 3,4,5. RSS is relatively conservatively in FWER control especially when $n$ is small. When there is no population
2.5. Numerical examples

The sample size is 400. Simulation is repeated for 500 times. “RSS” split samples for 50 times. “ASYM” is the standard logistic regression with Wald test. “GC” denotes the average genomic control inflation. “PERM” is MaxT permutation test with 100 times permutations. The red horizontal line is at value 0.05. GC value is the genomic control coefficient which indicates the strength of population stratification as a confounding factor.

Figure 2.1: FWER of multiple testing methods
2.5. Numerical examples

The sample size is 400. Simulation is repeated for 500 times. “RSS” split samples for 50 times. “ASYM” is the standard logistic regression with Wald test. “GC” denotes the average genomic control inflation. “PERM” is MaxT permutation test with 100 times permutations. The red horizontal line is at value 0.05. GC value is the genomic control coefficient which indicates the strength of population stratification as a confounding factor.

**Figure 2.2:** PSR of multiple testing methods
2.5. Numerical examples

Figure 2.3: Multiple testing methods results using Mutual Information

The sample size is 400. Simulation is repeated for 500 times. “RSS” split samples for 50 times. “PERM” is MaxT permutation test with 100 times permutations. The red horizontal line is at value 0.05. GC value is the genomic control coefficient which indicates the strength of population stratification as a confounding factor.

stratification, PERM has the most precise control of FWER. In Model 3 and 4, when the degree of population stratification is very high, PERM tends to choose a lot more false positives and is followed by ASYM. On the other hand, RSS has a good FWER control while GC becomes to conservative. In terms of selection power (Figure 2.2), PERM is the most powerful method. RSS have little power when $m$ is small, i.e. $m = 100$. But when $m$ is large, it is more powerful than ASYM and GC and close to PERM in Model 1 and Model 2. Since ASYM and PERM lose FWER control in Model 3,4,5, the power comparison is not fair. However, despite of the high FWER for ASYM and PERM in Model 3 and 4, RSS is still competitive with ASYM when $m$ is
large. When no population stratification is present (i.e. Model 1,2), RSS is less powerful than permutation tests but more than ASYM. In the cases of population stratification, RSS is still more powerful than asymptotic tests with genomic control, while the rest fail to control the FWER.

Figure 2.3 shows the results of mutual information. Since mutual information is non-parametric, only results of RSS and PERM are shown here. In general, we find that logistic regression is more powerful than mutual information. This was expected since logistic regression is the underlying true model. The results are similar as the logistic regression case. RSS is more conservative but yet hold FWER in control when population stratification is strong. And PERM fails to do so. In terms of power, PERM is more powerful, but the gap is narrowed when $m$ is large.

Despite of that RSS has weaker power than PERM, RSS is computational fast due to fewer number of resamplings. Here we run RSS with $b = 50$ sampling splittings and 100 times permutation for PERM. For RSS the number of sample splitting only affects the power. RSS still retains error control and consistency when $b = 1$ (Lemma 2.4). On the other hand, the number of permutation in PERM is strictly related to the precision of the P-value and therefore is linked to FWER control. Here we only run PERM with 100 times permutation because it is very computational demanding. It is fewer than the standard 1000 times. But the error it caused does not explain its large FWER in Model 3 and 4. Comparing to the standard PERM, RSS is 20 times faster.

In summary, RSS is a fast non-parametric variable selection methods. It is competitive to the asymptotic test and permutation test but robust to population stratification.
2.5. Numerical examples

2.5.2 GWAS breast cancer example

RSS requires additional treatments when it is applied to real GWAS dataset. Linkage-Disequilibrium (LD) compromises the exchangeable assumption 2.1 of test statistics and reduces the power of RSS. To fix this, certain pruning for LD is necessary, i.e. filtering highly correlated SNPs. In the following example, RSS shows promising potential for GWAS if the SNPs in strong LD to the signal is removed.

We look at an Australian breast cancer case and control GWAS. The dataset have 512, 297 SNPs and 491 patients. Here, we focus on a ESR1 GENE related pathway with 366 SNPs on the genome. ESR1 Gene has been found associated with Breast cancer in both gene expression studies and GWAS [55, 134].

The genomic control $\lambda = 1.05$ for this pathway suggests insignificant population stratification. Asymptotic multiple testing finds only one SNP, rs3778080, that is significantly associated with the phenotype. Its Bonferroni corrected P-value is 0.022. Meanwhile, permutation (MaxT procedure with 10,000 permutations) finds its P-value to be 0.0073, and RSS($m = 50$) finds its P-value to be 0.119. Asymptotic methods and permutation tests both identify the SNP rs3778080 to be significantly associated with breast cancer within the context of the 366 SNPs. However, they all fail to meet the genome wide significant criteria that is adjusted for all 512,297 SNPs in the study. On the other hand, RSS fails to identify this SNP even with respect to 366 SNPs. We suspect this is due to the uneven correlation between those 366 SNPs which would break the exchangeable assumption of the test statistics.

In GWAS, the correlation structure between SNPs are called linkage-disequilibrium (LD). Figure 2.4 illustrates the LD structure of these 366 SNPs. We note that SNPs close to each other tend be correlated. Among top 20
SNPs (Table 2.2), quite a few SNPs are highly correlated to the top SNP rs3778080. Among 366 SNPs, 21 SNPs are in strongly LD to SNP rs3778080 with $D' > 0.99$. But their associations with the phenotype are not significant. There are two possible explanations. One is that these 21 SNPs are truly associated with the phenotype, but the signal is very weak due to the sample size. Another possibility is that they are not associated with the phenotype but only correlated to the SNP rs3778080 or other confounders. In the second case, the RSS assumption 2.1 is violated. If we remove those 21 SNPs from the variable selection problem, RSS finds the P-value of rs3778080 to be $2.63 \times 10^{-5}$ and no other SNPs are significant. If we use Bonferroni correction for genome-wide significance, the adjusted P-value will be $2.63 \times 10^{-5} \cdot \frac{512207}{512297} \sim 0.039$, which is genome-wide significant.

When being applied to real GWAS datasets, RSS requires special treatments for LD. LD compromises the exchangeable assumption 2.1 of test statistics and reduces the power of RSS. To fix this, certain pruning for LD is necessary. Linkage disequilibrium based SNP pruning is a procedure implemented in PLINK [91]. Basically, for a set of highly correlated and closely located SNPs, the procedure keeps one SNP and filters out others based on either the variance inflation factors (VIF) in a local linear regression model, or their pairwise correlation measurement. Pruning reduces the test power on a single variant. But multiple testing will benefit from pruning since the total number of tests are reduced. In Section 3.1, we will briefly introduce what is LD, how it affects GWAS and the popular methods dealing with LD.
2.5. Numerical examples

Figure 2.4: LD structure

Pairwise $D'$ of 366 SNPs. The vertical line indicates the location of rs3778080.
<table>
<thead>
<tr>
<th>Method</th>
<th>CHR</th>
<th>BP</th>
<th>adj. ASYM P</th>
<th>PERM P</th>
<th>RSS P</th>
<th>(D') to rs3778080</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3778080</td>
<td>6</td>
<td>152385236</td>
<td>0.013</td>
<td>0.007</td>
<td>0.119</td>
<td>0.000</td>
</tr>
<tr>
<td>rs2207232</td>
<td>6</td>
<td>152340288</td>
<td>0.118</td>
<td>0.064</td>
<td>0.432</td>
<td>1.000</td>
</tr>
<tr>
<td>rs13216134</td>
<td>6</td>
<td>152328484</td>
<td>0.174</td>
<td>0.096</td>
<td>0.468</td>
<td>0.985</td>
</tr>
<tr>
<td>rs3778084</td>
<td>6</td>
<td>152388039</td>
<td>0.183</td>
<td>0.099</td>
<td>0.551</td>
<td>1.000</td>
</tr>
<tr>
<td>rs3778089</td>
<td>6</td>
<td>152393761</td>
<td>0.183</td>
<td>0.099</td>
<td>0.514</td>
<td>1.000</td>
</tr>
<tr>
<td>rs3778082</td>
<td>6</td>
<td>152387664</td>
<td>0.195</td>
<td>0.104</td>
<td>0.498</td>
<td>1.000</td>
</tr>
<tr>
<td>rs13203975</td>
<td>6</td>
<td>152333104</td>
<td>0.379</td>
<td>0.197</td>
<td>0.542</td>
<td>0.984</td>
</tr>
<tr>
<td>rs6932864</td>
<td>6</td>
<td>152376475</td>
<td>0.393</td>
<td>0.203</td>
<td>0.613</td>
<td>1.000</td>
</tr>
<tr>
<td>rs3020418</td>
<td>6</td>
<td>152345162</td>
<td>1.000</td>
<td>0.512</td>
<td>0.737</td>
<td>1.000</td>
</tr>
<tr>
<td>rs3775777</td>
<td>4</td>
<td>70709510</td>
<td>1.000</td>
<td>0.657</td>
<td>0.735</td>
<td>0.291</td>
</tr>
<tr>
<td>rs4149534</td>
<td>4</td>
<td>70718005</td>
<td>1.000</td>
<td>0.657</td>
<td>0.723</td>
<td>0.291</td>
</tr>
<tr>
<td>rs3775770</td>
<td>4</td>
<td>70724270</td>
<td>1.000</td>
<td>0.722</td>
<td>0.738</td>
<td>0.396</td>
</tr>
<tr>
<td>rs4149527</td>
<td>4</td>
<td>70724996</td>
<td>1.000</td>
<td>0.722</td>
<td>NA</td>
<td>0.396</td>
</tr>
<tr>
<td>rs3775768</td>
<td>4</td>
<td>70725112</td>
<td>1.000</td>
<td>0.722</td>
<td>NA</td>
<td>0.396</td>
</tr>
<tr>
<td>rs4147581</td>
<td>11</td>
<td>67351585</td>
<td>1.000</td>
<td>0.758</td>
<td>NA</td>
<td>0.041</td>
</tr>
<tr>
<td>rs3778099</td>
<td>6</td>
<td>152418575</td>
<td>1.000</td>
<td>0.768</td>
<td>NA</td>
<td>0.984</td>
</tr>
<tr>
<td>rs1238574</td>
<td>4</td>
<td>70709023</td>
<td>1.000</td>
<td>0.821</td>
<td>NA</td>
<td>0.624</td>
</tr>
<tr>
<td>rs10082248</td>
<td>1</td>
<td>209867116</td>
<td>1.000</td>
<td>0.828</td>
<td>NA</td>
<td>0.087</td>
</tr>
<tr>
<td>rs1881668</td>
<td>4</td>
<td>70725456</td>
<td>1.000</td>
<td>0.857</td>
<td>NA</td>
<td>0.056</td>
</tr>
<tr>
<td>rs3020411</td>
<td>6</td>
<td>152343763</td>
<td>1.000</td>
<td>0.909</td>
<td>NA</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Table 2.2:** Results on top 20 SNPs among 366 SNPs

Adj ASYM P the asymptotic P-value after Bonferroni correction for FWER. PERM P are P-values of permutation MaxT test. RSS P are RSS P-values. PERM P and RSS P don’t need to be adjusted for FWER. Due to sparsity constrain, only a few RSS P are calculated.
2.6 Conclusion

In this chapter, we have presented a novel non-parametric method for variable selection, Rank Stability Selection (RSS). It splits the sample into two halves and compares the rankings of the covariate association statistics between two sub-samples. RSS relies on the exchangeable assumption. It assumes the association statistics of covariates are exchangeable under the null hypothesis. This assumption accepts a non-trivial mixture of sub-populations. Population structure is a significant confounding factor in GWAS. Under the exchangeable assumption, RSS can adapt the population structure, and deliver reliable results without estimating or correcting population structure.
2.7 Supplementary

2.7.1 Chapter notations

\( X_i \) The \( i \)th explanatory variables.
\( Y \) The response variable.
\( A \) Set of indices of all explanatory variables.
\( S \) Set of indices of explanatory variables associated with the response variable.
\( N \) Set of indices of explanatory variables not associated with the response variable.
\( \widehat{N} \) The estimation of \( N \).
\( m \) The number of total explanatory variables.
\( n \) The number of observations (sample size).
\( d_i \) The \( i \)th sample vector of the response variables and all the explanatory variables.
\( D \) A sample of \( n \) sample vectors.
\( y \) The vector of all sample points of \( Y \).
\( x_i \) The vector of all sample points of \( X_i \).
\( \phi \) An association test statistic.
\( \phi_i \) The association test statistic between \( y \) and \( x_i \).
\( D^{(1)} \) The first subsample of \( D \).
\( \phi_i^{(1)} \) The test statistic applied on the \( D^{(1)} \).
\( R_i^{(1)} \) The rank of \( \phi_i^{(1)} \).
\( V_i^{(1)} \) The order of \( \phi_i^{(1)} \).
\( r^{(1)} \) The vector of all \( R_i^{(1)} \) for \( i = 1 \ldots m \).
\( v^{(1)} \) The vector of all \( V_i^{(1)} \) for \( i = 1 \ldots m \).
\( \xi \) The test statistic of RSS.
2.7.2 Brown’s method for combining dependent P-values

The Fisher’s method combines the P-values of independent hypothesis tests under the same hypothesis. Let $U_{1,j}, \cdots, U_{m,j}$ denote the P-values of hypothesis $H_{j-1}$ calculated from $m$ independent random split. If the null hypothesis is true and $U_{1,j}, \cdots, U_{m,j}$ are independent of each other, then according to the Fisher’s method we have

$$G_j = -2 \sum_{i=1}^{m} \log(U_{i,j}) \sim \chi^2_{2m}. \tag{2.15}$$

However, P-values of RSS in each sample splitting are not independent. Sample splitting results are conditional on the original full sample, and correlations between results are expected. The Brown’s method extends the Fisher’s method to allow hypothesis tests to be dependent\[12\]. It uses $G^*_j = c_1 \chi^2_{c_2}$ to approximate $G_j$ under the null hypothesis. $c_1$ and $c_2$ are chosen so that the first two moments of $G^*_j$ matches those of $G_j$. Under the null hypothesis $U_{i,j} \sim \text{Unif}(0, 1)$, and we have

$$E(G_j) = \sum_{i=1}^{m} E(-2 \log(U_{i,j})) = 2m$$

$$\text{Var}(G_j) = \sum_{i=1}^{m} \text{Var}(-2 \log U_{i,j}) + \sum_{k \neq l} \text{Cov}(-2 \log U_{k,j}, -2 \log U_{l,j})$$

$$= 4m + 4 \sum_{k \neq l} \rho_{k,l},$$

where $\rho_{k,l}$ is the correlation coefficient of $-2 \log U_{k,j}$ and $-2 \log U_{l,j}$ for all $j = 1, \cdots, m$.

$$\begin{cases} E(G_j) = E(G) \\ \text{Var}(G_j) = \text{Var}(G) \end{cases} \quad \Rightarrow \quad \begin{cases} 4m = c_1 c_2 \\ 4m + 4 \sum_{k \neq l} \rho_{k,l} = 2c_2^2 c_2 \end{cases}$$

$$\Rightarrow \begin{cases} c_1 = \frac{1}{2} + \frac{1}{m} \sum_{k \neq l} \rho_{k,l} \\ c_2 = \frac{2}{m + \sum_{k \neq l} \rho_{k,l}} \end{cases} \tag{2.16}$$
In order to find \( c_1 \) and \( c_2 \), all we need is to estimate \( \rho_{k,l} \). \( U_{k,i} \) measures the significance of association signals captured by \( \phi(\cdot, \cdot) \) among \( \{\phi_{V_i}^{(j)}, \cdots, \phi_{V_i}^{(j)}\} \), for \( j = 1, 2 \) at the \( k \)th sample split. Let \( Q_{k,V_i} \) be the P-value of \( \phi_{V_i} \) evaluated at the 1st subsample of the \( k \)th split under the \( J_{V_i} \). \( U_{k,i} \) and \( Q_{k,V_j} \) both measures the strength of the strongest signal defined by \( \phi \) at sample split \( k \) in different ways. We expected they are closely positive correlated.

On the other hand, \( \rho_{k,l} \) reflects the correlation between \( U_{k,i} \) and \( U_{l,i} \) when \( H_{i-1} \) is true. \( \rho_{k,l} \) is a relatively easy quantity to estimate, since it does not depend on the first two moments of the P-values. We could use \( Q_{k,V_i} \) and \( Q_{l,V_j} \) to estimate \( \rho_{k,l} \). When \( H_{i-1} \) is true and according to the assumption A2, \( Q_{k,V_{j_1}} \) and \( Q_{k,V_{j_2}} \) should be exchangeable uniformly distributed on (0,1) for \( j_1, j_2 \geq i \). Consequently, \( \{(Q_{k,V_{j}}, Q_{l,V_{j}}) : j \geq i\} \) can be regarded as some exchangeable random vectors for \( j \geq i \).

Then we can estimate \( \rho_{k,l} \) using sample correlation coefficient of \( (Q_{k,V_{j}}, Q_{l,V_{j}}) \) for \( j \geq i \). It is easy to show that the Pearson’s sample correlation coefficient is unbiased for the exchangeable random vectors. Practically, we may use the \( 2r \) variables in the middle of \( V_i \), i.e.

\[
X_{V_{i_{\lfloor P/2 \rfloor-r}}} \cdots X_{V_{i_{\lfloor P/2 \rfloor+r}}}
\]

Let \( M_{2r} \) denote their labels. Under the sparsity assumption A3, we know that \( M_{2r} \subset N \) for sure. Let \( G_{k,j}^* = -2 \log(Q_{k,j}) \). And we have

\[
\hat{\rho}_{k,l} = \frac{\sum_{j \in M_{2r}} (G_{k,j}^* - \overline{G}_k^*)(G_{l,j}^* - \overline{G}_l^*)}{\sqrt{\sum_{j \in M_{2r}} (G_{k,j}^* - \overline{G}_k^*)^2} \sqrt{\sum_{j \in M_{2r}} (G_{l,j}^* - \overline{G}_l^*)^2}}, \tag{2.17}
\]

where \( \overline{G}_k^* = \frac{1}{2r} \sum_{j \in M_{2r}} G_{k,j}^* \) and \( \overline{G}_l^* = \frac{1}{2r} \sum_{j \in M_{2r}} G_{l,j}^* \).

If the null distribution of \( \phi_i \) is unknown, i.e. \( Q_{k,i} \) is not available. We can
use empirical P-values $\hat{Q}_{k,i}$,

$$
\hat{Q}_{k,j} = \frac{1}{P} \sum_{i=1}^{P} 1(\phi_i^{(k,1)} \leq \phi_j^{(k,1)})
$$

(2.18)

### 2.7.3 Consistency of a single split RSS

**Theorem 2.5.** RSS with single split is consistent with respect to Type I error rate $\alpha$ if Condition 2.1 and $|S| \leq 2\sqrt{-\log(1-\alpha)P}$.

