# A Geometry-Based Multiple Testing Correction for Contingency Tables by Truncated Normal Distribution

by

**Tapati Basak**

**Thesis Supervisor: Professor Ryo Yamada, M.D., Ph.D.**

**DISSERTATION**

Submitted to the Unit of Statistical Genetics

of the Graduate School of Medicine in partial fulfillment of

the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

in

**Medicine**


**Kyoto University, Japan**

February 2019

# *Dedication*

To my son, Sadya, wise beyond his age

and

To my beloved husband, Dinesh, infinitely supportive

# Declaration

I hereby declare that this Ph.D. thesis entitled "A Geometry-Based Multiple Testing Correction for Contingency Tables by Truncated Normal Distribution" was carried out by me for the degree of Doctor of Philosophy in Medicine under the guidance and supervision of Professor Ryo Yamada, M.D., Ph.D., Statistical genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Japan.

This is an original work and has not published anywhere. For the present thesis, which I am submitting to the University, no degree or diploma or distinction has been conferred on me before, either in this or in any other University.

Place: Kyoto, Japan (Tapati Basak)
Date: February 2019 Ph.D. Student

# Acknowledgements

There are many people that I would like to thank who have supported me throughout the graduate school. First and foremost, I want to thank my supervisor, Professor Ryo Yamada, for his guidance and support over the past four years. Dr. Yamada provided me with excellent supervision and encouragement throughout my graduate training. He knows my weakness and helps me to improve my research experience. Without his patience and guidance, my thesis work would not have been accomplished.

I would like to thank the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan for providing the financial support of my research with the Monbukagakusho Scholarship.

I would like to thank all the past and current members of the Statistical Genetics lab and all the members of Human Disease Genomics, who worked to maintain the database of the Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. I especially thank Mr. Kazuhisa Nagashima, Mr. Satoshi Kajimoto and Mr. Takahisa Kawaguchi for their contributions to my thesis project.

Lastly, I would like to thank my family who has supported me over the last four years. My husband and son are with me throughout the graduate school with their patience, understanding, and support. Without them, I would not have accomplished my graduate school goal. Also, my mother always encourages me on my way to achieving goals.

This work was supported by the Core Research for Evolutional Science and Technology (CREST) (Grant number: JPMJCR1502) from Japan Science and Technology Agency (JST). The Nagahama study was supported by a university grant; the Center of Innovation Program; the Global University Project; Grant-in-Aid for Scientific Research (Grant Numbers: 25293141, 26670313, 26293198, 17H04182, 17H04126) from the MEXT in Japan; the Practical Research Project for Rare/Intractable Diseases (Grant Numbers: ek0109070, ek0109196, ek0109348); the Comprehensive Research on Aging and Health

Science Research Grants for Dementia R&D (Grant Number: dk0207006); the Program for an Integrated Database of Clinical and Genomic Information (Grant Number: kk0205008); the Practical Research Project for Life-Style related Diseases including Cardiovascular Diseases and Diabetes Mellitus (Grant Numbers: ek0210066, ek0210096) from Japan Agency for Medical Research and Development (AMED); and the Takeda Medical Research Foundation.

# Abstract

Inference procedure is a critical step of experimental researches to draw scientific conclusions especially in multiple testing. The false positive rate increases unless the unadjusted marginal $p$-values are corrected. Therefore, a multiple testing correction is necessary to adjust the $p$-values based on the number of tests to control type I error. We propose a multiple testing correction of MAX-test for a contingency table, where multiple $\chi^2$-tests are applied based on a truncated normal distribution (TND) estimation method by Botev. The table and tests are defined geometrically by contour hyperplanes in the degrees of freedom ( df ) dimensional space. A linear algebraic method called spherization transforms the shape of the space, defined by the contour hyperplanes of the distribution of tables sharing the same marginal counts. So, the stochastic distributions of these tables are transformed into a standard multivariate normal distribution in df-dimensional space. Geometrically, the $p$-value is defined by a convex polytope consisted of truncating hyperplanes of test's contour lines in df-dimensional space. The TND approach of the Botev method was used to estimate the corrected $p$. Finally, the features of our approach were extracted using a real GWAS data.

**Keywords: Contingency table; Convex polytope; MAX-test; Multiple testing; Type I error; Truncated normal distribution**
MSC 2010 Classification Codes: **62H17, 52B40, 62F03, 62H10**

# Contributions

**Chapter 1:** This chapter is the original contribution by the author.

**Chapter 2:** The content of sections 2.4.1.2 and 2.5 are the original contributions.

**Chapter 3:** The section 3.4 is the original contribution by the author.

**Chapter 4:** This chapter contains a review work.

**Chapter 5:** The simulated example of integral evaluation in section 5.2.1 is the original contribution by the author.

**Chapter 6:** The total contents of this chapter are the original contributions.

**Chapter 7:** This chapter contains an original contribution.

**Chapter 8:** This chapter is the original contribution.

**Chapter 9:** The contents of this chapter are the original contribution by the author.

# Contents

# List of figures

# List of abbreviations

| Abbreviations | Phrase |
|---|---|
| $p_{sig}$ | $5.0 \times 10^{-08}$ |
| $H_0$ | Null hypothesis |
| $H_a$ | Alternative hypothesis |
| BC | Barycentric |
| $cdf$ | Cumulative distribution function |
| CATT | Cochran-Armitage trend test |
| df | Degrees of freedom |
| GWAS | Genome-wide association studies |
| GLM | Generalized linear model |
| IS | Importance sampling |
| LD | Linkage disequilibrium |
| MAX-test | Maximum test |
| MERT | Maximin efficiency robust tests |
| MHT | Mantel-Haenszel test |
| M-O-C-S | Maximum-of-chi-squares |
| $pdf$ | Probability density function |
| Sph-Btv | Spherization and Botev's approach |
| SNP | Single-nucleotide polymorphism |
| SOV | Separation of variables |

| | |
|---|---|
| S-O-C-S | Sum-of-chi-squares |
| TND | Truncated normal distribution |
| VRE | Vanishing relative error |

# Chapter 1

# Introduction

## 1.1 Statistical background

Multiple testing problem occurs when a set of simultaneous statistical tests are considered. In many situations, more than one or even a large number of hypotheses are simultaneously tested, which is referred to as multiple comparisons (Hsu 1996). For example, in case of high-dimensional data obtained from the field of genetics, medicine, molecular biology, bioinformatics, agricultural science etc. (Johnson et al. 2010). In statistical hypothesis testing, a pre-defined confidence level generally applies only to each individual test. But, multiple testing considers the same confidence level for the whole family of simultaneous tests. The probability of committing false statistical inferences considerably increase when more than one hypothesis is simultaneously tested at a given confidence level (McDonald 2014; Chen et al. 2017). In general, where all the null hypotheses ($H_0$) are independent and also supposed to be true, the statistical inference of committing at least one false rejection will become inevitable even if 100 hypotheses are individually tested at confidence level $\alpha$ =0.05. Estimation of this error rate is more complex if the hypotheses are correlated and not all of them are true (Chen et al. 2017). Hence, the unadjusted or marginal $p$-values are no longer a useful quantity for this inference process as we are testing many features at the same time.

For example, the genetic tendency to develop a disease can be defined from the association tests. Individual SNP tests are performed in GWAS with a genetic model [Bush & Moore, 2012], and the same cutoff $p_{sig}$ or $Pr$(genome-wide type I error) = $5.0 \times 10^{-08}$ is used for all SNP tests with consideration of control type I error [Kanai et al., 2016]. A gene is