**Proof.** Lemma 2.2 shows that probability 1, i.e. any covariates with indices belong to $S$ have higher rank than those belong to $N$. In this case, among top $|S|$ variables, we will have $\xi \leq \left\lfloor \frac{|S|}{2} \right\rfloor$. The asymptotic upper bound of P-value of hypothesis $H_0$ is

$$
u_{0,*} \leq \lim_{P \to \infty} \Pr \left( \xi \leq \left\lfloor \frac{|S|}{2} \right\rfloor \right)
$$

$$
\leq 1 - \exp \left( -\frac{\left( \frac{|S|}{2} \right)^2}{P} \right)
$$

$$
\leq 1 - \exp^{\log(1-\alpha)P/P} \leq \alpha.
$$
2.7.4 Supplementary tables and figures

Figure 2.5: FDR of multiple testing methods
## 2.7. Supplementary

| Method     | $P$ | $\lambda_{GC}$ | FWER | FDR  | PSR  | $|\hat{S}|$ |
|------------|-----|-----------------|------|------|------|----------|
| ASYM       | 100 | 1.08            | 0.022| 0.0157| 0.35 | 1.06     |
| GC         | 100 | 1.08            | 0.014| 0.0104| 0.31 | 0.93     |
| Permutation| 100 | 1.08            | 0.046| 0.0310| 0.42 | 1.25     |
| RSS        | 100 | 1.08            | 0.002| 0.0039| 0.19 | 0.56     |
| M.I.Permutation | 100 | 1.08 | 0.046 | 0.0391 | 0.33 | 0.98 |
| M.I.RSS    | 100 | 1.08            | 0.000 | 0.0000 | 0.13 | 0.39     |
| ASYM       | 500 | 1.02            | 0.014| 0.0107| 0.26 | 0.77     |
| GC         | 500 | 1.02            | 0.012| 0.0093| 0.25 | 0.74     |
| Permutation| 500 | 1.02            | 0.034| 0.0260| 0.31 | 0.93     |
| RSS        | 500 | 1.02            | 0.018| 0.0190| 0.29 | 0.87     |
| M.I.Permutation | 500 | 1.02 | 0.050 | 0.0602 | 0.24 | 0.72 |
| M.I.RSS    | 500 | 1.02            | 0.010| 0.0186| 0.19 | 0.58     |
| ASYM       | 2500| 1.00            | 0.024| 0.0288| 0.17 | 0.50     |
| GC         | 2500| 1.00            | 0.016| 0.0207| 0.16 | 0.49     |
| Permutation| 2500| 1.00            | 0.058| 0.0675| 0.21 | 0.62     |
| RSS        | 2500| 1.00            | 0.040| 0.0522| 0.19 | 0.58     |
| M.I.Permutation | 2500 | 1.00 | 0.034 | 0.0689 | 0.15 | 0.44 |
| M.I.RSS    | 2500| 1.00            | 0.010| 0.0251| 0.13 | 0.39     |

Table 2.3: Simulation results for Model 1

| Method     | $P$ | $\lambda_{GC}$ | FWER | FDR  | PSR  | $|\hat{S}|$ |
|------------|-----|-----------------|------|------|------|----------|
| ASYM       | 100 | 1.13            | 0.012| 0.00167| 0.527 | 4.74     |
| GC         | 100 | 1.13            | 0.006| 0.00079| 0.472 | 4.25     |
| Permutation| 100 | 1.13            | 0.044| 0.01120| 0.657 | 5.91     |
| RSS        | 100 | 1.13            | 0.002| 0.00515| 0.070 | 0.63     |
| M.I.Permutation | 100 | 1.13 | 0.056 | 0.01027 | 0.585 | 5.27 |
| M.I.RSS    | 100 | 1.13            | 0.000 | 0.00000 | 0.078 | 0.70     |
| ASYM       | 500 | 1.02            | 0.016| 0.00289| 0.398 | 3.59     |
| GC         | 500 | 1.02            | 0.010| 0.00183| 0.391 | 3.52     |
| Permutation| 500 | 1.02            | 0.046| 0.00765| 0.532 | 4.79     |
| RSS        | 500 | 1.02            | 0.014| 0.00302| 0.471 | 4.24     |
| M.I.Permutation | 500 | 1.02 | 0.054 | 0.01410 | 0.457 | 4.11 |
| M.I.RSS    | 500 | 1.02            | 0.016| 0.00640| 0.385 | 3.47     |
| ASYM       | 2500| 1.01            | 0.008| 0.00188| 0.324 | 2.91     |
| GC         | 2500| 1.01            | 0.008| 0.00201| 0.319 | 2.87     |
| Permutation| 2500| 1.01            | 0.034| 0.00863| 0.434 | 3.91     |
| RSS        | 2500| 1.01            | 0.034| 0.01076| 0.410 | 3.69     |
| M.I.Permutation | 2500 | 1.01 | 0.042 | 0.01271 | 0.369 | 3.32 |
| M.I.RSS    | 2500| 1.01            | 0.012| 0.00431| 0.322 | 2.90     |

Table 2.4: Simulation results for Model 2
### 2.7. Supplementary

| Method          | $P$ | $P_{GC}$ | FWER | FDR  | PSR | $|\hat{S}|$ |
|-----------------|-----|----------|------|------|-----|--------|
| ASYM            | 100 | 1.21     | 0.040| 0.022| 0.39 | 1.18   |
| GC              | 100 | 1.21     | 0.008| 0.005| 0.30 | 0.89   |
| Permutation     | 100 | 1.21     | 0.088| 0.044| 0.45 | 1.34   |
| RSS             | 100 | 1.21     | 0.000| 0.000| 0.19 | 0.58   |
| M.I.Permutation | 100 | 1.21     | 0.058| 0.039| 0.35 | 1.06   |
| M.I.RSS         | 100 | 1.21     | 0.000| 0.000| 0.15 | 0.44   |
| ASYM            | 500 | 1.15     | 0.096| 0.089| 0.27 | 0.81   |
| GC              | 500 | 1.15     | 0.018| 0.024| 0.20 | 0.59   |
| Permutation     | 500 | 1.15     | 0.138| 0.102| 0.33 | 0.99   |
| RSS             | 500 | 1.15     | 0.052| 0.050| 0.28 | 0.83   |
| M.I.Permutation | 500 | 1.15     | 0.110| 0.109| 0.26 | 0.78   |
| M.I.RSS         | 500 | 1.15     | 0.018| 0.024| 0.19 | 0.57   |
| ASYM            | 2500| 1.14     | 0.076| 0.107| 0.17 | 0.52   |
| GC              | 2500| 1.14     | 0.008| 0.022| 0.11 | 0.33   |
| Permutation     | 2500| 1.14     | 0.124| 0.145| 0.22 | 0.67   |
| RSS             | 2500| 1.14     | 0.054| 0.084| 0.19 | 0.58   |
| M.I.Permutation | 2500| 1.14     | 0.132| 0.206| 0.16 | 0.47   |
| M.I.RSS         | 2500| 1.14     | 0.020| 0.043| 0.12 | 0.37   |

**Table 2.5:** Simulation results for Model 3

| Method          | $P$ | $P_{GC}$ | FWER | FDR  | PSR | $|\hat{S}|$ |
|-----------------|-----|----------|------|------|-----|--------|
| ASYM            | 100 | 1.81     | 0.194| 0.0739| 0.483| 4.35   |
| GC              | 100 | 1.81     | 0.000| 0.0000| 0.282| 2.53   |
| Permutation     | 100 | 1.81     | 0.302| 0.0964| 0.611| 5.50   |
| RSS             | 100 | 1.81     | 0.000| 0.0000| 0.075| 0.68   |
| M.I.Permutation | 100 | 1.81     | 0.262| 0.0824| 0.545| 4.90   |
| M.I.RSS         | 100 | 1.81     | 0.004| 0.0025| 0.088| 0.80   |
| ASYM            | 500 | 1.63     | 0.220| 0.1276| 0.398| 3.58   |
| GC              | 500 | 1.63     | 0.010| 0.0027| 0.243| 2.19   |
| Permutation     | 500 | 1.63     | 0.344| 0.1563| 0.530| 4.77   |
| RSS             | 500 | 1.63     | 0.052| 0.0171| 0.370| 3.33   |
| M.I.Permutation | 500 | 1.63     | 0.302| 0.1384| 0.470| 4.23   |
| M.I.RSS         | 500 | 1.63     | 0.046| 0.0161| 0.316| 2.84   |
| ASYM            | 2500| 1.53     | 0.240| 0.1713| 0.334| 3.01   |
| GC              | 2500| 1.53     | 0.004| 0.0020| 0.184| 1.65   |
| Permutation     | 2500| 1.53     | 0.362| 0.2092| 0.440| 3.96   |
| RSS             | 2500| 1.53     | 0.084| 0.0366| 0.318| 2.86   |
| M.I.Permutation | 2500| 1.53     | 0.312| 0.1862| 0.378| 3.40   |
| M.I.RSS         | 2500| 1.53     | 0.062| 0.0262| 0.265| 2.38   |

**Table 2.6:** Simulation results for Model 4
| Method                | $P$  | $\lambda_{GC}$ | FWER  | FDR   | PSR   | $|\hat{S}|$ |
|-----------------------|------|----------------|-------|-------|-------|---------|
| ASYM                  | 100  | 1.13           | 0.028 | 0.0157| 0.42  | 1.26    |
| GC                    | 100  | 1.13           | 0.006 | 0.0051| 0.35  | 1.04    |
| Permutation           | 100  | 1.13           | 0.056 | 0.0300| 0.47  | 1.42    |
| RSS                   | 100  | 1.13           | 0.000 | 0.0000| 0.19  | 0.57    |
| M.I.Permutation       | 100  | 1.13           | 0.054 | 0.0339| 0.40  | 1.19    |
| M.I.RSS               | 100  | 1.13           | 0.000 | 0.0000| 0.13  | 0.38    |
| ASYM                  | 500  | 1.05           | 0.048 | 0.0368| 0.29  | 0.86    |
| GC                    | 500  | 1.05           | 0.026 | 0.0221| 0.26  | 0.78    |
| Permutation           | 500  | 1.05           | 0.098 | 0.0720| 0.34  | 1.02    |
| RSS                   | 500  | 1.05           | 0.042 | 0.0367| 0.30  | 0.91    |
| M.I.Permutation       | 500  | 1.05           | 0.072 | 0.0648| 0.27  | 0.80    |
| M.I.RSS               | 500  | 1.05           | 0.018 | 0.0224| 0.21  | 0.62    |
| ASYM                  | 2500 | 1.04           | 0.040 | 0.0533| 0.20  | 0.59    |
| GC                    | 2500 | 1.04           | 0.018 | 0.0296| 0.17  | 0.51    |
| Permutation           | 2500 | 1.04           | 0.084 | 0.0920| 0.25  | 0.75    |
| RSS                   | 2500 | 1.04           | 0.040 | 0.0511| 0.23  | 0.68    |
| M.I.Permutation       | 2500 | 1.04           | 0.072 | 0.1099| 0.18  | 0.53    |
| M.I.RSS               | 2500 | 1.04           | 0.022 | 0.0366| 0.14  | 0.42    |

Table 2.7: Simulation results for Model 5
In GWAS, the Bonferroni correction adjusts P-values for multiple testing. It is a conservative method to control the Family-wise Type I Error Rate (FWER). Here we propose a dimension reduction method to reduce the number of tests and therefore improve the multiple testing power. The method assumes a latent variable model and use Independent Component Analysis (ICA) to extract independent components as estimation of those latent variables. By targeting the haplotype blocks, we are able remove large amount of the redundancy information. Haplotype Blocks are blocks of SNPs in strong linkage-disequilibrium (LD) on the genome. When applied to the iCOGS breast cancer dataset with European ancestors, the method reduces 210935 SNPs to 57403 independent components (ICs). We found 26 loci associated with breast cancer using those ICs in addition to the standard GWAS approach and literature findings.

3.1 Introduction

In Genome Wide Association Studies (GWAS), multiple testing is the core statistical method to search for phenotype associated genetic variations. In this chapter, we develop a dimension reduction method on correlated genetic variants. The Bonferroni correction adjusts P-values for multiple testing. It is
3.1. Introduction

a conservative method to control the Family-wise Type I Error Rate (FWER) based on number of tests. Linkage Disequilibrium (LD), correlation between genetic variations, is a genome wide phenomenon in human being. Testing all the SNPs disregarding LD will lead to over-correction for multiple testing, and therefore weaken the power. Therefore, testing less number of less dependent variants would lower the Bonferroni correction threshold and improve multiple testing efficiency and variable selection power.

To improve the efficiency, the simplest strategy is to test only “tag” SNPs. A “tag” SNP serves as a surrogate for the haplotype block it lies in. A haplotype block is a small region on the genome where the genetic variants lies in it are highly correlated. It can be selected by comparing pair-wise LD score (i.e. \( r^2 \)) with base pair distance constrained [91]. This approach significantly reduces the number of tests. However, some information within haplotype blocks would be lost. As an alternative, we may still test for all the SNPs, but use an effective number of independent tests for Bonferroni correction instead. For example, simpleM [40] counts number of principal components that captures majority of the genetic variation, and \( K_{eff} \) [86] estimates FWER based on pairwise correlations between SNPs. Another approach is to test the joint effects of SNPs within some pre-defined SNP sets. SNP sets can be genes, haplotype blocks, etc. For each SNP set, a statistical model is built and estimated. Then these models are tested against the null model i.e. no statistical association between the set of SNPs and the phenotype. In this case, Bonferroni correction only performs on number of SNP sets instead of number of SNPs. Logistic kernel machine trains a logistic additive model including a kernel on SNPs within some SNP set. By choosing different kernel, it can test linear effects, interaction effects (epistasis) and nonlinear effects within SNP sets [70, 129]. Logistic kernel machine is equivalent to generalized mixed
3.1. Introduction

regression model[71]. Essentially, it treats the kernel of SNPs as random effects. SNP-based biological pathway analysis also performs association analysis on SNP sets [120, 121]. Pathway analysis focuses on joint effect of set of genes with biological connection. Since we only search for individual locus associated with the phenotype in this paper, pathway based methods is outside the scope of this chapter.

In this chapter, we adopt error correction code(ECC) hypothesis on haplotype blocks and apply Independent Component Analysis(ICA) to filter out redundant information between related genetic variants. This is essentially a dimension reduction technique to reduce the number of genetic variants.

ECC originates in communication theory. When a message is transmitted through a noisy communication channel, its information would be contaminated. ECC inserts parity bits(redundancies) into the original message making itself more robust to communication noise. The GWAS ECC hypothesis states that genetic variants are messages with ECC that transmitted through generations. We create a genomic communication model under this hypothesis. In this model, SNPs are ECC of genetic messages. From population point of view, ECC can be used to explain linkage-disequilibrium(LD) structure observed in human genome. Furthermore, ECC suggests genetic messages are latent variables of the SNPs rather than SNPs themselves. We call these latent variables genetic switches, and SNPs are linear combination of genetic messages plus communication noise. Using the latent variables directly for association analysis would greatly reduce the dimension of the problems, and filter out noise. Our aim is to induce those latent variables from SNPs. Instead of decoding SNPs along the genome, we focus on haplotype blocks. Haplotype blocks here refer to blocks of SNPs in strong LD on the genome. They are the regions with excessive redundancies and mostly likely to be ECC.
3.1. Introduction

Figure 3.1: Manhattan Plot of iCOGS breast cancer dataset

The Y-axes is $-\log_{10} P$-value, the horizontal dashed line is the Bonferroni threshold for genome wide significance. The red points denote the genome wide significant haplotype blocks selected by RSS IC.

We choose to use Independent Component Analysis (ICA) to extract the latent variables within the haplotype blocks. We call the extracted features independent components (ICs). ICA is a useful technique to separate linear mixtures of independent signals [59]. We show that Principal Component Analysis (PCA) is able to extract most information of genetic switches from SNPs given number of ICs. Then we apply the fastICA algorithm [58] to further separate the uncorrelated signals into statistically dependent ICs. In short, ICs are representatives of SNPs in haplotype blocks. They can be used as covariates in the association analysis. As the sizes of haplotype blocks is still small, usually under 200k base pair, each haplotype block can still point to the underlying genes on the genome. ICA under ECC hypothesis can be regarded as a dimension reduction method for GWAS.

We apply this dimension reduction method to the iCOGS breast cancer dataset with European ancestors. iCOGS dataset belongs to a large cohort study COGS [16]. The dataset has 87063 individual in total. Applying ICA under the ECC hypothesis, the number of association tests is reduced to
3.1. Introduction

57403 ICs from 210935 SNPs. Comparing to standard SNP approach, our ICA approach is able to capture most of the peaks on the Manhattan plot (Figure 3.1), but also discovers some additional loci comparing to the standard approach (Figure 3.8). We have found that additional 26 suspicious loci (Table 3.1) are associated with breast cancer apart from standard GWAS approaches and literature findings.

We also create two simulation studies to evaluate the reliability of iCOGS dataset results. Firstly, we study the performance of ICA under ECC hypothesis. We simulate one causal haplotype where SNPs in it are generated from linear combinations of a few latent variables. The phenotype is associated with one of the latent variables. We find that, in the casual haplotype block, ICs have smaller P-values than SNPs with both adjusted for multiple testing, and the lead IC (the IC with the smallest P-value) have better prediction power (AUC) on the testing dataset than the lead SNP. Furthermore, the number of ICs within a block is insensitive to the genotyping density. When the number of SNPs increases within the block, ICA approaches enjoy better prediction power and suffers less penalization from the multiple testing than the standard SNP approach. This favors ICA approaches when a GWAS using dense genotyping SNP arrays.

Since ECC hypothesis is difficult to test, in the second simulation study, we then examine the robustness of ICA method when the hypothesis is false, i.e. SNPs themselves are directly associated with the phenotype not the latent variables. We choose some random subsample of the iCOGS genotype dataset as covariates and generate the phenotype from a random selection of SNPs. We find that the ICA approach is more conservative than the standard approach. It selects relatively less causal haplotype blocks. A haplotype block is casual if at least one of the SNP in it is casual. FIXME and much
3.2. A population genomics communication model

less non-causal haplotype blocks correlated to casual one than the standard approach.

In summary, we found ICA is a reliable method to remove redundancies and preserves signals in the real GWAS dataset. When SNPs are directly associated to the phenotype (2nd simulation study), ICA approaches are expected to be more conservative than the standard approach, i.e. less number of variables found significant. However, in the iCOGS results, despite that substantial amount of genetic variations are removed, ICA approaches select as many associated regions as the standard SNP approach. ICA approaches also are also able to recover a few of the significant SNPs reported in the literature in addition to the standard approach. These all suggest that ECC on haplotype blocks is a useful working hypothesis. It can be used to aid the standard GWAS approach to search for additional findings. At last, it is worthwhile to further investigate the true underlying functional linkage between SNPs in strong LD.

3.2 A population genomics communication model

In GWAS, DNA sequences can be regarded as genetic messages passing through generations. In human, we observe strong linkage disequilibrium structures over the genome. This suggests large amount of redundancy inside the genetic messages. A communication model would help us understand the flow of genetic information, and therefore be used to extract crucial and compact genetic information for our association analysis.
3.2. A population genomics communication model

An iterative communication model. This is a communication model that integrates genetic variants and phenotypes in GWAS framework. It also emphasizes the communication noise accumulated when genetic messages are transmitted through generations. The key argument here is that the redundancies we observed among human genomes are byproducts of some mechanism fighting this information corruption. Our aim is to squeeze out redundancies and filter redundancies in the genetic messages through the communication model. Consequently, we are able to test the more compact extracted information for association analysis to the phenotype.

Section 1.3 introduces relevant works on modeling genetic information using communication models in the information theory. Their aim is to reverse engineer the DNA-RNA-protein translation process and unfold the biological meaning of DNA sequences. However, in GWAS we only need to model the genetic variations with respect to the population rather than the exact biological meaning of the message.

To simplify the problem, we adapt the GWAS convention that assumes the SNPs capture all the genetic variations in the population. This assumption rules out insertion and deletion, etc. Let \( x = (X_1, \ldots, X_m)^T \) be random vector of SNPs of an individual, then the population genetic information is defined by the collection of SNP random vectors of individuals.

Here, we introduce a GWAS communication model(Figure 3.2). The model describes a flow of genetic information(SNPs) through generations. To begin with, we have a generation consisting of individuals. Each individual is characterized by a collection of countable phenotypes, \( \{Y_1, Y_2, \ldots\} \). Phenotypes may be associated with the SNPs carried by themselves, or the environmental influence, or both of them. We assume that the collection of phenotypes capture all the information of an individual with respect to the population,
and then embryos of the next generation is created through mating. We call it the replication stage. Replication here refers to copying and remixing the population genetic information to the next generation through recombination process. Then the next generation develops from the embryos. We use a “black box” genetic decoder to describe how genetic information within the embryos is used to build up new individuals. This model is iterative and absent of any encoder. Individuals themselves are the outcome of the genetic messages, and also serve as the media. Furthermore, they are responsible for making new generations and therefore pass on the messages.

**Figure 3.2:** GWAS communication model.

**Error correction code model.** As discussed in Section 1.3, noise is unavoidable and essential to evolution. In an ideal noise-free model, if we assume Hardy-Weinberg Equilibrium (HWE) then the allele frequencies of SNPs remain constant between generations. So does the total entropy of the population genotypes. This means that the model is sustainable and able to transmit the same genetic information perfectly throughout generations. However, communication noise would break this equilibrium. We argue that error correction is a necessary mechanism when communication noise is taken into
3.2. A population genomics communication model

Here we setup a latent variable model that introduces Error Correction Code (ECC). Under the latent variable model, the genetic variants we observed are linear combinations of the genetic latent variables. The model could explain the redundancies we observed in the genetic variants. Thanks to those redundancies, communication errors can be corrected in the decoding stage.

In the ECC model, we assume DNA sequences are the channel code of the underlying genetic source code. First of all, we assume the phenotype is associated with the genetic switches through a generalized linear regression model, i.e.

\[ g \left( EY \mid \mathbf{u} \right) = \beta^\top \mathbf{u} \]  

(3.1)

where, \( \mathbf{u} \) of size \( q \times 1 \) (\( q \leq p \)) is the source code vector (genetic switches), \( g \) is a link function, \( \beta \) are some vector of coefficient parameters.

Under the ECC model, \( \mathbf{u} \) is not directly coded in the DNA sequences. It is processed through an ECC encoder. We use \( \mathbf{x} \) to denote the channel code in the DNA sequences.

\[ \mathbf{x} = d( \mathbf{G} \mathbf{u} + \mathbf{e} ) \]  

(3.2)

where \( \mathbf{G}_{m \times q} \) is some unknown channel encoder matrix, \( \mathbf{e} = (e_1, \cdots, e_p)^\top \) is the communication error and \( d \) is an element-wise discretization function. Take SNP as an example, \( d \) can be in form of the sum of two indicating functions. If we rewrite \( \mathbf{v} = (V_1, \cdots, V_m)^\top = \mathbf{G} \mathbf{u} + \mathbf{e} \), then \( \mathbf{x} \) have the following form,

\[ \mathbf{x} = (\cdots, X_i, \cdots)^\top = d \left( (\cdots, V_i + e_i, \cdots)^\top \right) \]

\[ = (\cdots, \mathbf{1}(V_i > c_{i,1}) + \mathbf{1}(V_i > c_{i,2}), \cdots)^\top, \]

for some constants \( c_{i,1}, c_{i,2}, i = 1, \cdots, m \).
3.3 Independent Component Analysis on ECC

We further assume $G$ is a full-row-rank matrix. Otherwise, information would be lost. In channel coding theory and applications, ECC encoder generally take one of two forms, namely the linear block codes and convolutional codes. They are both linear operations over the original source code[102].

### 3.3 Independent Component Analysis on ECC

According to the previous model, the vector of SNPs $x$ is the discretization of continuous channel code vector $v$, and $v$ is some linear combination of $u$ plus noise $e$. The task of Independent Component Analysis (ICA) is recover the source code $u$, i.e. to find some independent components $s = Cv$ as an estimator of $u$. The components of $s$ is expected to be independent so that association analysis on them instead of components of $v$ would be more direct, and would yield higher power.

ICA has two steps, data whitening and maximizing non-gaussianity. The whitening pre-processes data through a linear transformation. It decorrelates the data, and also reduces the dimension of the data. Let matrix $K_{p \times k}$ denote the linear transformation. $z = K^\top v$ are the reduced vector with size $k$, and the covariance matrix of $z$, Cov$(z)$ is a diagonal matrix. Theorem 3.1 shows that if the columns of $K$ are the first $k$ principal component of $v$ with largest variances then the maximum mutual information $I(z, u)$ is preserved in the least square sense[22, 59].

The correlation between components of $z$ equals to 0 is only the necessary condition of independence. Although $z$ preserves most information of the $u$, each component of $z$ could still be a linear combination of $u$. Testing the associations between components of $z$ and $Y$ would still be sub-optimal. ICA algorithms searches for some matrix $B_{k \times k}$ so that the components $s =$
Bz would be least dependent on each other. In the end, we can use the independent components $s = Cv = BK^t v$ for association analysis.

### 3.3.1 Whitening

Whitening is the data pre-processing stage in PCA. It uses Principal Component Analysis (PCA) to decorrelate and reduce the dimension of $v$. After whitening, $z$ is a reconstruction of $v$ with its correlation matrix being diagonal. Principal Component Analysis (PCA) is a statistical procedure that converts a vector of variables into a vector of uncorrelated variables through linear transformation. The covariance matrix of $v$ is

$$Q_v = E\left((v - E(v))(v - E(v))^\top\right). \tag{3.3}$$

$Q_v$ is positive definite. PCA performs singular value decomposition on it, i.e.

$$Q_v = W\Lambda W^\top, \tag{3.4}$$

where $W$ is an orthogonal matrix with $W^\top W = I$, and $\Lambda$ is a diagonal matrix filled with the eigenvalues of $Q_v$, $\lambda_1, \ldots, \lambda_p$. The $k$th column of $W$ is the $k$th eigenvector of $Q_v$ corresponding to eigenvalue $\lambda_k$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. The columns of $Wv$ are referred to principal components in PCA. The columns of $Wv$ are called principal component scores of $v$. According to the optimal reconstruction principal of PCA [22] (Section 3.1.2), if $K$ is consistent of first $k$ columns of $W$, then the mean square error $||v - z||^2$, is minimized with respect to $k$. In this sense, $z$ is the best reconstruction of $v$ with dimension reduced to $k$.

Here we are interested in whether $z$ can maximize the mutual information $I(z, u)$. If the error term $e$ does not exist, then $v$ is just linear combination of $r$ independent variables and the linear space of $v$ has rank $r$ as well. $I(z, u)$
3.3. Independent Component Analysis on ECC

will be maximized if \( k \geq r \). When we add the noise \( e \) into the equation, we consider to minimize the information loss \( \Delta I = I(v, u) - I(z, u) \). Since \( I(v, u) \) is independent of \( K \), it is equivalent to maximize \( I(z, u) \). Theorem 3.1 shows that if \( u \) is normally distributed, then \( \Delta I \) can be minimized with respect to \( k \); if \( u \) is not normally distributed, then the a upper bound of \( \Delta I \) can be minimized. The upper bound is based on the normal approximation of \( v \) with the same covariance matrix.

**Theorem 3.1.** Let \( z = K^\top v \). \( K \) is some full column rank \( m \times r \) matrix and \( K^\top K = I_k \). \( v \) is decomposed into the projection on to the space spanned by \( z \), \( v_z \), and the compliment \( v_c \). Suppose \( e \) is a vector of i.i.d normally distributed random variables, then we have

- \( H(v_c) + \frac{m-k}{2} \log(2\pi e) \) is an upper bound of the information loss, \( \Delta I = I(v, u) - I(z, u) \).

- If the columns of \( V \) are the first \( r \) eigenvectors correspond to the first \( r \) eigenvalues with largest absolute values then
  
  - If \( v \) is multivariate normal, \( H(v_c) \) will be minimized.
  
  - If \( v \) is not multivariate normal, let \( \nu \) be the normal approximation of \( v \) with same covariance matrix, then \( H(\nu_c) \geq H(v_c) \) will be minimized.

**Proof.** Section 3.2.2 in [22] shows that when information source is contaminated by white noise, PCA is able to minimize the upper bound of information loss. Here based on their work, we show that the result is still valid when the information source is first encoded by an ECC encoder matrix \( G \) and then contaminated by white noise \( e \).
Since $G$ is a deterministic function, $I((v, z), u) = I(z, u)$. $\Delta I$ can be rewritten as $I((v, z), u) - I(v, z)$. Then we decompose $v$ into the subspace span by $z$ and its compliment.

$$v = v_z + v_c,$$

where $v_z = V(V^T V)^{-1} z$. We can write $v_z = P v$ for some projection matrix $P = V(V^T V)^{-1} V^T$. [89] shown that

$$H(v, z) - H(v, z, u) = H(v_c, z) - H(v_c, z, u)$$ \tag{3.6}

Then we have

$$\Delta I = I((v, z), u) - I(z, u)$$

$$= H(v, z) + H(u) - H(v, z, u) - H(z) - H(u) + H(z, u)$$

$$= H(v_c, z) - H(v_c, z, u) - H(z) + H(z, u)$$

$$= H(v_c|z) - H ((I - P)(Gu + e)|K^T(Gu + e), u)$$

$$= H(v_c|z) - H ((I - P)e|K^T e, u)$$

$$= H(v_c|z) - H ((I - P)e)$$

$$= H(v_c|z) - \frac{m - k}{2} \log(2\pi e)$$

$$\leq H(v_c) - \frac{m - k}{2} \log(2\pi e)$$ \tag{3.7}

Note that $(I - P)$ is a projection matrix itself. $v_c$ and $z$ belongs to kernel and range of $P$ respectively. If $v$ is normal, then $v_c$ and $z$ are independent. In other words, the upper bound (3.7) equals to $\Delta I$ if $v$ is multivariate normal. Similarly, $(I - P)e$ and $K^T e$ are independent as well since $e$ is multivariate normal. $\frac{1}{2} \log(2\pi e)$ is the entropy of a standard normal distribution.

PCA minimizes the upper bound of $H(v_c)$. Since PCA minimizes the mean square error of $v_c$, the entropy of $v_c$ is would be minimized if $v$ is
multivariate normal. When \( v \) is not multivariate normal, \( H(\nu_c) \) is an upper bound of \( H(v_c) \) and it is minimized by PCA. Therefore, the upper bound \( \Delta I \) would be minimized by KPCA.

**Choice of \( k \).** \( k \) controls how much information is filtered out in the orthogonal space defined by \( W \). Large \( k \) will preserve more information of \( u \) but also includes noise \( e \). On the contrary, small \( k \) will remove both the noise and the information of \( u \). If the signal noise is greater than one, then we expect that the variances of principle components corresponding to \( Gu \) are greater than the variances of components of \( e \). Therefore we can use a mixture model to model \( \lambda \) with a mixture of one uniform and distribution and one normal distribution. In the ECC model, we assume components of \( e \) have equal variances. The normal distribution with a lower mean and lower variance models the components of \( \lambda \) corresponding to \( e \). The uniform distribution with a higher mean and variance models the components of \( \lambda \) corresponding to \( u \).

We can use EM algorithm to find an estimator of \( k \) given \( \lambda \).

**Tightness of the lower bound.** The tightness of the bound can not be determined in that distribution of \( u \) is unknown. However, if the distribution of \( v \) is distant to normal distribution then it is less likely for it to be a mixture of \( u \). According to central limit theorem, in a board sense, the distribution of mixture of independent variables tends to be normally distributed. In this sense, Theorem 3.1 gives a tight bound in the worst scenario when signals are fully mixed. In the following subsection, we will further discuss the connections between mixture of independent variables and normality.
3.3 Independent Component Analysis on ECC

3.3.2 ICA: maximize the non-gaussianity

By choosing an appropriate $k$, whitening allow us to decorrelate $v$ and remove certain degree of white noise. However, $z$ could still be a mixture of $u$ and be statistically dependent on zero correlation. Here we seek some orthogonal rotation matrix $B_{k \times k}$ so that $s = (S_1, \cdots, S_k)^\top = Bz$ would least likely to be a mixture of genetic switches. Through rotation, we further reduce the dependence between components of $z$. This would help further boost the power of multiple testing in GWAS.

We use Independent Component Analysis (ICA) to separate underlying signals in $z$. ICA separate independent variables from their linear combination by maximizing the non-gaussianity. According to central limit theorem, the sum of i.i.d. random variable with finite variance is asymptotically normally distributed. When mixing independent random variables with finite variance, in a board sense, the distribution would become more like Gaussian than their original distributions. Intuitively, the farther a variable is from a Gaussian distribution, the less likely it is a mixture of independent variables. Non-gaussianity describes the distance between a distribution and its Gaussian approximation with equal mean and variance. Neg-entropy is a measure of non-gaussianity.

In single variant case, for random variable $S$, let $f_S$ be its density function, the differential entropy of $S$ is

$$H(S) = - \int f_S(\eta) \log f_S(\eta) d\eta. \quad (3.8)$$

And the neg-entropy of $S$, $J(S)$ is

$$J(S) = H(\nu_S) - H(S). \quad (3.9)$$

$\nu_S$ is the normal approximation of $S$ with the same mean and variance. Let
3.3. Independent Component Analysis on ECC

\[ S = b^\top z \] for some vector \( b \). We can find one independent component of \( z \) by maximizing the neg-entropy.

For vector \( s \), we can find \( s \) by search for matrix orthogonal \( B \) that maximum the sum of the marginal entropy of \( s \) i.e.

\[
\hat{J}(s) = \sum_{i=1}^{m} H(\nu_{S_i}) - \sum_{i=1}^{m} H(S_i).
\] (3.10)

Maximize the sum of marginal neg-entropy is directly linked to minimize the mutual information between components of \( s \). The mutual information within \( s \) is defined as

\[
I(s) = I(S_1, \cdots, S_k) = \sum_{i=1}^{k} H(S_i) - H(S_1, \cdots, S_k)
\] (3.11)

Note that \( |\det B| = 1 \) and rotating with \( B \) preserves entropy. Both the entropy \( s \) and \( \nu_s = (\nu_{S_1}, \cdots, \nu_{S_k})^\top \) are invariant subject to \( B \). Maximizing \( \hat{J}(s) \) is equivalent to minimizing \( \sum_{i=1}^{k} H(S_i) \) and therefore is equivalent to minimizing \( I(s) \). By rotating \( z \), we seek matrix \( B \) that minimizes the dependency between components in \( s \).

\( H(S) \) would be hard to compute, since we have no assumption on the distribution of \( u \) at the first place. We can use the following approximation[58] of \( J \),

\[
J(S) \approx [E\{G(S)\} - E\{G(\nu)\}]^2,
\] (3.12)

where \( G \) could be either \( G_1 \) or \( G_2 \) with

\[
G_1(s) = \frac{1}{a_1} \text{logcosh}(a_1 s)
\] (3.13)

\[
G_2(s) = -\exp(-\frac{s^2}{2}),
\] (3.14)

and \( 1 \leq a_1 \leq 2 \) is a tuning parameter. Then (3.10) can be approximated by :

\[
\sum_{i=1}^{k} \{E\{G(s_i)\} - E\{G(\nu_i)\}\}^2.
\] (3.15)
To maximize (3.15) with respect to $B$, we can use a gradient descent method. The gradient would be

$$\frac{\partial \hat{J}(s)}{\partial b_i} = 2\alpha_i E \{ z g(b_i^T z) \},$$

where $b_i$ is the $i$ column of $B$, $\alpha_i = E \{ G(b_i^T z) - G(\nu_i) \}$ and $g$ is the derivative of $G$. In addition to the standard gradient decent algorithm, another symmetric orthogonalization step is necessary[59]. This step can be done through:

$$B \leftarrow (BB^T)^{-\frac{1}{2}} B.$$  \hspace{1cm} (3.17)

Let $BB^T = Q^T \Lambda Q$, where $Q$ is some orthogonal matrix, and $\Lambda$ is a diagonal matrix with $(\lambda_1, \ldots, \lambda_r)$. Then $(BB^T)^{-\frac{1}{2}} = P^T \Lambda^{-\frac{1}{2}} P$, and $\Lambda^{-\frac{1}{2}}$ is a diagonal matrix with value $(\lambda_1^{-\frac{1}{2}}, \ldots, \lambda_r^{-\frac{1}{2}})$. (3.17) ensures in each iteration $s$ would be decorrelated and converge to different underlying genetic switches. The algorithm can be summarized as follows:

1. Set number of independent component to estimate to be $r$.
2. Initialize $B = (b_1, \ldots, b_r)^T$.
3. For $i = 1, \ldots, r$,
   a) Gradient decent: $b_i \leftarrow b_i - \frac{\partial \hat{J}(Bz)}{\partial b_i}$
   b) Normalization: $b_i \leftarrow \frac{b_i}{||b_i||}$
4. symmetric orthogonalization: $B \leftarrow (BB^T)^{-\frac{1}{2}} B$
5. If $B$ does not converge, go to step 3.

Details can be found in Chapter 9 of [59]. Software FastICA[58] is an efficient implementation of ICA, and will be used in this chapter.
3.3. Extracting from haplotype blocks

In this chapter, we propose to perform ICA at the haplotype block level on the genome where the LD is in its greatest strength. Haplotype block is a small region on genome where SNPs inside it are in strong LD to each other. A haplotype block is usually less than $200\text{k}$ base pair wide\cite{91}. Degree of LD is directly linked to redundancy level of the genetic messages within the haplotype block. Applying ICA on the haplotype block level is an efficient way of removing redundancies. As an alternative, we may apply it genome-wide, at chromosome level, at pathway level, and at gene level.

For association analysis, testing independent components (ICs) extracted from haplotype blocks for association would give us several advantages. The total number of hypothesis tests would be significantly reduced. As haplotype blocks are still relatively small on the genome, we can pinpoint phenotype associated locus on the genome. It also gives us better interpretation of the associations within a haplotype block. In a Manhattan plot, for example Figure 3.1, we search for SNPs with pikes of P-values piercing through the Bonferroni threshold. The pikes are results of strong local correlation. The reason of this practice is that human genome is in massive linkage disequilibrium (LD) structures. Even if a casual variant is not directly genotyped, its variation may still be captured by the nearby genotyped SNPs due to the LD. If a significant SNP is not in any LD structure, i.e. the surrounding SNPs are not significant, then it is more likely to be noise or some artifacts. In common practice, only the most significant SNPs in the spikes is reported. We refer to this as the lead SNP approach. However, we can not distinguish whether the other SNPs are significant simply due to LD or they are actually carrying other less important information. On the other hand, we can test different ICs extracted from this...
3.3. Independent Component Analysis on ECC

haplotype block separately. Since ICs are expected to be independent of each other, we won’t have this confusion.