reported to be associated with a phenotype when any SNP of the gene has been identified as positive relative to this $p_{sig}$. However, the structure of the genes are not even over the genome from the statistical standpoint based on the total number of SNPs per gene. As a result, genes having different numbers of SNPs have different probabilities of being detected as false positives in association studies. For example, if a gene has no SNP in it, there is no chance of it being found associated and no chance of being detected falsely. If a gene has 1 SNP in it, there is a small chance of it being detected falsely. If a gene has more than 1 SNP, the probability of type I error becomes larger. In addition to that, the LD pattern in each gene varies substantially, so, the LD structure of individual genes must be considered in this relationship between the numbers of SNPs per gene and $p$ because the SNP tests are mutually dependent on the LD structures of genes, and the degree of dependency per gene is heterogeneous over the genome. This heterogeneity directs the distribution of $p$-values, as well as type I error. In general, $p$ becomes higher for a gene with weaker LD structure or more SNPs. Also, this violation of independence assumption limits the conventional correction approach's ability to control type I error efficiently.

## 1.2 Problems statement

The correction of $p$-value is obvious for multiple testing to control the type I error (Bender and Lange 2001). Many correction methods are proposed. However, these corrections are not always easy and straightforward in terms of the selection of proper method suitable for the various experimental properties and study purposes (Chen et al. 2017). For example, the genome-wide association studies (GWAS) consider simultaneous testing of individual single-nucleotide polymorphism (SNP) of a gene and picks up positive genes when any one of SNPs in the gene is associated with a trait (Bush and Moore 2012). The contingency table tests are generally used to analyze the dichotomous case-control traits, where the deviation is measured from independence under $H_0$ of no association between the phenotype and SNP genotype classes. One of the most celebrated form of this test is the popular chi-square ($\chi^2$) test (Bush and Moore 2012). Selection of the largest test statistic (MAX-test) from different genetic models is a powerful approach as it provides safeguard against model uncertainty (González et al. 2008).

Due to rapid advances in genotyping technology and standardized guidelines for reporting statistical evidence, the multitude of comparisons made in a GWAS may result false positives (type I errors). When testing millions of SNPs in a single gene association study will become the standard, consideration of multiple comparisons is an essential part of determining statistical significance (Johnson et al. 2010). Moreover, correction of multiple testing $p$-values are also required for the SNPs in linkage disequilibrium (LD) with each other (Nyholt 2004). The closely spaced SNPs frequently yield high correlation because of extensive LD. Therefore, when association studies are conducted with many SNPs, the tests performed on each SNP are usually not independent, depending on the correlation structure among the SNPs. This violation of the independence assumption limits the Šidák and Bonferroni correction's ability to control the type I error effectively (Gao et al. 2008). The permutation test can be applied (Ritchie et al. 2001; Hoh et al. 2001). But, an enormous number of permutations is required to accurately estimate small $p$-values, constituting a computational burden (Knijnenburg and Wessels 2009).

We combined the geometry of multi-way contingency table (Slavkovi´c and Fienberg 2010) and the $\chi^2$-test (Yamada and Okada 2009; Hirosawa et al. 2010) to develop a linear algebraic transformation called spherization. Geometrically, the table and tests are defined by their contour hyperplanes in a df-dimensional space. The spherization is used to convert the shape of the space, defined by the contour hyperplanes of the distribution of tables sharing the same marginal counts. Hence, the stochastic distributions of tables are transformed into a standard multivariate normal distribution in df-dimensional space that is able to address tables with smaller $p$-values because of their symmetricity with respect to distance and direction. If MAX-test is performed for a set of $H_0$, the tables with the same test statistic values are located in the df-dimensional space in the shape of a convex polytope, in which the corrected $p$-values correspond to the sum of the probability out of the convex polytope of the truncated normal distribution (TND). The TND approach by Botev (2017) was applied to estimate the multiple testing corrected $p$-values. The method was applied to a real GWAS data and the additive model tests of individual SNPs were repeated for a SNP set.

## 1.3 Aims and scope

The main aims of this research are:

i) to propose a method of estimating the multiple testing corrected $p$-values for the MAX-based tests of the contingency tables.

ii) to extract the features of the proposed test using a GWAS real data.

The key advantage of the new method is the ability to select the appropriate number of independent markers to be tested by defining a variance-covariance matrix, which can handle multiple $\mathrm{df} = 1$ tests in a unified fashion at higher dimensions by considering the individual structures of genes.

## 1.4 Thesis overview

The thesis is organized into nine chapters.

Chapter 2 presents the detailed of both the algebraic and the geometric presentation of contingency tables using simplices and polytopes. A general relationship between the contingency tables, simplices and the barycentric (BC) coordinates is also illustrated here, which aims to search the answer of the question that how the BC coordinates can be used to define the distribution of the tables inside the simplex. The simplex-based presentation of tables in the df-dimenaional space is also presented here. This chapter concludes how the joint distribution of the barycenter can completely determine the distribution of the tables for fixed conditional probabilities and marginal counts and how the simplex-based rotational geometry can be used to present the categorical variables and their combinations in a geometric space.

Chapter 3 represents the description of the most celebrated maximum test (MAX-test) and its distribution. A brief description of the mathematical derivation of the MAX-test in the context of three popular genetic models (additive, dominant and recessive) is given along with its underlying asymptotic distribution. The generalized MAX-test for multiple additive models that was used in this research is also outlined here. In summary, the knowledge

about the distribution of the MAX-statistic could provide a useful solution for the computational burden issue that arises from the estimation of the $p$-values using the conventional permutation procedure and the geometric definition of the test provide a powerful insight about the underlying distribution.

Chapter 4 discusses about some existing literatures on GWAS multiple testing that are considered as the developments and the modifications of the MAX-tests from different perspectives. A general conclusion that can be drawn from this literature is that most of the existing tests are two-sided and these tests are not suitable when the alternative hypothesis $(H_a)$ becomes restricted. Also, defining the underlying distribution of the MAX-test statistic is a big challenge when the mode of inheritance and the population variance are unknown.

Chapter 5 aims to discuss about the accuracy improvement of the Monte Carlo Integration through the application of importance sampling (IS) via minimax tilting approach. A comparison of the accuracy of the simple Monte Carlo and the IS methods for estimating an integral is demonstrated with a numerical example. The discussion concludes that the use of IS along with the minimax tilting provides an estimator with relatively smaller variation in estimation.

Chapter 6 proposes a linear algebraic transformation called spherization for the shape of the space defined by the table's and the test's contour lines. A brief description of this technique for the geometry of tables and $\chi^2$-statistic to develop the spherical geometry of the $\chi^2$-statistic for the proportion trend test is given. The geometric features of the shape of the truncated distribution in the transformed space also presented here. The idea generalization from the single to multiple testing is illustrated in the context of algebraic and geometric concepts. In conclusion, the spherization overcomes the difficulties of handing asymmetric space of the distribution of tables by converting to symmetric one using the rotational geometry of spheres.

Chapter 7 describes the proposed Sph-Btv approach and the experimental designs to evaluate the method. The experimental design includes a brief description of the real data.

The web source about the availability of the real data and R-codes used in this study are also provided here. The Sph-Btv method is a combination of the spherization and the minimax tilting approach that can be applied to multiple testing problems in higher dimensions.

Chapter 8 presents the results and a discussion of the new method evaluation. This chapter provides the $p$-values estimated for the SNP subsets characterized in Chapter 7 by the Sph-Btv approach. The assessment of $p$-values considering the gene-wise structures for multiple $\chi^2$-cutoffs is presented here. The evaluation indicates that the $p$-values are heterogeneous over the genome based on the gene's individual structure.

The final chapter contains a brief summary with some concluding remarks. The limitations and some future directions of this research are also outlined here.

## 1.5 Computations

The preprocessing of the real genotype data (viz. marker selection, recoding, pairwise $r^2$ calculation, etc.) were carried out in PLINK 1.9. All the programs used to generate the results in this thesis were written in R programming language System Version 3.2.4. To write R-codes for our proposed method, the R built function were used under the package, "TruncatedNormal" [Botev, 2017] of version 1.0.