At last, when applying ICA, we may use SNP $X$ directly as $\hat{V}$. The value of $X$ denotes the number of minor allele an individual has at a locus. In the logistic regression, this approach is equivalent to a multiplicative risk model. Essentially, the model assumes having two minor alleles instead of one would double the log odds ratio of having the disease. The multiplicative risk model is widely accepted in the GWAS [126, 127].
3.4 Results on iCOGS breast cancer dataset

The Collaborative Oncological Gene-environment Study (COGS)\textsuperscript{a} was funded by the European Commission and it initiated its 7th Framework Programme in May 2009. COGS aims to identify both genetic and environmental risk factors in three hormone-related cancers, i.e. breast cancer, ovarian cancer, and prostate cancer. At the time, it was the largest GWAS project including more than 200,000 individuals. It consists of seven consortia, among which the largest three are the Breast Cancer Association Consortium (BCAC)\textsuperscript{b}, the Ovarian Cancer Association Consortium (OCAC)\textsuperscript{c} and the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL)\textsuperscript{d}. Among the consortia, BCAC alone has a collection of 45 studies from all over the world. iCOGS is a custom built genotyping array in COGS\textsuperscript{e}. It was designed to extend the knowledge of the associations between the genetic variants with three hormone-related cancers: breast, ovarian and prostate cancers. The whole project genotyped more than 250,000 subjects. iCOGS chips includes over 200,000 SNPs.

In this chapter, we only focus on the dataset on the breast cancer and patients with European ancestors. Within that dataset, one patient has a

\textsuperscript{a} Collaborative Oncological Gene-environment Study. Collaborative Oncology Gene-environment Study. URL: http://www.cogseu.org/ (visited on 09/01/2016).
\textsuperscript{b} Breast Cancer Association Consortium. Genetic Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, UK.. URL: http://apps.ccge.medschl.cam.ac.uk/consortia/bcac/index.html (visited on 09/01/2016).
\textsuperscript{c} Ovarian Cancer Association Consortium. Genetic Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, UK.. URL: http://apps.ccge.medschl.cam.ac.uk/consortia/ocac/index.html (visited on 09/01/2016).
\textsuperscript{d} Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome. Centre for Cancer Genetic Epidemiology, University of Cambridge, UK.. URL: http://practical.ccge.medschl.cam.ac.uk (visited on 09/01/2016).
\textsuperscript{e} O. G. Bahcall. COGS project and design of the iCOGS array. 2013. URL: http://www.nature.com/icogs/primer/cogs-project-and-design-of-the-icogs-array/ (visited on 04/12/2015).
missing phenotype, and is therefore removed. The base pair position of SNPs is based on NCBI36/hg18(Mar 2006)\textsuperscript{f}. Some SNPs share the same base pair position and major allele. Only the first SNP is kept. SNPs that do not belong to chromosomes (with label 24) are removed as well. After the cleaning, 210935 SNPs are in the dataset with 87063 individuals (43947 cases, and 43116 controls).

### 3.4.1 Haplotype block estimation and IC extraction

Here we assume the Error Correction Code hypothesis on all haplotype blocks. Haplotype block is constructed according SNPs pairwise linkage disequilibrium and base-pair distance\cite{39}. This method is implemented in software PLINK\cite{91}. We choose $D'$ as the LD metric, and restrict the maximum block size to be 300kb, the minimum MAF to be 0.05. A pair of SNPs are in strong LD if $D'$ is greater than 0.90. The proportion of strong LD pairs within a haplotype block should be greater than 0.70. Each Haplotype block has minimum 2 SNPs. The SNPs outside haplotype blocks are ignored (None of the those SNPs has raw logistic regression P-value passing the Bonferroni correction threshold.) We obtained 28673 haplotype blocks in total. Then we apply ICA on each haplotype block. Number of ICs in each block is determined by fitting a mixture model of Gaussian and uniform distribution on the variances of PCA principle components. 57403 ICs are extracted from all haplotype blocks. This reduces the dimensions to of the multiple testing problem to 27.2%. Figure 3.3 summarizes meta information of haplotype blocks and extracted ICs. The majority of haplotype blocks only have less than 10 SNPs. This is due to the limited size of iCOGS array. The mode of haplotype block widths is around

10k base pair. Most of ICs explained more than 80% of the total variance of haplotype blocks. The ICs over SNPs ratio is similar between chromosomes.

Figure 3.3: Summary of haplotype blocks and independent components
(a) Histogram of log10 haplotype block sizes in SNPs. (b) Histogram of log10 haplotype block width in number of base pairs. (c) Histogram of proportion of variance explained by extracted independent components within each haplotype block. (d) Number of features within each chromosome.

3.4.2 Population stratification analysis
In the statistical analysis, we test for the difference in allele frequencies between cases and controls. However, the difference would be influenced by systematic differences of frequencies in case and control samples. This possible confounding factor is referred to as population stratification. Population stratification would inflate the tails of null distribution of the test statistics,
3.4. Results on iCOGS breast cancer dataset

Figure 3.4: Population structure of iCOGS breast cancer dataset

X and Y axes are first two principal components of the genotype dataset. The color of each point denotes one sub-study. 40 sub-studies in total.

and cause P-values to be smaller than that is expected under the asymptotic null hypothesis.

Despite of the huge sample size of iCOGS, population stratification is a
major concern because of the 40 sub-studies contributed to the dataset. To illustrate the population structure, we perform principal component analysis on the whole genotype dataset. Figure 3.4 plots the dataset on the first two principal component cores. The various colors denote different sub-studies. Whilst the butterfly pattern looks beautiful, it causes problems. The genomic control inflation $\lambda$ for multiple testing is 2.14 for SNPs and 1.92 for ICs. This suggests very strong population stratification. It should be corrected before further association analysis.

### 3.4.3 New findings using ECC and RSS

We use logistic regression model on ICs and multiple testing with Rank Stability Selection (RSS) to search for locus associated with the breast cancer. We first fit a logistic regression model (3.18) for each IC we have identified. Instead of performing a Wald test on the coefficient $\beta_{i,1}$ for the $i$th IC, we perform RSS for all the $\beta_{i,1}$ of ICs in one chromosome at a time and calculate their corresponding $P$-values. As shown in Section 2.3, RSS is a non-parametric resampling technique for multiple testing. It has no assumptions on the distribution of $\beta_{i,1}$, and is robust to population stratification as a confounding factor.

$$\logit E(Y|IC_i) = \beta_{i,0} + \beta_{i,1} IC_i.$$ (3.18)

As the result, we found 89 haplotype blocks containing ICs associated with the breast cancer with genome-wide significance. Among those regions, we identified 24 loci that could neither be found using SNPs nor have been found in the literature[85]. Details of those haplotype blocks are listed in Table 3.1.

Those 24 new findings can be explained by two level of improvements of our method. The first level is that some genetic communication noise is filtered out during the ECC dimension reduction stage by ICA. We can see the
improvements through the P-values between standard methods on SNPs (GC PCA SNP P) and standard method on ICs (GC PCA IC P). The second level is we use the stability of the signal rather than the magnitude to determine its significance. Although noise is reduced, signal may be weakened by ICA as well. RSS takes advantage over the increase in the signal noise ratio. This can be seen through the P-values between standard methods on ICs (GC PCA IC P) and RSS on ICs (RSS IC P).

All those 24 SNPs are close to Hardy-Weinberg Equilibrium 3.6[91]. We map the SNPs to 14 Genes using dbSNP[109]. In the gene ontology enrichment analysis[84], no gene set is found significant. In the pathway analysis, ESR1 and MAP2K3 Genes has appeared on the KEGG breast cancer pathway[63]. We have also looked up the human protein atlas[116], and find that gene RAB3GAP1, CNTN6, MAP2K3, ESR1, ACTG1 have evidence associated with breast cancer cells at protein level.

In the following section, we compare our approach with the standard GWAS procedure. We find that RSS on ICs is relatively conservative on population stratification correction; RSS on ICs also shares a large proportion of findings with standard procedures and literature findings. This suggests RSS on ICs is a reliable method with its own characteristics. These 24 new findings found by ECC and RSS have a good chance to be truly linked to breast cancer.
<table>
<thead>
<tr>
<th>Lead SNP</th>
<th>chr</th>
<th>bp(k)</th>
<th>MAF</th>
<th>width(k)</th>
<th># SNPs</th>
<th>V.E.</th>
<th>GC PCA SNP P</th>
<th>GC PCA IC P</th>
<th>RSS IC P</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2311859</td>
<td>1</td>
<td>15360</td>
<td>0.141</td>
<td>1</td>
<td>2</td>
<td>0.087</td>
<td>8.5e-03</td>
<td>8.0e-11</td>
<td>2.7e-08</td>
<td>KAZN</td>
</tr>
<tr>
<td>rs6730157</td>
<td>2</td>
<td>135874</td>
<td>0.347</td>
<td>110</td>
<td>2</td>
<td>0.574</td>
<td>8.7e-01</td>
<td>4.8e-01</td>
<td>5.2e-09</td>
<td>RAB3GAP1</td>
</tr>
<tr>
<td>rs3213943</td>
<td>2</td>
<td>136468</td>
<td>0.170</td>
<td>157</td>
<td>2</td>
<td>0.613</td>
<td>6.6e-03</td>
<td>6.4e-02</td>
<td>8.3e-11</td>
<td>R3HDM1</td>
</tr>
<tr>
<td>rs6430298</td>
<td>2</td>
<td>148991</td>
<td>0.434</td>
<td>47</td>
<td>2</td>
<td>0.104</td>
<td>5.0e-06</td>
<td>1.4e-07</td>
<td>6.6e-08</td>
<td>MBD5</td>
</tr>
<tr>
<td>rs4698932</td>
<td>2</td>
<td>106494</td>
<td>0.482</td>
<td>1</td>
<td>2</td>
<td>0.515</td>
<td>2.2e-04</td>
<td>7.1e-08</td>
<td>1.1e-07</td>
<td>-</td>
</tr>
<tr>
<td>rs6830464</td>
<td>4</td>
<td>131605</td>
<td>0.409</td>
<td>2</td>
<td>5</td>
<td>0.178</td>
<td>3.6e-05</td>
<td>6.9e-08</td>
<td>4.6e-08</td>
<td>-</td>
</tr>
<tr>
<td>rs7443354</td>
<td>5</td>
<td>25880</td>
<td>0.067</td>
<td>4</td>
<td>4</td>
<td>0.099</td>
<td>3.2e-02</td>
<td>3.4e-13</td>
<td>6.6e-07</td>
<td>-</td>
</tr>
<tr>
<td>rs380346</td>
<td>5</td>
<td>58073</td>
<td>0.377</td>
<td>17</td>
<td>2</td>
<td>0.056</td>
<td>3.9e-04</td>
<td>1.1e-08</td>
<td>3.9e-07</td>
<td>RAB3C,LOC105378986</td>
</tr>
<tr>
<td>rs1543403</td>
<td>6</td>
<td>152429</td>
<td>0.488</td>
<td>1</td>
<td>3</td>
<td>0.998</td>
<td>4.8e-07</td>
<td>1.7e-07</td>
<td>4.3e-07</td>
<td>ESR1</td>
</tr>
<tr>
<td>rs4593472</td>
<td>7</td>
<td>130665</td>
<td>0.351</td>
<td>5</td>
<td>2</td>
<td>0.849</td>
<td>1.2e-04</td>
<td>2.3e-05</td>
<td>7.0e-07</td>
<td>LINC-PINT</td>
</tr>
<tr>
<td>rs10252584</td>
<td>7</td>
<td>134040</td>
<td>0.390</td>
<td>1</td>
<td>20</td>
<td>0.065</td>
<td>3.2e-01</td>
<td>1.3e-04</td>
<td>3.8e-08</td>
<td>-</td>
</tr>
<tr>
<td>rs720475</td>
<td>7</td>
<td>144077</td>
<td>0.248</td>
<td>5</td>
<td>2</td>
<td>0.120</td>
<td>3.1e-06</td>
<td>1.6e-05</td>
<td>3.7e-08</td>
<td>ARHGEF5</td>
</tr>
<tr>
<td>rs799890</td>
<td>8</td>
<td>117216</td>
<td>0.194</td>
<td>69</td>
<td>9</td>
<td>0.330</td>
<td>5.8e-05</td>
<td>1.8e-06</td>
<td>5.3e-09</td>
<td>LINC00536</td>
</tr>
<tr>
<td>rs67397162</td>
<td>8</td>
<td>129182</td>
<td>0.166</td>
<td>78</td>
<td>2</td>
<td>0.422</td>
<td>1.8e-06</td>
<td>1.9e-06</td>
<td>1.4e-09</td>
<td>-</td>
</tr>
<tr>
<td>rs10896050</td>
<td>11</td>
<td>65578</td>
<td>0.186</td>
<td>8</td>
<td>8</td>
<td>0.679</td>
<td>4.4e-06</td>
<td>5.2e-08</td>
<td>8.7e-06</td>
<td>-</td>
</tr>
<tr>
<td>rs4545589</td>
<td>11</td>
<td>87019</td>
<td>0.318</td>
<td>2</td>
<td>3</td>
<td>0.115</td>
<td>2.1e-06</td>
<td>2.2e-16</td>
<td>4.2e-09</td>
<td>TMEM135</td>
</tr>
<tr>
<td>rs2887022</td>
<td>12</td>
<td>96025</td>
<td>0.315</td>
<td>4</td>
<td>3</td>
<td>0.891</td>
<td>6.7e-07</td>
<td>6.6e-07</td>
<td>3.1e-06</td>
<td>-</td>
</tr>
<tr>
<td>rs2295882</td>
<td>14</td>
<td>91858</td>
<td>0.238</td>
<td>34</td>
<td>10</td>
<td>0.351</td>
<td>3.4e-06</td>
<td>5.3e-06</td>
<td>2.6e-07</td>
<td>CCD8C</td>
</tr>
<tr>
<td>rs8064086</td>
<td>16</td>
<td>67013</td>
<td>0.125</td>
<td>1</td>
<td>2</td>
<td>0.183</td>
<td>1.8e-05</td>
<td>1.7e-07</td>
<td>1.8e-05</td>
<td>-</td>
</tr>
<tr>
<td>rs735753</td>
<td>16</td>
<td>80644</td>
<td>0.223</td>
<td>6</td>
<td>7</td>
<td>0.857</td>
<td>3.4e-07</td>
<td>8.8e-08</td>
<td>1.5e-06</td>
<td>CDYL2</td>
</tr>
<tr>
<td>rs11651154</td>
<td>17</td>
<td>21203</td>
<td>0.381</td>
<td>10</td>
<td>11</td>
<td>0.086</td>
<td>4.1e-04</td>
<td>3.2e-09</td>
<td>1.3e-07</td>
<td>MAP2K3</td>
</tr>
<tr>
<td>rs7503278</td>
<td>17</td>
<td>79482</td>
<td>0.196</td>
<td>5</td>
<td>3</td>
<td>0.085</td>
<td>7.8e-02</td>
<td>2.4e-07</td>
<td>3.1e-06</td>
<td>ACTG1,FSCN2</td>
</tr>
<tr>
<td>rs635269</td>
<td>18</td>
<td>8407</td>
<td>0.414</td>
<td>4</td>
<td>6</td>
<td>0.081</td>
<td>7.6e-03</td>
<td>1.2e-06</td>
<td>8.3e-12</td>
<td>-</td>
</tr>
<tr>
<td>rs1175745</td>
<td>18</td>
<td>24571</td>
<td>0.392</td>
<td>110</td>
<td>2</td>
<td>0.803</td>
<td>1.6e-04</td>
<td>1.0e-04</td>
<td>1.7e-08</td>
<td>CHST9</td>
</tr>
</tbody>
</table>

Table 3.1: 24 additional haplotype blocks found significant using ecc

Haplotype blocks found significant by both GC PCA IC and rss ic are listed. The lead SNP is the SNP with smallest p-value in a haplotype block. width(k) is the base pair width of the block. #SNPs is the number of SNPs within the block. V.E. is the proportion of the variance explained by the selected IC. “P” standards for P-value for the corresponding method.
3.5 Comparison between methods on the iCOGS dataset

We further compare our RSS IC method with the standard GWAS approach with adequate population stratification control. We find that RSS IC is relatively conservative. However, RSS IC is able to find different associated regions in comparison with the standard approach.

3.5.1 Population stratification correction

To correct population stratification for multiple testing, we adopt and compare three methods, Genomic Control (GC) (Section 1.2.6), Principal Component analysis (PCA) (Section 1.2.6), and Rank Stability Selection (RSS) (Section 2.3). GC (Section 1.2.6) is the simplest method among all. It is easy to compute but has the weakest test power. The PCA method captures genetic variation shared within the population using its principal component scores. Adding the top few scores as covariates into the logistic regression model (3.19) would efficiently remove population stratification effect on $\beta_1$.

\[
\logit E(Y|X_i) = \beta_0 + \beta_1 X_i + \sum_{j=1}^{j+1} \beta_j PC_j
\]  

(3.19)

Both GC and PCA methods are commonly used in GWAS. They can be applied to both SNPs and ICs. On the other hand, as discussed in (Section 2.5.2), the LD structure between SNPs would break the exchangeable assumption of RSS. We will apply RSS to the ICs only.

To investigate the strength of population stratification control, we use genomic control $\lambda_{GC}$ as the measurement and P-value QQ plot as the illustration. Firstly, we compare their performance using SNPs. For PCA, the first seven
principal component scores are used as covariates in the logistic regression model for each SNP. PCA itself reduces $\lambda$ to 1.44 from 2.14, which is still not adequate. Therefore, applying GC over the PCA P-values is necessary. Figure 3.5 is the QQ plot of three methods’ P-values. PCA by itself is not much different from raw P-values. We can see PCA GC and GC both have $\lambda = 1$, but PCA GC have more discoveries pass the Bonferroni threshold. This makes PCA GC a sufficient and efficient method for population stratification correction.

Then we look at ICs. The raw ICs have $\lambda = 1.92$. Here we compare GC IC, GC PCA IC with RSS IC. Since RSS only calculates P-values for the top few ICs, its genomic control $\lambda$ is not computable. However, in Figure 3.6, the number of significant ICs found by three methods are close. This suggests that RSS is sufficient to control population stratification.

For further association analysis, we compare the results of GC PCA on SNPs(GC PCA SNP), GC PCA on independent components(GC PCA IC), with RSS on independent components(RSS IC) since they are adequate to correct the population stratification.

**3.5.2 Association analysis**

Here we compare the number of genome-wide significant loci founded by three methods; whether they share findings among each other; with how those findings are compared to the literature findings. We set the family wise error rate (FWER) to be 0.05 and the correspondence Bonferroni threshold for genome-wide significance is $2.37 \times 10^{-7}$ for SNPs and $9.14 \times 10^{-7}$ for ICs.