# Chapter 2

# Simplicial geometry of contingency tables

## 2.1 Introduction

This chapter presents the geometrical concepts of contingency tables and its relation to the simplex. Beginning with [Fienberg, 1968] and [Fienberg & Gilbert, 1970], several authors have exploited the geometric representation of contingency tables, in terms of quantities such as margins and odds ratios, both for the proof of statistical results and to gain deeper understanding of models used for contingency table representation. The geometric structure of statistical models for case-control studies was given by [Carlini & Rapallo, 2005]. There has also been considerable recent interest in geometric descriptions of contingency table models and analytical tools, from highly varying perspectives [Slavković & Fienberg, 2010]. This is a simplex-based geometric view of the tables using the BC coordinates system [Möbius, 1827] which is a simplification of table's geometry by using the meaningful coordinates. The BC coordinates help to define the distributions of the tables from table geometry by handling the points inside a simplex where the generalized BC coordinates are defined in terms of a polytope (generalized form of a simplex in higher dimensions). This concept of simplex-based geometry of tables can also be demonstrated in the $df$-dimensional space of the corresponding tables which is smaller than the original dimension that the tables have. This geometry provides the simplest way to present the categorical variables and their combinations of a table by simplices and rotation matrices. The rotational geometry provides a useful way for dimension reduction and to transform the space of the stochastic distribution of the tables to a symmetric space which is easier to handle from statistical viewpoints.

The organization of the present chapter is as follows. Section 2.2 represents the algebraic and the geometric description of the polytopes and section 2.3 explores the concept of the BC coordinates system, respectively. In section 2.4, we introduce the algebraic and geometric interpretation of contingency tables in terms of simplex and BC coordinates. Section 2.5 demonstrates the simplex-based geometry of tables in $df$-dimensional space. Finally, Section 2.6 contains some concluding remarks.

## 2.2 Simplicial geometry of polytopes

A polytope is a geometric object defined from vertices, edges, faces, cells and so on (Figure 1). The possible ways to define a polytope includes simplex (the equilateral triangle and the regular tetrahedron), hypercubes or measure polytopes (the square and the cube), orthoplexes or cross polytopes (the square and regular octahedron).

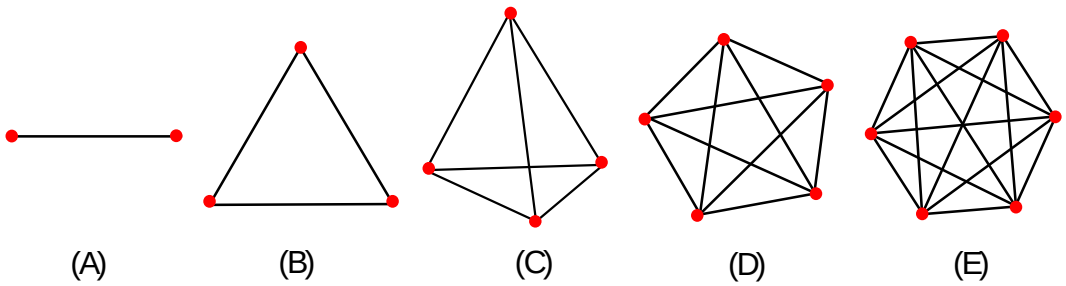



**Figure 1: The simplices.** (A) 1-simplex. (B) 2-simplex. (C) 3-simplex. (D) 4-simplex. (E) 5-simplex.

A simplex is a generalization of the notion of a triangle or a tetrahedron to arbitrary dimensions. The boundary of a $n$-simplex has $(n+1)$ 0-faces (polytope vertices), $n(n+1)/2$ 1-faces (polytope edges), and $\binom{n+1}{m+1}$ $m$-faces, which is a binomial coefficient where $(m+1)$ are the subset of the $(n+1)$ points. Specifically, a $n$-simplex is an $n$-dimensional polytope which is the convex polytope of its $(n+1)$ vertices can be denoted by $\Delta^n$. For example, 0-simplex is a point, 1-simplex is a line, 2-simplex is a triangle,

3-simplex is a tetrahedron, 4-simplex is a 5-cell, 5-simplex and so on. Figure 1 shows the simplices from dimension 1 to 5.