First as an overview of the findings, we look at the Manhattan plot (Figure 3.7) breaking down into chromosomes for three methods. We observe massive association signals and find that quite a few SNPs have low P-values(down to
3.5. Comparison between methods on the iCOGS dataset

Figure 3.5: QQ-plot of SNP P-values

X-axis is the expected P-value quantile under uniform distribution. Y-axis is the empirical quantile of the observed P-values. The black line indicate the observed P-value under null hypothesis. The grey horizontal line is Bonferroni threshold for genome wide significance. $\lambda_{GC}$ is inside the parentheses next to the methods’ names.
3.5. Comparison between methods on the iCOGS dataset

Figure 3.6: QQ-plot of IC P-values

X-axis is the expected P-value quantile under uniform distribution. Y-axis is the empirical quantile of the observed P-values. The black line indicate the observed P-value under null hypothesis. The grey horizontal line is Bonferroni threshold for genome wide significance. $\lambda_{GC}$ is inside the parentheses next to the methods' names.
3.5. Comparison between methods on the iCOGS dataset

The three methods all agree on most of spikes of p-values passing their Bonferroni correction. Table 3.2 lists number of findings in terms of SNPs, haplotype blocks, regions and regions corrected for suspicious SNPs. GC PCA SNP find more haplotype blocks than the other two methods. However, not all haplotype blocks found by GC PCA SNP are of our interest. A SNP is suspicious if none of surrounding SNPs is significant by themselves (p-values ≤ 0.01). This suggests the signals of these SNPs are not captured by surrounding LD, and therefore more likely to be noise or artefacts. In Figure 3.7, we find that two IC methods have far less suspicious regions. Furthermore, if two haplotype blocks are both genome wide significant, and they are within 200k base pairs, we merge them and call it a “region”. After corrected for suspicious SNPs, the three methods find similar number of genome wide significant regions. Two IC methods find a few more than the SNP method.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SNP</th>
<th>Haplotype Block</th>
<th>Regions</th>
<th>Regions*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC PCA on SNPs</td>
<td>1681</td>
<td>116</td>
<td>61</td>
<td>48</td>
</tr>
<tr>
<td>GC PCA on ICs</td>
<td>-</td>
<td>96</td>
<td>57</td>
<td>51</td>
</tr>
<tr>
<td>RSS on ICs</td>
<td>-</td>
<td>89</td>
<td>56</td>
<td>49</td>
</tr>
</tbody>
</table>

Table 3.2: Associated loci by methods

Regions* are regions with suspicious SNPs removed. A SNP is suspicious if none of the nearby 50 SNPs have GC PCA p-value less than 0.01. 25 suspicious SNPs are found among all findings. If two genome wide significant SNPs are less than 200k base pairs apart, then they belong to the same region.
3.5. Comparison between methods on the iCOGS dataset

Figure 3.7: Manhattan plot by chromosomes

The subtitles correspond with the chromosomes. X-axis is the base pair position scaled by the length of the chromosomes. Y-axis is a scale different by chromosome as well. The dashed horizontal line is the fixed Bonferroni threshold across all chromosomes. Black points belong to the P-values of GC PCA SNP. Red points belong to the P-value of RSS IC. Blue points belong to the P-value of GC PCA IC.
3.5. Comparison between methods on the iCOGS dataset

Despite the number of significant regions are close, the actual regions found is quite different. Figure 3.8 shows that each methods contain some findings that are unique to themselves. We further compare these findings with the literature on breast cancer GWAS [85] (Figure 3.9). Among the three methods, GC PCA IC is able to recover most of the findings in the literature. In Figure 3.9(b), no more than half of SNPs are found to be significant pursuant to the results of all three methods. This would be for two reasons. Firstly, the original study is a meta analysis combining multiple ethnic groups, while our dataset only includes patients with European ancestors. Secondly, [85] only uses top 7 principal components to correct for population stratification (genomic control $\lambda = 1.20$). We conduct a much stringent population stratification using both PCA and GC (Genomic Control $\lambda = 1$).

Figure 3.8: Comparison of findings between methods
3.6 Numerical examples

We further construct two numerical simulation examples to validate the discoveries we found on the iCOGS dataset about the three methods.

Since ECC hypothesis is difficult to test on the GWAS dataset, we conduct two simulation scenarios to evaluate performance of ICA under the casual IC assumption and the casual SNP assumption. The ICA approach stands for using extracted ICs as covariates for association analysis. The standard approach uses SNPs directly as covariates for association analysis. The first simulation study assumes the ECC hypothesis, which means the disease casual variables are some latent variables of the SNPs. We simulate one haplotype block, and compare the best feature selected within the block by the ICA approach with the one by standard approach. We find that the ICA

Figure 3.9: Comparison of findings with literature findings

On breast cancer, there were 27 SNPs(a) were reported significant prior to the COGS project and additional 41 SNPs(b) were found significant based on a meta analysis including iCOGS dataset and addition 9 studies [85]. The number at the right bottom corner of the Venn graphs indicates number of SNPs fails to be found by all three methods.
approach has better prediction power under the ECC hypothesis. The second simulation study assumes that some SNPs are casual variables. We focus on the robustness of the ICA approach when its underlying model is false. We find that ECC is a relatively conservative method.

At last, given the results of two simulation studies, we further speculate that the 21 additional locus we found by RSS IC are supporting evidence of the existence of ECC in the human genome.

3.6.1 Simulation on haplotype block under the ECC hypothesis

We assume that within a haplotype block, there are $q$ number of independent genetic switches, and $p$ number of SNPs. Let the genetic switch vector $\mathbf{u} = (u_1, \ldots, u_q)^\top$ following Bernoulli distribution with frequency $\mathbf{\phi} = (\phi_1, \ldots, \phi_q)^\top$. Then we generate a random channel encoder matrix $\mathbf{G}$ with entries following uniform distribution $\text{Unif}(14, 16)$. The allele frequencies of SNPs is proportionate to the $\mathbf{G}\mathbf{u} + \mathbf{\varepsilon}$, where elements of $\mathbf{\varepsilon}$ are i.i.d random vector with $\text{Norm}(0, 4)$. Then we simulate phenotype $y$ with the logistic model using only the first genetic switch as the causal covariate. To evaluate the performance, we compare the ICA approach with the standard approach. The ICA approach selects the lead IC in the block, while the standard approach selects the lead SNP in the block. We compare the P-values of the lead features on the training dataset, and Area Under ROC curve (AUC) on an independent testing dataset. We conduct experiment by varying $p$, $q$ and $\phi_i$.

**Results**  This simulation shows that under the ECC hypothesis, the ICA approach is better in recovering the latent casual genetic switch than the lead SNP approach. Table 3.3 shows the results where the causal genetic switch is
a common variate (frequency between 0.1 and 0.5). Table 3.4 and 3.4 show the results when the causal genetic switch is a rare variate (frequency = 0.04). “Oracle” denotes the causal genetic switch. The oracle P-value and AUC are not dependent on \( q \) and \( p \). In table 3.4, the large oracle P-value is caused by the large standard error of the rare variate. It indicates that the larger the \( p \), the denser the genotype data is. Because of the multiple testing correction, the lead SNP approach loses its significance dramatically when \( p \) increases. However, the number of independent component estimated by ICA is stable. Therefore, ECC suffers much less significance loss due to dense genotype data. In terms of prediction power AUC, the ICA approach benefits much more from increasing \( p \) than the lead SNP approach. In both cases, overall the ICA approach has higher AUCs.

### 3.6.2 Simulation under the casual SNPs hypothesis

We select SNPs in chromosome 6 from the iCOGS dataset as our simulation covariates. The HLA region in chromosome 6 has a strong link to the immune system \(^8\) which makes it an interesting region to test with. Then a random subsample of 1000 individuals are selected for the simulation. Among 14027 SNPs on chromosome 6, we estimate 1780 haplotype blocks, and 2572 independent components are extracted using ICA. We randomly choose \( q \) haplotype blocks as causal blocks, and select one SNP from each block as casual SNPs. Then we generate phenotype \( y \) using those \( q \) SNPs with the same effect size. We compare the results by using three methods are compared. GC SNP, GC IC and RSS IC. GC SNP and GC IC perform multiple tests on logistic

<table>
<thead>
<tr>
<th>$q$</th>
<th>$p$</th>
<th>$r^2$</th>
<th>Oracle P</th>
<th>Lead SNP P</th>
<th>Lead IC P</th>
<th>Oracle AUC</th>
<th>Lead SNP AUC</th>
<th>Lead IC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0.60</td>
<td>2.0e-08</td>
<td>6.78e-06</td>
<td>1.49e-07</td>
<td>0.705</td>
<td>0.698</td>
<td>0.706</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>0.61</td>
<td>2.3e-08</td>
<td>5.39e-06</td>
<td>2.63e-07</td>
<td>0.713</td>
<td>0.709</td>
<td>0.711</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>0.67</td>
<td>4.3e-09</td>
<td>9.18e-07</td>
<td>1.42e-08</td>
<td>0.718</td>
<td>0.712</td>
<td>0.718</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>0.64</td>
<td>8.8e-09</td>
<td>3.60e-07</td>
<td>3.98e-09</td>
<td>0.713</td>
<td>0.710</td>
<td>0.712</td>
</tr>
<tr>
<td>1</td>
<td>200</td>
<td>0.64</td>
<td>2.5e-08</td>
<td>1.03e-07</td>
<td>2.19e-08</td>
<td>0.711</td>
<td>0.708</td>
<td>0.715</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>0.72</td>
<td>6.7e-09</td>
<td>5.32e-04</td>
<td>2.59e-04</td>
<td>0.715</td>
<td>0.622</td>
<td>0.634</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>0.69</td>
<td>9.7e-03</td>
<td>2.21e-02</td>
<td>4.24e-03</td>
<td>0.712</td>
<td>0.623</td>
<td>0.633</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>0.67</td>
<td>3.1e-08</td>
<td>2.18e-02</td>
<td>2.05e-03</td>
<td>0.706</td>
<td>0.616</td>
<td>0.633</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>0.71</td>
<td>4.1e-07</td>
<td>5.45e-02</td>
<td>3.77e-03</td>
<td>0.711</td>
<td>0.620</td>
<td>0.631</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>0.71</td>
<td>1.4e-08</td>
<td>4.11e-02</td>
<td>1.73e-03</td>
<td>0.717</td>
<td>0.629</td>
<td>0.640</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.69</td>
<td>2.6e-09</td>
<td>5.71e-02</td>
<td>1.94e-02</td>
<td>0.725</td>
<td>0.596</td>
<td>0.611</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.74</td>
<td>1.9e-07</td>
<td>1.95e-01</td>
<td>3.12e-02</td>
<td>0.708</td>
<td>0.590</td>
<td>0.596</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>0.73</td>
<td>7.6e-08</td>
<td>1.61e-01</td>
<td>1.95e-02</td>
<td>0.700</td>
<td>0.588</td>
<td>0.601</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>0.72</td>
<td>1.8e-08</td>
<td>2.49e-01</td>
<td>1.29e-02</td>
<td>0.711</td>
<td>0.589</td>
<td>0.599</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>0.75</td>
<td>5.2e-08</td>
<td>1.43e-01</td>
<td>4.89e-03</td>
<td>0.711</td>
<td>0.594</td>
<td>0.605</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.79</td>
<td>3.8e-09</td>
<td>3.37e-01</td>
<td>1.15e-01</td>
<td>0.701</td>
<td>0.559</td>
<td>0.564</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>0.76</td>
<td>2.3e-06</td>
<td>5.93e-01</td>
<td>9.96e-02</td>
<td>0.706</td>
<td>0.559</td>
<td>0.566</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>0.76</td>
<td>5.2e-08</td>
<td>1.00e+00</td>
<td>9.75e-02</td>
<td>0.703</td>
<td>0.561</td>
<td>0.562</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>0.78</td>
<td>3.5e-08</td>
<td>1.00e+00</td>
<td>8.22e-02</td>
<td>0.707</td>
<td>0.568</td>
<td>0.572</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>0.76</td>
<td>2.2e-09</td>
<td>1.00e+00</td>
<td>5.80e-02</td>
<td>0.718</td>
<td>0.563</td>
<td>0.573</td>
</tr>
</tbody>
</table>

Table 3.3: Experiment on haplotype block with ECC hypothesis (common variant)

Each experiment is conducted on one simulated haplotype block with $q$ genetic switches and $p$ SNPs. Each experiment has sample size 500 and is repeated for 100 times. Genetic switches is with frequency sampled from $\text{Unif}(0.1, 0.5)$. Only one switch is causal with odds ratio $= 4.5$. Average number of independent components extracted in each block is 2. $r^2$ is the average LD $r^2$ between SNPs. Oracle P denotes the P-value of causal genetic binary switch. All P-values are adjusted for multiple testing. AUC is calculated on independent test dataset with the same sample size. Oracle AUC is calculated using the causal genetic binary switch.
### Table 3.4: Experiment on haplotype block under ECC hypothesis (rare variant)

Each experiment is conducted on one simulated haplotype block with $q$ genetic switches and $p$ SNPs. Each experiment has sample size 1000 and is repeated for 100 times. The one and only one causal genetic switch has frequency $0.04$ with odds ratio $= 5 \times 10^{21}$, other genetic switch is sampled from frequency following $\text{Unif}(0.1, 0.5)$. Average number of independent components extracted in each block is 2. $r^2$ is the average LD $r^2$ between SNPs. Oracle P denotes the P-value of causal genetic binary switch. All P-values are adjusted for multiple testing. AUC is calculated on independent test dataset with the same sample size. Oracle AUC is calculated using the causal genetic binary switch.
regression coefficient with genomic control on SNPs and ICs respectively. RSS IC uses rank stability selection to select associated ICs. We evaluate the performance through the average number of causal haplotype blocks found, and the average number of non-causal haplotype blocks selected.

<table>
<thead>
<tr>
<th>s</th>
<th>GC SNP</th>
<th>RSS IC</th>
<th>GC IC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TP</td>
</tr>
<tr>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.98</td>
<td>2.67</td>
<td>0.77</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>7.79</td>
<td>0.86</td>
</tr>
<tr>
<td>5</td>
<td>4.54</td>
<td>7.11</td>
<td>3.46</td>
</tr>
<tr>
<td>5</td>
<td>4.91</td>
<td>14.87</td>
<td>3.98</td>
</tr>
<tr>
<td>10</td>
<td>7.22</td>
<td>9.25</td>
<td>5.41</td>
</tr>
<tr>
<td>10</td>
<td>8.81</td>
<td>16.99</td>
<td>6.83</td>
</tr>
</tbody>
</table>

**Table 3.5**: Experiment on chromosome 6 under casual SNP hypothesis

$s$ is the number of casual SNPs. $\beta$ is the log odds of the risk alleles. TP stands for average number of true causal haplotype blocks found. FP stands for the average the number of false causal haplotype blocks found. ADD is the additional number causal blocks found by two IC methods combined to the lead SNP methods. Experiment is repeated for 100 times.

**Results** In this simulation study, casual variables are SNPs instead of some latent variables. We compare the performance of the ICA approaches against the standard approach where ECC hypothesis is no longer valid. Due to the strong LD structure, measuring whether pre-defined casual SNP is selected by a method is not quite meaningful. Quite a number of SNPs in strong LD to the casual SNP will become significant as well. This would lead to high error rate for the standard approach. To assess the performance, we examine whether the haplotype block of the pre-defined casual SNP is chosen. In general, the ICA approaches (RSS IC and GC IC) are more conservative than the standard approach. The difference between RSS IC and GC IC is insignificant, and RSS IC is the most conservative method. From Table 3.5, we find that despite that the ICA approaches have weaker power, they contain much fewer number of
3.7. Discussions

errors comparing to the standard approach. Here we suspect the multiple testing power gain is less than the effect of information loss due to dimension reduction under the false model.

Furthermore, the ICA approaches are much more precise in finding causal haplotype blocks, where GC SNP selects far more non-associated blocks. This phenomenon is not likely to be caused by strong LD threshold for haplotype blocks. For the iCOGS dataset we have already chosen a low LD threshold\(^b\) to estimate haplotype blocks. In this setting, LD between haplotype blocks is weak.

3.7 Discussions

The error correction code hypothesis provides an intuitive model for dimension reduction in GWAS. It is a robust and extendable framework with several merits.

Evidence of ECC in the iCOGS dataset  By comparing the results of the iCOGS dataset and simulation studies, we find strong evidence supporting the existence of ECC in some haplotype blocks. Through two simulation studies, we have established that the ICA approaches have greater power than the standard approach where the ECC hypothesis is true, and more conservative when ECC hypothesis is false. However on the iCOGS dataset, two ICA approaches (GC PCA IC, RSS IC) select a few more associated loci than the standard approach. Furthermore, under the casual SNP assumption, GC SNP is able to find most of casual SNPs found by the ICA approaches but not vice versa. But on the iCOGS dataset, the sets of associated loci selected

\(^b\)PLINK settings for the proportion of “strong LD” pairs within a haplotype block is 0.70, the default value is 0.95[91].
3.7. Discussions

by different methods differ significantly (Figure 3.8). This suggests that at least some of casual variables are not likely to be SNPs.

**Stability Rank Selection**  RSS is a non-parametric multiple testing procedure introduced in Section 2.3.2. It is robust to population structures confounding factors and requires no additional estimation or calculation. On the iCOGS dataset, it has similar number of significant haplotype blocks as GC PCA does (Figure 3.6). This verifies its claim. Another characteristic of RSS is that stability signals are different from significant signals. RSS is able to find signals in addition to the full sample association analysis. This is reflected in Figures 3.8 and 3.9. 10 regions found by RSS IC is absent from the other two methods, and 4 of them are found to be significant in the literature. On the iCOGS dataset, RSS proves itself as a reliable and simple procedure for multiple testing.

**Dense genotype data**  Recently, rare variant has attracted a lot of interest[44]. One by the product of rare variant analysis would significantly increase the number of SNPs in a GWAS. This can be achieved by using ultra high density arrays \(^1\) or imputation methods. With imputation on 1000 Genome project[17], the number of imputed SNPs will easily exceed 1 million. This imposes much higher pressure on multiple testing, and the genome wide significance will be raised from \(5 \times 10^{-7}\) to \(5 \times 10^{-8}\). Meanwhile, the LD structure is invariant to the density of genotype data. The total number of haplotype blocks with respect to a population would still be the same. The ICA approaches would easily avoid the hustle brought by ultra dense genotype data. Since iCOGS only have 210935 SNPs, we expect that the effect

---

of dimension reduction would increase 5-fold for imputed data.

**Epistasis analysis** The ICA approach reduces the dimension of the problem for epistasis analysis geometric increases as the interaction order increases. Even though we only reduce the dimension of the problem by 4 times on the iCOGS dataset, in a two-way interaction analysis, the ICA approach will reduce the dimension of the problem by 25 times. In epistasis analysis, in order to test different types of interaction SNPs are discretized. In a logistic regression model, this would lead to 4 degrees of freedom for the interaction terms. In addition, using all 4 degrees of freedom to model interaction would end up in a saturated model. Another side effect of discretized SNPs is that two-way interaction contingency tables would create a lot of dummy variables with low counts. This is a typical rare variant problem but with much larger quantity. On the other hand, if genotype density was high, there would be enough variation in ICs, and ICs can be treated as continuous variables. We can simply use the polynomials to model the interaction effects.

Another possible approach would be extending the SNP sets from haplotype blocks to genes or even pathways. As mentioned in [71], kernel methods can be used to detect interaction effects within the SNP sets. We can implement kernel principal component analysis[103] at the whitening stage and then, the ICs would both include linear and nonlinear genetic effects.

**PCA and permutation test** In Section 3.5.1, we have shown that using genotype principal component scores as covariates in the logistic regression can adjust the population stratification. With adequate number of principal components, we can apply permutation test on the coefficient of the SNPs. According to simulation studies in Section 2.5.1, permutation tests have the greatest test power among four methods. In that sense, PCA and permutation
test would be a promising combination. However, our RSS and dimension reduction method still have several advantages over the PCA and permutation test approach. Firstly, determination of adequate number principal components is a non-trivial task. On the other hand, dimension reduction and RSS are relatively straightforward in application. Secondly, RSS is much less computational intensive than permutation tests. The error rate of RSS does not depend on the number of resampling. In our simulation, we find that 50 times of random sample splitting is good enough. On the contrast, permutation tests generally require 1000 times of permutation. Thirdly, in theory, PCA only accounts for linear population structure. RSS by default, assumes exchangeability which is a more general condition.

Future research In Section 3.4.3, we have listed 26 significant SNPs that is new to the breast cancer GWAS literature. In future research, we should aim to validate these findings on independent GWAS breast cancer datasets.