The simplex presents the simplest possible polytope in any given space. The standard $n$-simplex is the subset of $\mathbb{R}^{n+1}$ formed from $(n+1)$ unit vectors,

$$\Delta^n = \left\{ (x_0 + \cdots + x_n) \in \mathbb{R}^{n+1} : \sum_{i=0}^{n} x_i = 1, x_i \geq 0, \text{for all } i \right\} \tag{2.1}$$

The $(n+1)$ vertices of the standard $n$-simplex are the points $v_i \in \mathbb{R}^{n+1}$, where

$$\begin{aligned} v_0 &= (1, 0, 0, \dots, 0), \\ v_1 &= (0, 1, 0, \dots, 0), \\ &\vdots \\ v_n &= (0, 0, 0, \dots, 1). \end{aligned} \tag{2.2}$$

There is a canonical map from the standard $n$-simplex to an arbitrary $n$-simplex with vertices $(v_0, \dots, v_n)$ as,

$$(x_0 + \cdots + x_n) \mapsto \sum_{i=0}^{n} x_i \, v_i \tag{2.3}$$

The coefficients $x_i$ are the BC coordinates of a point in the $n$-simplex. The general simplex (2.3) is an affine $n$-simplex and the canonical map is an affine transformation. More generally, there is a canonical map from the standard $(n-1)$-simplex in Eq. (2.3) (with $n$-vertices) onto any polytope with $n$-vertices can be presented from a similar Eq.to (2.3) with modifying the index,

$$(x_1 + \cdots + x_n) \mapsto \sum_{i=1}^{n} x_i \, v_i \tag{2.4}$$

The $x_i$'s in Eq. (2.4) are defined with respect to a polytope instead of a simplex are called generalized BC coordinates where every polytope is expressed as the image of a simplex: $\Delta^{n-1} \twoheadrightarrow \mathbf{P}$ with the barycenter **P**. A $n$-dimensional polytope can be projected onto a $n$-dimensional Euclidean space. More detailed explanation of the BC coordinates is provided in section 2.3.

### *2.2.1 Convex polytope*

The convex polytopes are the simplest and the special case of a polytope having the additional property of also being a convex set of points in the $n$-dimensional space $\mathbb{R}^n$ (Figure 2). A convex $n$-polytope is sometimes treated as a surface or $(n-1)$-manifold. Mathematically, the quantity $\sum_{i=1}^{n} x_i\, v_i$ on the left-hand side of the Eq. (2.4) is a polytope with $n$-vertices and it will be a convex polyhedron if it is the set of solutions to a finite system of linear inequalities for any subset $S \in x_i, i = 1,2,,\ldots,n$ of $\mathbb{R}^{\mathrm{n}}$ and will be a convex polytope if it is bounded. This kind of polytope form the basis for several different generalizations of the concept of polytopes either bounded or unbounded.

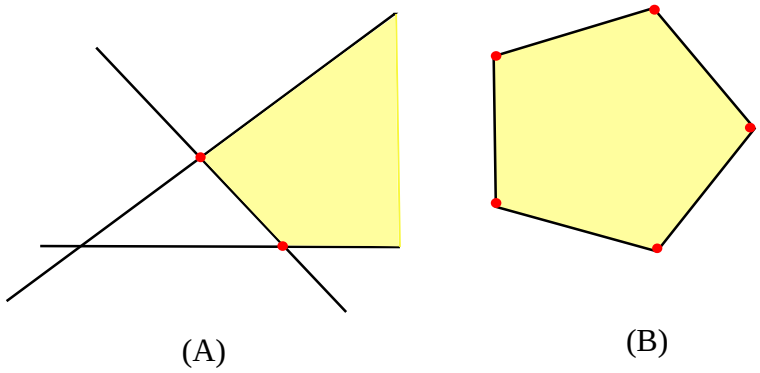



**Figure 2: The 2-dimensional convex polytopes.** (A) A $\mathcal{H}$-polytope. (B) A $\mathcal{V}$-polytope.

A convex polytope is sometimes defined as the intersection of a set of half-spaces. This definition allows a polytope to be neither bounded nor finite. For example, polytopes defined in the linear programming of linear algebra. A strong motivation for the study of

convex polytopes comes from practically significant areas such as combinatorial optimization, linear programming, and computational geometry. Based on the suitability of the problem definition convex polytopes can be an intersection of half spaces ($\mathcal{H}$-polytope, Figure 2A) or vertex representation ($\mathcal{V}$-polytope, Figure 2B).