3.8 Appendix
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>SNP</th>
<th>Allele 1</th>
<th>Allele 2</th>
<th>Allele counts</th>
<th>Observed heterozygosity rate</th>
<th>Expected heterozygosity rate</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2311859</td>
<td>A</td>
<td>G</td>
<td>982/8480/28558</td>
<td>0.223</td>
<td>0.237</td>
<td>2.623e-28</td>
</tr>
<tr>
<td>2</td>
<td>rs6730157</td>
<td>G</td>
<td>A</td>
<td>5557/17962/19418</td>
<td>0.4183</td>
<td>0.4479</td>
<td>3.147e-42</td>
</tr>
<tr>
<td>2</td>
<td>rs3213943</td>
<td>A</td>
<td>C</td>
<td>1329/11273/30329</td>
<td>0.2626</td>
<td>0.2718</td>
<td>4.197e-12</td>
</tr>
<tr>
<td>2</td>
<td>rs6430298</td>
<td>A</td>
<td>C</td>
<td>7370/20532/13453</td>
<td>0.4965</td>
<td>0.4892</td>
<td>0.00248</td>
</tr>
<tr>
<td>4</td>
<td>rs4698932</td>
<td>G</td>
<td>A</td>
<td>9224/20394/10042</td>
<td>0.5142</td>
<td>0.4998</td>
<td>9.076e-09</td>
</tr>
<tr>
<td>4</td>
<td>rs6830464</td>
<td>A</td>
<td>G</td>
<td>6539/17438/13763</td>
<td>0.4621</td>
<td>0.4817</td>
<td>2.59e-15</td>
</tr>
<tr>
<td>5</td>
<td>rs7443354</td>
<td>A</td>
<td>G</td>
<td>183/5242/37514</td>
<td>0.1212</td>
<td>0.1221</td>
<td>0.00248</td>
</tr>
<tr>
<td>5</td>
<td>rs380346</td>
<td>C</td>
<td>A</td>
<td>5676/19634/15147</td>
<td>0.4853</td>
<td>0.4726</td>
<td>6.388e-08</td>
</tr>
<tr>
<td>6</td>
<td>rs1543403</td>
<td>G</td>
<td>C</td>
<td>10026/21088/11820</td>
<td>0.4912</td>
<td>0.4991</td>
<td>0.0009763</td>
</tr>
<tr>
<td>7</td>
<td>rs4593472</td>
<td>A</td>
<td>G</td>
<td>5532/19716/17688</td>
<td>0.4592</td>
<td>0.4599</td>
<td>0.745</td>
</tr>
<tr>
<td>7</td>
<td>rs10252584</td>
<td>G</td>
<td>A</td>
<td>5929/16403/13712</td>
<td>0.4551</td>
<td>0.4767</td>
<td>8.874e-18</td>
</tr>
<tr>
<td>7</td>
<td>rs720475</td>
<td>A</td>
<td>G</td>
<td>2870/16079/23991</td>
<td>0.3745</td>
<td>0.379</td>
<td>0.01258</td>
</tr>
<tr>
<td>8</td>
<td>rs799890</td>
<td>C</td>
<td>G</td>
<td>1529/13057/28350</td>
<td>0.3041</td>
<td>0.3049</td>
<td>0.5905</td>
</tr>
<tr>
<td>8</td>
<td>rs67397162</td>
<td>A</td>
<td>G</td>
<td>1151/11426/30360</td>
<td>0.2661</td>
<td>0.2686</td>
<td>0.05454</td>
</tr>
<tr>
<td>11</td>
<td>rs10896050</td>
<td>A</td>
<td>C</td>
<td>1430/12743/28755</td>
<td>0.2968</td>
<td>0.2974</td>
<td>0.6969</td>
</tr>
<tr>
<td>11</td>
<td>rs4545589</td>
<td>G</td>
<td>A</td>
<td>3914/17881/17557</td>
<td>0.4544</td>
<td>0.4399</td>
<td>5.973e-11</td>
</tr>
<tr>
<td>12</td>
<td>rs2887022</td>
<td>G</td>
<td>A</td>
<td>4114/18285/20537</td>
<td>0.4259</td>
<td>0.4268</td>
<td>0.6349</td>
</tr>
<tr>
<td>13</td>
<td>rs2295882</td>
<td>G</td>
<td>A</td>
<td>2629/15827/24473</td>
<td>0.3687</td>
<td>0.3705</td>
<td>0.2974</td>
</tr>
<tr>
<td>16</td>
<td>rs8064086</td>
<td>G</td>
<td>C</td>
<td>232/5637/17715</td>
<td>0.239</td>
<td>0.2252</td>
<td>2.779e-23</td>
</tr>
<tr>
<td>16</td>
<td>rs735753</td>
<td>C</td>
<td>G</td>
<td>2092/14331/26516</td>
<td>0.3338</td>
<td>0.3382</td>
<td>0.006422</td>
</tr>
<tr>
<td>17</td>
<td>rs11651154</td>
<td>G</td>
<td>A</td>
<td>5634/17771/14034</td>
<td>0.4747</td>
<td>0.4748</td>
<td>0.9479</td>
</tr>
<tr>
<td>17</td>
<td>rs7503278</td>
<td>G</td>
<td>A</td>
<td>1383/12303/24352</td>
<td>0.3234</td>
<td>0.3177</td>
<td>0.0004074</td>
</tr>
<tr>
<td>18</td>
<td>rs635269</td>
<td>G</td>
<td>A</td>
<td>6894/20785/13612</td>
<td>0.5034</td>
<td>0.4868</td>
<td>4.031e-12</td>
</tr>
<tr>
<td>18</td>
<td>rs1175745</td>
<td>C</td>
<td>A</td>
<td>6808/20674/15454</td>
<td>0.4815</td>
<td>0.4797</td>
<td>0.4445</td>
</tr>
</tbody>
</table>

**Table 3.6: Hardy-Weinberg Equilibrium test result**

Although some P-values are very much, the difference between Observed and expected heterozygosity rates are very close. This is due to the large sample size of the iCOGS dataset. Any small difference would result a small P-value.
Predicting disease risk given the genotype is one of the essential tasks of GWAS. For complex diseases, multi loci on genome could be associated with the disease, and it would require a multivariate statistical model to accurately predict the risk. A model selection criterion can be used to select the best model based on observations. However, given up to millions of variants in GWAS, classical model selection criterion such as AIC and BIC would fail to select the correct model but favour overfitting ones. We propose a novel model selection criterion, correction of Stochastic Complexity Criterion (SCCc). SCCc is derived from the Minimum Description Length (MDL) principles. Under ultra-high dimensional settings, SCCc imposes a sparse notation for variables included in the model and it maintains its model selection consistency by adding a penalty term that corresponds to this sparse notation. We will show by proof that it is variable selection consistent, and also by simulation that it has a stable FWER in finite samples.

4.1 Introduction

Previous chapters have demonstrated new methods for multiple testing in GWAS.

In multiple testing, only one genetic variant is tested for association at a
time.

For more complex diseases such as cancers, this simple approach might not be sufficient to explore the underlying genetic associations.

Since most of Mendelian diseases have been examined already [118], the new trend in GWAS is to dig deep in with more complex models, such as multi-variate analysis, additive model with interaction terms [74, 117, 128]. This imposes new challenges in GWAS. With $m$ genetic variants, the simple additive effective combination of variables is of $O(2^m)$. Therefore, the multiple testing approach on complex statistical models is simply not feasible. The Bonferroni threshold would be too high to leave any test power. On the other hand, an information criterion measures the goodness of a model given observations, and the best model has the lowest value. In more details, they are estimations of Kullback Leibler divergence between the working model and the underlying true model. Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) are most widely used due to their simplicity and ability to penalize on the complexity of the models [38]. Although information criterion does not have inference abilities such as Type I error rate, they are usually model selection consistent which says they will select the correct statistical if it exists in the pre-defined model space with probability one when the sample size is large enough.

Denote $n$ as the sample size and $m$ as the total number of covariates. With fixed $m$, AIC and BIC are model selection consistent which means selecting the true model when the sample size goes to infinite [94]. However, in $m \gg n$ problems such as GWAS, they would become too liberal and tend to over select noise variables. Modified BIC [110, 132], and Extended BIC (EBIC) [14, 15] tackle this problem by introducing an additional penalization term on the size of the potential model space. For example, for $m$ covariates, there
are at least $2^m$ models available for selection. Each model corresponds to a unique combination of covariates included in the model. Chen and Chen[15] have proved the consistency of EBIC for generalized linear models when $m = O(\exp(n^\kappa))$ as $n \to \infty$ for $\kappa \in (0,1/3)$ under certain conditions. However, there is a parameter $\lambda$ inside EBIC which controls the strength of the additional penalization term. Chen and Chen suggests $\lambda = 0.5$ based on their simulation studies. We will show that from our numerical examples, a fixed value $\lambda$ is not a good choice if we want to control the variable selection errors consistently. The value of $\lambda$ forms yet another optimization problem.

In this paper, we propose a novel model selection criterion based on Stochastic Complexity Criterion (SCC)[92, 93, 96]. It is a correction of SCC(SCCc) in high dimensional data problems. In parallel to AIC and BIC, under the Minimal Description Length (MDL) principles, SCC evaluates model goodness by their ability to compress the observed data with aid of a family of statistical models[99–101]. In other words, SCC has the shortest code length of the compressed data. SCC is usually more liberal than BIC with a finite sample size[51]. To revise SCC under the high dimensional problem, we add a part of code to SCC by extending the MDL principles to the model structure space. Let model structure be the set of covariates with non-zero coefficients in the logistic regression models. The model structure space is collection of all model structures. In $m \gg n$ problems, the complexity of model structures becomes enormous. Therefore, we could save significant code length by using a sparse notation of covariates with non-zero coefficients if the model structure is relatively small. We first derive an optimal code for the additional part under MDL principals. However, to calculate its code length, the knowledge of the underlying true model is necessary.

We then propose a sub-optimal code derived from the null model hy-
4.1. Introduction

pthesis, which assumes that no covariates are associated with the response variable. On that basis, we are able to control the Type I error $\alpha$ (the chances of selecting false models under the null model hypothesis). This features a novel attempt to bring statistical inference into model selection criteria.

SCCc can be calculated by either permutation techniques or asymptotic approximations. The asymptotic version is

$$SCCc : - \log \text{likelihood} + \frac{k}{2} Q(\chi^2_{df=1}, \alpha/m)$$  \hspace{1cm} (4.1)

where $k$ is the number of covariates in the model with non-zero coefficients, $Q$ is the right quantile function and $\alpha$ is the target Type I error rate. We will also show that under reasonable conditions, SCCc has the same consistent properties as EBIC, i.e. the probability of selecting the true model structure equals to one when $n \to \infty$. In comparison, with EBIC, SCCc’s $\alpha$ is a more intuitive parameter than $\lambda$, and through numerical examples, we will show that it is consistent in error control regarding varying $m$.

Introducing error control enables SCCc to select models within a local region on the genome (i.e. models built on SNPs within certain pathway) but still reach genome wide significance. To achieve this, we can simply substitute the number of sample genetic variants $m$ with the number of genome wide variants.

In addition, there are discussions in GWAS between the rare genetic variants (i.e. SNPS with low minor allele frequencies(MAF)) versus common genetic variants [6, 23, 44, 104]. For rare variants, SNP set based methods are preferred to single site analysis[69]. In additional to current methods, multivariate logistic regressions can be used to find significant regions with rare variants. However, we find standard information criteria tend to include more false positives for rare variants and permutation based SCCc would still
4.2 Variable selection and model selection criteria

Since this thesis focuses on case control study designs, we restrict the variable selection and model selection problems under the logistic models. Let $Y$ be the response variable and $x = (X_1, \cdots, X_m)$ be the vector of covariates. We assume the associations between them follows a logistic regression model, i.e.

$$\logit(E(Y|x)) = \beta_0 + \beta^T x,$$

where $\beta = (\beta_1, \cdots, \beta_m)^T$ and $\beta_0$ are real valued coefficients, i.e. the parameters of the logistic regression model. Details can be found in Section 1.2.1.

To study the associations between $Y$ and $x$, we are only interested in the non-zero elements in $\beta$ and $\beta_0$ is not of our concern. Let $\Theta$ denote the parameter space of $\beta$ in $\mathbb{R}^m$. We assume $\beta$ is identifiable in $\Theta$. Let $M$ be the union of all subsets of indices of the covariates. A model structure $S \in M$ is a set of indices. It also denotes a family of logistic regression models with the corresponding non-zero coefficients. The variable selection problem is to find the $S_t$ which specifies all the non-zero components of $\beta$ given a dataset.

Given data and a $S$, we can fit $\hat{\beta}$ using maximum likelihood estimator (MLE). Let vector $y^n$ and design matrix $X^n$ denote the $n$ i.i.d. random variables of $Y$ and corresponding covariate vectors $x$. If we are restricted to MLE and $n$ is large enough then given $S$ we can fit a unique logistic regression model with $\hat{\beta}_0$ and $\hat{\beta}_S$, where $\hat{\beta}_i = 0$ for any $i \notin S$. Now we can estimate $S_t$ with the logistic regression model corresponds to $\hat{S}$ that describes the dataset the best. To evaluate the goodness of a logistic model given structure $S$ and the
dataset, AIC and BIC are commonly accepted model selection criteria. They are derived from the KL-divergence between the current model $S, \hat{\beta}_S$ and the ideal true model $S_t, \beta$. They consist of negative log-likelihood, and a penalty term proportion to the degree of complexity of $\hat{\beta}_S$.

$$AIC : -2\ell + 2k, \quad BIC : -2\ell + k \log(n)$$

(4.3)

Where $\ell$ is the log likelihood of $y^n$ given $\hat{\beta}_S$, $k$ is the number of covariates with non-zero coefficients in the model. Given a model selection criterion, we regard that the model have the minimum score as the best model. From (4.3), we can see BIC penalizes heavier on the model complexity than AIC with moderate size of $n$. However, in $m \gg n$ problems, even BIC is too liberal, i.e. in favour of over-fitted models[14, 15]. To tackle this problem, Modified BIC (MBIC)[132] and Extended BIC (EBIC)[14, 15] further extend the penalty term by adding $O(k \log p)$.

$$EBIC : -2\ell + k \log(n) + 2k\lambda \log m, \quad \text{with } \lambda \in [0, 1]$$

(4.4)

In this chapter, we develop a novel model selection criterion based on Minimal Description Length(MDL) principles to avoid the selection of any tuning parameter.

### 4.2.1 Stochastic complexity and two part codes

In MDL principles, Stochastic Complexity(SC) has the shortest binary code length of the observations obtainable with the help of some model class[98, 100, 101]. The model class used in SC is a family of parametric models specified by a real valued parameter vector. Recall in logistic regression models, given $x$ the distribution of $Y$ is specified by $\beta_0$ and $\beta$.

Given a dataset of $y^n$ and $X^n$, SC can be calculated with two parts of code. The first part is the code of data given the probability model(specified by $\beta$).
4.2. Variable selection and model selection criteria

The second part is the code of parameters $\beta$. However, $\beta$ is unknown but we can use the MLE $\hat{\beta}$ instead. The two-part-code length can be denoted as $L_{SC}$, such that

$$L_{SC}(y^n) = L_1(y^n | \hat{\beta}) + L_2(\hat{\beta})$$

SC is the shortest (4.5) obtainable. Qian(1998)[92] gives a computable derivation in (4.6). We call it Stochastic Complexity Criterion (SCC), and will use it for the shortest two part code throughout the paper.

$$SCC : -\ell + \frac{1}{2} \log |I_n(\hat{\beta})| + \sum_{j=1}^{k} \log \left( |\hat{\beta}_j| + \varepsilon n^{-1/4} \right)$$

where $\hat{\beta}$ is the maximum likelihood estimator of $\beta$ and $I_n$ is the corresponding sample fisher information matrix. Both BIC and SCC includes sample size $n$ in the penalization term. Qian proofs under some condition that BIC and SCC are strongly consistent while AIC is not[94]. In practice, AIC is the most liberal, i.e. tend to select more covariates, while BIC is stricter than SCC. Therefore, SCC suffers the same problem as BIC which is too liberal when $m$ is relatively large comparing to $n$.

4.2.2 Stochastic complexity corrected for sparsity

In the $m \gg n$ problems, we adopt the sparsity assumption, i.e. most elements of $\beta$ is zero. Otherwise, it would be impossible to estimate $\beta$ since the degree of freedom of the model would be larger than the sample size. Under the sparsity assumption, substantial code length, $L_2(\beta)$ would be saved if we only code the non-zero coefficients in $\beta$ like the sparse matrix notation in computational soft-wares. In this case, we can introduce an additional code to denote the index of those non-zero coefficients, equivalently, the model structure $S$. We use $L_3(S)$ to denote the corresponding code length, and we
4.2. Variable selection and model selection criteria

have the Sparse Stochastic Complexity as

\[ L_{\text{sparse}}(y^n, S) = L_1(y^n|\hat{\beta}_s) + L_2(\hat{\beta}_s) + L_3(S) \]  

(4.7)

Given model structure \( S \), SCC minimizes \( L_1 \) and \( L_2 \). For \( L_{\text{sparse}} \) to be applicable, we need to find a way to code \( S \) for all \( S \in M \). We introduce the following two approaches under the MDL principles.

The Bayesian approach  Here we give a heuristic way to determine \( L_3(S) \) for \( S \in M \) using Bayesian approach. In Bayesian approach, we would like to find a model structure that maximizes the posterior probability. Then we propose a prior distribution on \( M \) so that the we always select the correct model structure when \( n \) is large enough.

First of all, in this discussion, \( y^n \) and \( S \) are all conditioned on \( X^n \). To begin with, we would like to maximize the posterior probability of a model structure given data, i.e.

\[ \Pr(S|y^n) = \frac{\Pr(y^n|S) \cdot \Pr(S)}{\Pr(y^n)}. \]  

(4.8)

\[ \max_S \Pr(S|y^n) \Leftrightarrow \min_S \{- \log(\Pr(y^n|S)) - \log(\Pr(S))\} \]

Let \( \mathcal{P} \) be a universal probability space induced from \( y^n \) conditioning on \( X^n \). We set \( \pi(S) \) to be a prior distribution that equals to the chances of a model structure \( S \) maximizes the \( \Pr(y^n|S) \) i.e.

\[ \pi(S) = \Pr_S \left( S = \arg\min_{T \in M} \Pr(y^n|T) \right). \]  

(4.9)

Then we know that \( S_t \) and MLE \( \hat{\beta}_{S_t}, \hat{\beta}_{S_t,0} \) would maximize \( \Pr(y^n|S) \) and so does \( S_t \) would maximize \( \pi(S) \). Therefore, with prior \( \pi(\cdot) \), \( S_t \) would maximize the posterior probability.
4.2. Variable selection and model selection criteria

For $L_{\text{sparse}}$, we can set $L_3$ be corresponds to a similar prior $\pi^*(\cdot)$. $\pi^*(\cdot)$ is the chance of a model structure would minimize SCC, i.e.

$$L_3 = -\log(\pi^*(S))\text{ and }$$

$$\pi^*(S) = \Pr\left(S = \arg \min_{T \in M} \text{SCC}(y^n, T)\right). \quad (4.10)$$

This would automatically make $L_3$ optimal prefix code with redundancy free, i.e.

$$\sum_{S \in M} e^{-L_3(S)} = 1. \quad (4.11)$$

And similarly, we will have $S_t$ minimizes $\text{SCC}(y^n, T)$ and $L_3(T)$ for $T \in M$. $\pi^*(S)$ is a promising solution for the third part code length. When $m$ gets larger, the model structure space grows at rate of $O(2^m)$ and the prior probability of incorrect models $1 - \pi(S_t)$ would also increase at the same speed. $L_3$ would serve as a penalty component for ultra high dimensions.

The minimax regret approach Another way to construct the third part code length is the minimax regret approach. The regret of $S$ is the code length could be saved if using the best model structure $S_t \in M$ regarding to $y^n$ given $X^n$. Note that the best model is not necessary the true model, and the regret is a random variable as well. Minimax regret is a model selection strategy that selects the optimal model which minimizes the maximum regret. It is widely used in decision theories and MDL. It chooses the model which minimizes the opportunity cost in the worst scenario.

We want to construct a $\hat{L}_3(\cdot)$ so that the true model would always minimize the maximum regret. Since $L_3$ is subject to our specification, we find some $L_3$ that would set maximum regret for $S_t$ to be zero. Since maximum regret is non-negative, this ensures that $S_t$ minimizes the maximum regret for all $S \in M$. Under this configuration, $S_t$ would minimize $L_{\text{sparse}}$ regardless of size of $m$. 

112
4.2. Variable selection and model selection criteria

Conditioning on $X^n$, define $\text{Reg}(S, y^n)$ to be the regret of model $S$.

$$\text{Reg}(y^n, S) = L_{\text{sparse}}(y^n, S) - \min_{T \in M} L_{\text{sparse}}(y^n, T)$$  \hspace{1cm} (4.12)

If we use SCC as the first two part code, then we have

$$\max \text{Reg}(S) = \max_{y^n \in Y^n} \left\{ L_{\text{sparse}}(y^n, S) - \min_{T \in M} L_{\text{sparse}}(y^n, T) \right\}$$  \hspace{1cm} (4.13)

The final step is to construct $L_3$ so that $S_t$ minimizes the maximum regret. Suppose $L_3(S_t) = c_0$, then we have

$$L_3(S) = \max_{y^n \in Y^n} \left\{ \text{SCC}(y^n, S_t) - \text{SCC}(y^n, S) \right\} + c_0$$  \hspace{1cm} (4.14)

$c_0 > 0$ here is a normalized constant so that $L_3(\cdot)$ would be a prefix code, i.e. (4.11) would be true. Then, we have

$$\max \text{Reg}(S_t) = \max_{y^n \in Y^n} \left\{ L_{\text{sparse}}(y^n, S_t) - \min_{T \in M} L_{\text{sparse}}(y^n, T) \right\}$$

$$\leq \min_{T \in M} \max_{y^n \in Y^n} \left\{ L_{\text{sparse}}(y^n, S_t) - L_{\text{sparse}}(y^n, T) \right\}$$

$$\leq \min_{T \in M} \left\{ \max_{y^n \in Y^n} \left\{ \text{SCC}(y^n, S_t) - \text{SCC}(y^n, T) \right\} + L_3(S_t) - L_3(T) \right\}$$

$$\leq 0.$$  

Since regret is always non-negative, so is $\max \text{Reg}$. Therefore with (4.14), $S_t$ would minimize $\max \text{Reg}$ as desired.

The above two approaches are only MDL principles but not statistical methods. Both of them require the knowledge of $S_t$ which is what we are after at the first place. To work around this, we can either introduce some subjective prior information on $\pi(S)$ through the Bayesian approach. Or derive $L_3$ based on fair assumptions(or non-informative prior in Bayesian). One of the simplest approaches is a uniform code length for $S \in M$[51], so
4.2. Variable selection and model selection criteria

that all models have equal chances to be generate the shortest code length. “Luckiness principle”[47] can reduce substantial amount of code length by taking advantage of sparsity assumption. It groups models and use integer code to denote groups, uniform code to encode models within each group. This idea brings similar results to EBIC [14] with $\lambda = 1$. A derivation is given in Section 4.6.1.

In the following section, we will develop another a working estimation $\hat{L}_3$ based on the minimax regret approach. We develop $\hat{L}_3$ using regret in a specific scenario. That is when the null model is true.

4.2.3 Stochastic Complexity corrected under the null model

Here we propose a novel $\hat{L}_3$ suitable for high dimensional problems. Basically, it corrects SCC under the null model assumption. Assume we have no knowledge of the $S_t$ prior to the dataset, we can not construct a coding theme that requires knowledge of $S_t$. To tackle this problem, we adopt a conservative belief that selecting no variables is better than selecting the wrong variables. Following this belief, we design a $L_3(\cdot)$ so that the null model would be favoured if it is the correct model. Ideally, we wish to be able to choose the null model when the null model is true. However, even when the null model is true, we do not want to be too certain. Otherwise, we would always select the null model regardless of what the true model is. To do this, we want to introduce some noise so that we would select the null model with high confidence when the null model is true.

We denote $S_0 = \emptyset$ as the null model structure and its corresponding null model with $\beta_{S_0}$ i.e. $\beta_{S_0,i} = 0$ for $i = 1, \cdots, m$. Let $P_0$ is a universal probability
space induced from permutations of \( y^n \). The permuted response vector \( y^n_0 \) has the identical distribution with \( y^n \) but is independent of \( X^n \). Define \( \Delta_S \) to be the difference of the first-two-part code length between \( S \) and \( S_0 \) with respect to \( y^n_0 \) generated from \( \mathcal{P}_0 \), i.e.

\[
\Delta_S = SCC(y^n_0, S) - SCC(y^n_0, S_0).
\]  

(4.15)

Note that \( \Delta_S \) here is a random variable as well. Then we set \( L_3(S_0) = c_0 \) and \( L_3(S) - c_0 \) to be the \( \gamma \) right quantile of \( \Delta(y^n_0, S) \), i.e.

\[
L_3(S) - c_0 = Q(\Delta_S, \gamma) = \sup \left\{ \delta \left| \Pr_{\mathcal{P}_0}(\Delta_S \leq \delta) < \gamma \right. \right\}.
\]  

(4.16)

This correction under the null model (4.16) is closely related to the minimax regret approach (4.14). Instead of using the maximum regret, we use the \((1 - \lambda)\) quantile. After all the null model is not the true model. If we use maximum regret, then we would always select the null model ignoring the true model.

Using quantile also gives us some inference ability lack by other model selection criterion. We are able to bound the probability of selecting a non-null model while the null model is true.

By quantile definition, we immediately have

\[
\Pr_{\mathcal{P}_0}(L_{\text{sparse}}(y^n, S) < L_{\text{sparse}}(y^n, S_0)) < \gamma
\]  

(4.17)

By Boole’s inequality, we have

\[
\Pr_{\mathcal{P}_0}(\hat{S} \neq S_0) = \Pr_{\mathcal{P}_0} \left( \bigcup_{S \in M} \{L_{\text{sparse}}(y^n_0, S) < L_{\text{sparse}}(y^n_0, S_0)\} \right)
\]

\[
\leq \sum_{S \in M} \Pr_{\mathcal{P}_0} \left( L_{\text{sparse}}(y^n_0, S) < L_{\text{sparse}}(y^n_0, S_0) \right)
\]  

(4.18)

At this point, the \( L_3(\cdot) \) is computable but not quite feasible to estimate. We can estimate the quantile (4.16) using permutation technique. However,
the size of \( M \) is \( 2^m \). To estimate quantile for one \( S \) would require \( O(2^m) \) number of permutations. To solve this problem, we present a marginal coding mechanism for the 3rd part of code as an estimation of (4.16). The idea is that \( S \) is essentially a set of \( k = |S| \) elements. The code of \( S \) can be a concatenation of code of all its elements, and the code length of \( S \) could be the sum of code length of each element plus the code length of the size of \( S \). Let \( \mathcal{X}_i \) denotes the model structure that only the coefficient of the \( i \)th covariate is non-zero. We have

\[
L_4^*(S) = \sum_{i \in S} L_4(\mathcal{X}_i) + \log^*(k) + c_0, \tag{4.19}
\]

where \( L_4(\mathcal{X}_i) \) is the code length of the \( i \)th covariate and where \( \log^* k = \log k + \log \log k + \cdots + c_0 \), and \( c_0 \approx 2.865 \).

Then we define the margin two part code length difference, \( \overline{\Delta}_i \) as

\[
\overline{\Delta}_i = SCC(y^n_0, \mathcal{X}_i) - SCC(y^n_0, S_0) \tag{4.20}
\]

Consequently, we set

\[
L_4(\mathcal{X}_i) = Q(\overline{\Delta}_i, \gamma). \tag{4.21}
\]

Finally, we have our Stochastic Complexity Criterion corrected for high dimension problems under the null model (SCCc) defined as

\[
SCCc(y^n, S) = SCC(y^n, S) + \sum_{i \in S} Q(\overline{\Delta}_i, \gamma) \tag{4.22}
\]

For simplicity, \( \log^*(k) \) and \( c_0 \) is \( O(1) \) when \( n, m \to \infty \) and therefore, are not included in SCCc.

4.2.4 Asymptotic properties of SCCc

First of all, SCC is asymptotically equivalent to BIC. In SCC, the determinate of sample Fisher information matrix \( |I_n| = O(n^k) \) when \( n \to \infty \). Then we
have $\log(|I_n|) = k \log(n) + O(1)$. Therefore, $2SCC = -2\ell + k \log(n) + O(1)$ as $n \to \infty$, and SCC is asymptotic equivalent to EBIC, i.e. $\lim_{n \to \infty} \frac{2SCC}{BIC} = 1$.

Secondly, $L_4$ is asymptotically linked to the quantile of $\chi^2_{df=1}$. Let $\delta_j = L_1(y^n|\hat{\beta}_{S_0}) - L_1(y^n|\hat{\beta}_{X_j})$. From likelihood ratio test[125], if the null model $S_0$ is the true model, then $2\delta_j$ follows the $\chi^2_{df=1}$ distribution for $j = 1, \cdots, m$ when $n \to \infty$. And we have

$$\lim_{n \to \infty} \hat{L}_4(\chi^2_{x_i}) = 1$$  \hspace{1cm} (4.23)

At last, we will show that SCCc is consistent with EBIC when $n, m \to \infty$. BIC and SCC are consistent for logistic regression with fixed $m$[94]. EBIC is consistent with $m = O(\exp(n^\kappa))$ as $n \to \infty$ for $\kappa \in (0, 1/3)$[15]. Here we show that under similar conditions of EBIC, if we gradually shrink $\alpha$ as $m$ increases, SCCc is consistent as well.

**Theorem 4.1.** If the following condition is satisfied, then SCCc is consistent.

- condition A1-A6 in [15] (listed in Section 4.6.3)
- $\frac{1}{\sqrt{n}} \leq \alpha \leq 1 - \frac{1}{\sqrt{4 \log(m) + 6 + 2}}, m > 2$,
- sparsity assumption: $|S_t| = O(\log(n))$

**Proof.** Chen’s Theorem 2[15] provide a range for the 2nd plus 3rd part code length to preserve consistency when $|s_t| = O(\log(n))$.

$$k \log(m) \leq L_2(\hat{\beta}|S) + L_3(S) \leq k(\log(m) + \frac{1}{2} \log(n)), \hspace{1cm} (4.24)$$

If we prove SCCc is bounded by (4.24), then SCCc is consistent. For SCCc, asymptotically we have

$$\lim_{n \to \infty} \left\{ L_2(\hat{\beta}_S) + \hat{L}_3^*(S) \right\} = \frac{k}{2} \log(n) + \sum_{j \in S_t} \left\{ Q(\chi^2_{df=1}, \gamma) - \frac{1}{2} \log(n) \right\}$$

$$= kQ(\chi^2_{df=1}, \gamma). \hspace{1cm} (4.25)$$
4.3. Numerical examples

Inglot[60] gives a both upper and lower bounds for $Q_z(\gamma)$, where $z \sim \text{Norm}(0,1)$ and $\gamma \leq 0.1$:

$$\sqrt{2 \log(1/\gamma)} - \frac{\log(4 \log(1/\gamma)) + 2}{2 \sqrt{2 \log(1/\gamma)}} \leq z \leq \sqrt{2 \log(1/\gamma)} - \frac{\log(2 \log(1/\gamma)) + 3/2}{2 \sqrt{2 \log(1/\gamma)}}$$

(4.26)

By set $\frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{4 \log(p)+6+2}}$, Equation (4.25) is bounded by Equation (4.24).

By Theorem 1, we have an asymptotic version of SCCc,

$$SCCc* = -\ell + \frac{k}{2} Q_{x_{\gamma,1}^2} (\alpha/p)$$

(4.27)

**Proposition 4.2.** If we perform forward model selection using SCCc, then (4.18) is bounded above by $\alpha = \gamma m$ as $n \rightarrow \infty$.

### 4.3 Numerical examples

We construct a series of GWAS case-control simulation studies to evaluate the performance of SCCc with logistic regression models. The interest is to find out whether the best model selected by SCCc includes all the important covariates in $S_t$, and whether the covariates that are not in $S_t$ are included in the model selected by SCCc as well. In particular, we emphasize the comparison of the results between SCCc and EBIC. EBIC is also designed to deal with the high dimensional model selection problem. EBIC has a three-component structure similar to that of SCCc. Further more, the third component of both methods are $O(k \log m)$. Asymptotically, SCCc can be regarded as a special case of EBIC with a corresponding $\lambda^*$. Through simulation, we will demonstrate that SCCc has an advantage in controlling variable selection error with varying $m$ over EBIC with fixed $\lambda$. In other words, SCCc provides an estimation for a suitable $\lambda$ for EBIC.
4.3. Numerical examples

4.3.1 Simulation setup

We use $Y$ to denote the binary trait in interest. The covariates $\{X_1, \ldots, X_m\}$ are single nucleotide polymorphisms (SNPs) taking values from $\{0, 1, 2\}$. The value of a SNP is the number of minor alleles (with respect to the population) that one patient has at a specific locus on the genome. We assume that $Y$ follows a logistic regression model, i.e.

$$\logit E(Y | x) = \beta_0 + \mathbf{x}^T \beta_t,$$  \hspace{1cm} (4.28)

where $\beta_t = (\beta_{t,1}, \ldots, \beta_{t,m})$ is the coefficient vector. Let $S_t$ denotes the index set of SNPs associated with $Y$, and we have $\beta_{t,i} = 0$ if $i \notin S_t$. Here we are interested to know which SNPs are statistically associated with $Y$, i.e. which $\beta_{t,i}$ is non-zero for $i = 1, \ldots, m$.

For each SNP $X_i$, its value is simulated through a binomial distribution, $\text{Binom}(2, \theta_i)$. $\theta_i$ is the frequency of the reference allele at $X_i$ with respect to the population. We further simulate values of $\theta_1, \ldots, \theta_p$ from an independent Beta distribution, $\text{Beta}(\alpha = 2, \beta = 6)$. We focus on a relevantly small sample size being a few hundreds sample points. This is fairly common between general GWAS studies. In fact, the limited sample size is often where the challenge lies.

To study the performance of SCCc, we compare AIC, BIC, SCC, EBIC, SCCc(exact version), SCCc*(asymptotic approximation) in several numerical examples. From the following numerical examples, we will show that SCCc is as good as selecting the pre-defined true model as EBIC with certain $\lambda$. The problem with EBIC is that a fixed value $\lambda$ will not provide consistent error control over an increasing $m$. whilst this is not a problem for SCCc.

We use average positive selection rates (PSR) and false discovery rates (FDR) as the measurements of the power and error rates of different model selection methods.
methods. For each model, we repeat simulation for 1000 times to calculate the average PSR and FDR. Let \( \hat{S}_j \) be the model structure selected by some IC in the \( j \)th replication, then

\[
\text{PSR} = \frac{1}{R} \sum_{i=1}^{R} \frac{|\hat{S}_j \cap S_t|}{|S_t|}
\]

\[
\text{FDR} = \frac{1}{R} \sum_{i=1}^{R} \frac{|\hat{S}_j \setminus S_t|}{|\hat{S}_j|}
\]

\[
\text{FWER} = \frac{1}{R} \sum_{i=1}^{R} 1(\hat{S}_j \setminus S_t \neq \emptyset)
\]

Under the null model hypothesis, i.e. \( S_t = S_0 \), FDR is equivalents to Family-wise Error Rate (FWER). We measure the chances of selecting correct model structures where \( S \) is a correct structure if \( S \supseteq S_t \). We then use \( \mathcal{M}_c = \{ S | S \supseteq S_t \} \) to denote all correct structures. This captures the cases of selecting the correct covariates but over-fitting the model.

### 4.3.2 Computational time

SCCc is computational intensive. In order to calculate the quantile in the third part of the code, usually 1000 times permutation of the response variable is require. According to (4.22), it would require additional \( 1000 \cdot m \) times of the time of evaluating a single variate regression. On the contrary, the computational time of asymptotical version SCCc* is equalivelenent to AIC and BIC which is trival.

### 4.3.3 Exhaustive search with small \( m \)

We first perform an exhaustive search simulation study for a low dimensional problem. The sample size is \( n = 500 \). There are \( m = 10 \) SNP covariates. We simulate two cases. In the first case, \( Y \) is generated from the null model \( S_0 \).
4.3. Numerical examples

In the second case, $Y$ is generated from a logistic model of two non-zero coefficients, $1.5, -1.0$. The samples of $Y$ are all case and control balanced. In each experiment, there are 1000 replicates.

Table 4.1 shows the results when the true model is the null model. Both SCCc and EBIC variants controls the FWER under targeted $\alpha = 0.05$, while BIC and SCC fail to do so. This suggests that even in low dimension, BIC and SCC may over fit the model.

<table>
<thead>
<tr>
<th>Method</th>
<th>FDR(FWER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>0.111</td>
</tr>
<tr>
<td>EBIC$_{\lambda=0.25}$</td>
<td>0.050</td>
</tr>
<tr>
<td>EBIC$_{\lambda=0.5}$</td>
<td>0.032</td>
</tr>
<tr>
<td>EBIC$_{\lambda=1.0}$</td>
<td>0.009</td>
</tr>
<tr>
<td>SCC</td>
<td>0.072</td>
</tr>
<tr>
<td>SCCc</td>
<td>0.002</td>
</tr>
<tr>
<td>SCCc*</td>
<td>0.041</td>
</tr>
</tbody>
</table>

Table 4.1: Exhaustive search result for $S_t = S_0$

Table 4.2 shows the results when two covariates are truly associated with $Y$. Among all the methods, SCCc and EBIC$_{\lambda=0.5}$ demonstrated the best performance in selecting the true model structure and both SCCc and EBIC variants are good at controlling the FDR to a low level. SCC and BIC have the highest chances of selecting a correct model at the cost of high FDR. In summary, when $m$ is small, both SCCc and EBIC variants are competitive in selecting the true model structure with the error rate FDR under control.
4.3. Numerical examples

|       | $Pr(\hat{S} = S_0)$ | $Pr(\hat{S} = S_t)$ | $Pr(\hat{S} \in \mathcal{M}_e)$ | $|\hat{S}|$ | FDR     |
|-------|---------------------|---------------------|-------------------------------|---------|---------|
| BIC   | 0.000               | 0.843               | 0.942                         | 1.942   | 0.117   |
| EBIC$_{\lambda=0.25}$ | 0.000               | 0.879               | 0.928                         | 1.928   | 0.057   |
| EBIC$_{\lambda=0.5}$  | 0.001               | 0.892               | 0.914                         | 1.913   | 0.024   |
| EBIC$_{\lambda=1.0}$  | 0.001               | 0.872               | 0.878                         | 1.877   | 0.007   |
| SCC   | 0.000               | 0.458               | 0.985                         | 1.985   | 0.726   |
| SCCc  | 0.003               | 0.895               | 0.915                         | 1.912   | 0.020   |
| SCCc$^*$ | 0.000               | 0.886               | 0.923                         | 1.923   | 0.041   |

Table 4.2: Exhaustive search result for $S_t = \{1, 2\}$
4.3. Numerical examples

4.3.4 LASSO search with large $m$

Now we study those model selection criteria under a high dimensional model selection problem. When $m$ is large, it is not feasible to evaluate all models in the model space $\mathcal{M}$. We adopt Chen’s strategy [15] by using LASSO [113] to generate a regularization pathway of models and then use model selection criteria to determine the best model along the pathway. We simulate cases of $m = 100, 1000, 10000, 100000$. For $m = 100000$, the exact SCCc calculation is omitted due to computational cost, only SCCc* is provided.

![Figure 4.1: FDR under the null model structure](image)

X-axis is the $\lambda$ of EBIC. Results of SCCc and SCCc* are shown on the far right in the plot. The sample size is 500 case control balanced. 1000 replicates.

First we study the scenario when $S_t = S_0$. Chen suggests a fixed $\lambda = 0.5$ for their EBIC based on their simulation results [14]. In Figure 4.1, both SCCc
and SCCc* have FWER controlled under the target \( \alpha = 0.05 \). However, FWER of \( \text{EBIC}_{\lambda=0.5} \) increases when \( m \) increases. Furthermore, the FWER curve of \( \text{EBIC} \) with varying \( \lambda \) shifts upwards as well. This indicates that \( \text{EBIC} \) fails to control FWER consistently with a fixed \( \lambda \).

![Under simulation model](image)

**Figure 4.2:** Power analysis

The sample size is 500 case control balanced. 1000 replicates. True model structure includes 8 covariates.

Then we consider a logistic model with \( S_t = \{1, 2, 3, 4, 5, 6, 7, 8\} \) with coefficients \( \beta_t = (1.5, -1.4, 1.3, -1.2, 1.1, -1.0, 0.9, -0.8, 0, \cdots, 0) \). Figure 4.2 shows results of all models selected with the PSR versus FDR. Both SCCc and SCCc* have consistent FDR control around 0.05 regardless of \( m \). All SCCc family is under the \( \alpha = 0.05 \) threshold. On the other hand, \( \text{EBIC}_{\lambda=0.5} \) does not have consistent FDR control just as the previous null model cases. At
last, points of SCCc and SCCc* all sit on the corresponding EBIC curve with varying $\lambda$. This suggests that SCCc can serves as an equivalent estimation of $\lambda$ with consistent error control.
4.4 Real data example

In this section, we examine different model criteria for logistic regression models through a case-control study on genetic association on breast cancer. Cases and controls of the dataset come from Australia Breast Cancer Family Study[79] and Australian Mammographic Density Twins and Sisters Study[87] respectively. The dataset has 491 individuals (cases and controls). The genetic variation is measured in 366 SNPs with a Human610-Quad bead-chip array.

![LASSO regularization path of the breast cancer dataset](image)

**Figure 4.3:** LASSO regularization path of the breast cancer dataset

Path of top 50 SNPs is shown. The upper x-axis scale is the number SNPs included in the logistic regression model.

To compare how different model selection criteria would select their best logistic regression models, we first use LASSO[113] ("glmnet" package [37] in R) to obtain a LASSO regularization pathway (Figure 4.3) of logistic models.
Then we rank the top 50 SNPs according their number of appearances in the pathway.

| Score  | $|\hat{S}|$ |
|--------|-----------|
| AIC    | 620.1     | 16       |
| BIC    | 652.6     | 4        |
| 2 SCC  | 639.2     | 8        |
| EBIC$_{\lambda=0.25}$ | 658.6 | 1 |
| EBIC$_{\lambda=0.5}$  | 661.5 | 1 |
| EBIC$_{\lambda=1.0}$  | 666.6 | 0 |
| 2 SCCc$_{m=366}$       | 664.0 | 1 |
| EBIC$_{\lambda=0.25,m=1M}$ | 662.5 | 1 |
| EBIC$_{\lambda=0.5,m=1M}$ | 666.6 | 0 |
| EBIC$_{\lambda=1.0,m=1M}$ | 666.6 | 0 |
| 2 SCCc$_{m=1M}$         | 666.6 | 0 |

Table 4.3: Best model selected in the breast cancer dataset
SCC and SCCc are scaled by 2 for comparison with AIC, BIC, and EBIC. 1M = 1,000,000.

Then along the pathway, we perform forward model selection of those top 50 SNPs. AIC, BIC, EBIC, SCC, SCCc are calculated for all 50 models. Table 4.3 shows the best model selected according to different criteria. We find that SCCc and EBIC are the most conservative while AIC is the most liberal followed by SCC. The top SNP is rs3778080. As discussed in Section 2.5.2, according to multiple testing, rs3778080 is the only SNP that is dataset wise significant but not genome wide significant. EBIC and SCC both support this result by choosing different $m$. However, EBIC still requires a selection of suitable $\lambda$ while SCC provides consistent results without any parameters.

At last, we find that SCCc is reliable model selection criterion which can be well adapted to high dimension model selection problems in real applications.
4.5 Conclusion and discussions

In the high dimensional variable selection problems, we have introduced a novel model selection criterion, SCCc, for logistic regression model selection. SCCc is derived from SCC using the MDL principles. SCC is an estimation of the shortest two-part-code of the data which compresses the data with the help of a family of statistic parametric models. SCCc has an additional part of code that specifies the non-zero parameters. This allows SCCc to save substantial code length from SCC if the dimension of the parameter space is large and the parameter vectors are sparse. As a part of SCCc, we have proposed a computable code length for the third part of code derived from the null model where no covariates is associated with the response variable. This allows SCCc to control the Type I error, selecting a non-null model when null model is true. By doing so we hybrid model selection and model inference together, which is absent from other popular model selection criteria such as AIC, BIC. We have further shown that when $m \gg n$ and $n \to \infty$, our SCCc is able to select the true model structure with conditions.

In the numerical examples, SCCc is compared with EBIC, a Bayesian model selection criterion which is also developed for high dimensional problems. EBIC has an additional parameter, $\lambda$, which controls the degree of penalization on the total number of covariates $m$. We have found that a fix-valued $\lambda$ fails to consistently control the error rate i.e. FDR as $m$ increases. However, SCCc with no parameters retains such ability. On a case-control genotype dataset on breast cancer, both SCCc and EBIC can adjust for both dataset significance and genome-wide significance. SCCc proves consistent results with multiple testing procedures, while EBIC gives ambiguous results with a different selection of $\lambda$. 
Various generalizations and improvements can be made to SCCc. Firstly, SCCc can be easily applied to the family of exponential parametric models to suit different types of datasets. In addition, with an estimation of the effective degrees of freedom [38, 52], it can be applied to non-parametric models as well. Thirdly, we can use bootstrap techniques [45] to calculate the quantile under the empirical distribution instead of using permutation techniques for the null model. This would bring further improvements in terms of error control and variable selection power directly under the non-null true model. At last, we could apply SCCc at a much larger scale on the genome rather than a small region and test its performance on capture multiple weak signals with dependent structures.
4.6 Supplementary

4.6.1 A MDL approach of EBIC

Here we demonstrate that by using the MDL “luckiness” principle, we can derived a similar model selection criterion to EBIC. Since $M$ is only a finite set, the simplest approach to use uniform codes[51], i.e. $L_3^u = \log |M|$. However, with the equal length, $L_3$ is indifferent when comparing two models. Sparsity assumption and luckiness principle[47] is a perfectly matching couple and could be sued to solve this problem. Luckiness principle partitioned $M$ into multiple subsets, and order them in a pre-determined way. It uses an index code to denote the subsets and universal code for the elements in the. If we were “lucky”, the code for $S$ in higher order sets would be substantially shorter. Even for $S$ in the lowest order subset, the code would only slightly longer. If we choose the subsets to be models with same degrees of freedom and the per-determined order is from small degrees of freedom to large ones, then we can derive a $L_3$ that is similar to EBIC which imposes $k \log m$ as the equivalent $L_3$.

Let $\{M_i\}$ be a family of subsets of $M$, and nested according to their sparsity, i.e. $\bigcup_{i=0}^p M_i = M$, where $M_i = \{S | |S| \leq i\}$, and we denote $k = |S|$. In order to code $S$, we first encode $k$ as integers, then use uniform code for $S$ given $M_k$.

\[
L_3^1(S) = \log |M_k| + \log^* k \tag{4.29}
\]

where $\log^* k = \log k + \log \log k + \cdots + c_0$, and $c_0 \approx 2.865[99]$. If we were “lucky” i.e. $k$ is small, in other words $S$ is sparse, then $L_3^1(S)$ would be substantially shorter than $L_3^u(S)$. Even if the worst scenario, $k = m/2$, $L_3^1(S)$ is no more than $\log^* p$ longer than $L_3^u(S)$.
When \( m \gg k \), by Stirling approximation, \( \log \binom{m}{k} \approx k \log p + k - k \log k \). \( k \log m \) is the dominated term in \( L_3(S) \); since information criterion is for comparing models rather than estimate the exact MDL, we drop the rest terms, and have

\[
\hat{L}_3(S) = k \log m + k - k \log k + \log^* k \approx k \log m \tag{4.30}
\]

Meanwhile, EBIC has an equivalent penalize term \( \gamma k \log m \), where \( \gamma \in [0, 1] \). It is an approximation of \( \gamma \log \binom{m}{k} \). In some sense, the luckiness principle gives an MDL derivation of the EBIC.
4.6.2 Extended BIC conditions

Extended BIC consistency conditions for logistic regression[15]. These conditions are used in Theorem 4.1. Denote $\beta_S(S) = \{\beta_{S,j} | j \in S\}$, and $H_n(\beta_S) = \sum_{i=1}^{n} \sigma_i^2 x_i x_i^\text{intercal}$, where $x_i$ is the $i$th row of $X$, and $\sigma_i^2$ is the corresponding variance of $x_i^\text{T} \beta_S$. $|S| < K$ for some constant $K$. They are

A1 As $n \to \infty$, $m = O(n^\kappa)$ for some $\kappa > 0$.

A2 The size of $S_t$, $|S_t|$ is fixed.

A3 Let $B$ denote the parameter spaces of $\beta$. The interior of $B$ is not empty and $\beta_t \in B$.

A4 There exists $c_1, c_2 > 0$ so that for all sufficiently large $n$,

$$c_1 \leq \lambda_{\text{min}}(n^{-1}H_n(\beta_t(S))) \leq \lambda_{\text{max}}(n^{-1}H_n(\beta_t(S))) \leq c_2,$$

for all $S$ so that $|S| \leq K$, where $\lambda_{\text{min}}$ and $\lambda_{\text{max}}$ denote the smallest and the largest eigenvalues respectively.

A5 For any $\epsilon > 0$, there exists some $\delta > 0$ so that, when $n$ is sufficiently large,

$$(1 - \epsilon)H_n(\beta_t(S)) \leq H_n(\beta_S(S)) \leq (1 + \epsilon)H_n(\beta_t(S))$$

for all $S$ and $\beta_S$ so that $|S| \leq K$ and $||\beta_S(S) - \beta_t(S)||a$.

A6 Denote by $x_{ij}$ the $j$th column and $i$th row of the design matrix $X$.

$$\max_{1 \leq j \leq m} \max_{1 \leq i \leq n} \left\{ \frac{x_{ij}^2}{\sum_{i=1}^{n} x_{ij}^2 \sigma_i^2} \right\} = o((\log n)^{-1})$$
4.6.3 Rare genetic variants

In this section we show that the difference between asymptotic SCCc and permuted SCCc is significantly larger for rare genetic variants than for the common ones. Rare genetic variants in GWAS refer to SNPs with low MAF, i.e. low \( \eta \). With the application of the second generation sequencing technology, we are able to capture more genetic variation in the genome, and among large proportion of it is “rare” variants [44]. Table 4.4 shows how quantiles of log likelihood regrets varies with respect to their asymptotic approximation for rare variants.

Table 4.5, shows how type I error gets inflated in small \( p \) scenario with the same settings in Section 4.3.3.

Figure 4.4 and 4.5 demonstrates how rare variant affects the error control and power of SCCc in large \( p \) scenario. the sample size is 200 with case control balanced. 1000 replicates. All covariates are simulated SNPs with \( \eta = 0.05 \). As we illustrated in Table 4.5, SCCc failed to control FWER at \( \alpha = 0.05 \).

In Figure 4.5, there are two causal SNPs, with coefficient \((0.8, -0.7)\). Again, the error control is not as good as the results on non-rare MAF SNPs in Figure 4.5.
### Table 4.4: Exact log likelihood regret $\delta$ simulation

One SNP is generated with some MAF and sample size each time. $Y$ is independently generated with case control balance. The log likelihood regret is the difference between the null model log likelihood and the alternative model includes the unrelated SNP. 1000 replicates are generated for each scenario.

<table>
<thead>
<tr>
<th>MAF</th>
<th>Sample size $\gamma$</th>
<th>5e-02</th>
<th>5e-03</th>
<th>5e-04</th>
<th>5e-05</th>
<th>5e-06</th>
<th>5e-07</th>
<th>5e-08</th>
<th>5e-09</th>
<th>5e-10</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>100</td>
<td>4.12</td>
<td>6.15</td>
<td>7.04</td>
<td>7.24</td>
<td>7.27</td>
<td>7.27</td>
<td>7.27</td>
<td>7.27</td>
<td>7.27</td>
<td>7.27</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>2.65</td>
<td>5.03</td>
<td>7.41</td>
<td>9.93</td>
<td>12.35</td>
<td>14.85</td>
<td>17.42</td>
<td>19.95</td>
<td>22.4</td>
<td>29.10</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>2.68</td>
<td>5.02</td>
<td>7.48</td>
<td>9.89</td>
<td>12.37</td>
<td>14.83</td>
<td>17.35</td>
<td>19.82</td>
<td>22.24</td>
<td>49.99</td>
</tr>
<tr>
<td>0.25</td>
<td>100</td>
<td>2.68</td>
<td>5.02</td>
<td>7.48</td>
<td>9.89</td>
<td>12.37</td>
<td>14.83</td>
<td>17.35</td>
<td>19.82</td>
<td>22.24</td>
<td>49.99</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>2.65</td>
<td>5.03</td>
<td>7.41</td>
<td>9.93</td>
<td>12.35</td>
<td>14.85</td>
<td>17.42</td>
<td>19.95</td>
<td>22.4</td>
<td>29.10</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>2.26</td>
<td>4.44</td>
<td>6.68</td>
<td>8.93</td>
<td>11.25</td>
<td>13.54</td>
<td>15.88</td>
<td>18.17</td>
<td>20.54</td>
<td>199.57</td>
</tr>
<tr>
<td>0.45</td>
<td>100</td>
<td>2.57</td>
<td>4.88</td>
<td>7.28</td>
<td>9.71</td>
<td>12.14</td>
<td>14.59</td>
<td>17.06</td>
<td>19.60</td>
<td>22.13</td>
<td>36.50</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>2.36</td>
<td>4.56</td>
<td>6.85</td>
<td>9.18</td>
<td>11.48</td>
<td>13.84</td>
<td>16.20</td>
<td>18.55</td>
<td>20.91</td>
<td>72.41</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>2.22</td>
<td>4.37</td>
<td>6.60</td>
<td>8.87</td>
<td>11.14</td>
<td>13.42</td>
<td>15.73</td>
<td>18.04</td>
<td>20.35</td>
<td>144.52</td>
</tr>
</tbody>
</table>

Asymptotic $\chi^2$ Quantile: 1.92, 3.93, 6.05, 8.22, 10.4, 12.6, 14.8, 17.0, 19.3, $\infty$
Table 4.5: Type I error comparison with different allele frequency(AF), $S_t = S_0$

$\eta \sim \text{Beta}(2,6)$  \hspace{1cm} $\eta = 0.05$

<table>
<thead>
<tr>
<th></th>
<th>BIC</th>
<th>0.114</th>
<th>0.154</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBIC$\lambda=0.25$</td>
<td>0.064</td>
<td>0.088</td>
<td></td>
</tr>
<tr>
<td>EBIC$\lambda=0.5$</td>
<td>0.033</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>EBIC$\lambda=1.0$</td>
<td>0.008</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>BICc</td>
<td>0.019</td>
<td>0.026</td>
<td></td>
</tr>
<tr>
<td>SCC</td>
<td>0.746</td>
<td>0.818</td>
<td></td>
</tr>
<tr>
<td>SCCc</td>
<td>0.018</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td>SCCc*</td>
<td>0.047</td>
<td>0.072</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.4: FDR under the null model structure with rare variants
Figure 4.5: Power analysis for rare variants
In this thesis, we have developed three novel statistical methods to tackle three GWAS challenges. These methods are not isolated but are built onto each other which ultimately serve the same goal, which is to find a multi-loci predictive model that can explain the variations of the phenotype linked to the genotypes. The first method, Rank Stability Selection, is a non-parametric extension of the standard multiple testing widely used in GWAS. It examines the effect of one particular genetic variation on the phenotype at the time. The second method, the Error Correction Code dimension reduction method extends the scope from individual genetic variants to haplotype blocks. It examines independent information in terms of linear combinations of the genetic variants within a haplotype block. The last method, the Stochastic Complexity Criterion correction lays out a model selection foundation and it is also a convenient tool for us to explore more complex statistical GWAS models defined on the whole genome consisting of multiple genetic variants.

The biggest obstacle in this research is the curse of dimensionality. With finite number of objects, high dimensionality links to sparsity. The higher the dimension, the farther the objects are apart from each other and therefore, it is more difficult to gain statistical insights. It also brings challenges in computation. For multivariate model selection, the model space increases exponentially to the number of variate. The high dimensionality directly comes from the data. In GWAS, it is essentially the millions of SNPs.
We think a good method should scale well when the dimensionality of the problem increases. These three methods also demonstrate our continuous efforts in fighting against the curse. RSS has an efficient resampling algorithm that enables non-parametric statistical analysis with limited computation resource. ECC hypothesis directly reduces the dimensionality by creating intermediate surrogate variables preserving the association relationship with the phenotype. And SCCc refines the model selection tools for complex statistical model selection on ultra-high dimensions.

We also think a good statistical method should be relatively simple and easy to use. For RSS, due to its unique resampling algorithm design, it can also work with a wide range of association test statistics. Further more, estimation and adjustment for population stratification is not necessary. Meanwhile, one good advantage of SCCc over EBIC is that it does not require a tuning parameter but has a consisting model selection error control ability.

At last, these three methods have good potential for future research. RSS explores the concept of statistical stability in hypothesis testing. The sample splitting technique can also be applied to model selection criterion, for example, evaluating model selection criterion over subsamples. This would extend stability inference to complex association statistical models instead of hypothesis testing on marginal association statistics. Currently, the ECC dimension reduction only applies to haplotype blocks. If we can construct an experiment to test ECC in wider regions such as genes and pathways, then we can further reduce the dimensionality of the genotype data. For SCCc, it is merely a measurement to compare the goodness of two models. However, given the enormous size of the model space, we need smart ways to search the model space. For example, a stochastic search like Gibbs sample would significantly speed up the process[95].
Bibliography


Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Zhou, Zeyu

Title:
Statistical testing and selection by re-sampling in genome-wide association studies

Date:
2016

Persistent Link:
http://hdl.handle.net/11343/127538

File Description:
Complete Thesis

Terms and Conditions:
Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.