#### *2.2.1.1 Intersection of half spaces ($\mathcal{H}$-polytope)*

A $\mathcal{H}$-polytope is defined as an intersection of a finite number of half-spaces. There exist infinitely many $\mathcal{H}$-descriptions of a convex polytope (Figure 2A). But, the minimal $\mathcal{H}$-description is unique for a full-dimensional convex polytope. A $n$-dimensional closed $\mathcal{H}$-polytope can be defined from $m$ half-spaces using a linear inequality,

$$Ax \leq b \tag{2.5}$$

where $A$ is a $m \times n$ matrix, $x$ is a $n \times 1$ column vector of variables, and $b$ is an $m \times 1$ column vector of constants. The $\mathcal{H}$-description of Eq. (2.1) is,

$$\text{Conv}(S_{\mathcal{H}}) = \{x \in \mathbb{R}^n \mid {\lambda_i}^T x \leq b_i \quad \text{for } 1 \leq i \leq m\} \tag{2.6}$$

The coefficients of each row of $A$ and $b$ correspond to the coefficients of the linear inequality defining the respective half-space. For a full-dimensional polytope, each row in the matrix corresponds with a supporting hyperplane of the polytope, a hyperplane bounding a half-space that contains in the polytope. If not full dimensional, then the solution of (2.6) lies in a proper affine subspace of $\mathbb{R}^n$ and the polytope is restricted to this subspace. The intersection of arbitrary half-spaces need not be bounded but it can be defined as a bounded convex hull.

#### *2.2.1.2 Vertex representation ($\mathcal{V}$-polytope or convex hull)*

A bounded convex polytope as the convex hull of a finite set of points, where the finite set must contain the set of extreme points of the polytope (Figure 2B). For a compact convex polytope, the minimal $\mathcal{V}$-polytope is unique and it is given by the set of the vertices of the

polytope. For a finite point set $S \in x_i, i = 1,2,, \dots, n$ the $\mathcal{V}$-polytope is the set of all convex combinations of the points. The simplex in Eq. (2.1) can be expressed by assigning a non-negative weight ($\lambda_i$) for each point ($x_i$) as,

$$\mathrm{Conv}(S_{\mathcal{V}}) = \left\{ \sum_{i=1}^{n} \lambda_i x_i \mid (\forall i : \lambda_i \geq 0) \wedge \sum_{i=1}^{n} \lambda_i = 1 \right\} \tag{2.7}$$

where $\sum_{i=1}^{n} \lambda_i = 1$. For every choice of $\lambda_i$, the resulting convex combination is a point in the convex hull, and the whole convex hull can be formed by choosing coefficients in all possible ways. A $\mathcal{V}$-polytope is a generalization of an $\mathcal{H}$-polytope.

#### *2.2.1.3 $\mathcal{H}$-polytope vs $\mathcal{V}$-polytope*

The definitions of $\mathcal{V}$-polytopes and of $\mathcal{H}$-polytopes can be considered as equivalent after some conversions and considerations. For example, both can be regarded as the presentation from some finite quantities either inequalities or points. Every $\mathcal{V}$-polytope has a description of a finite system of inequalities, and every $\mathcal{H}$-polytope can be obtained as the convex hull of a finite set of points (its vertices) [Henk et al., 2017].

An $\mathcal{H}$-polytope has a finite number of extreme points (equal to its vertices), which is the convex hull of its set of vertices. Thus, an $\mathcal{H}$-polytope is a $\mathcal{V}$-polytope. Moreover, it has a finite number of $k$-faces (for $k = 0, \dots, d-2$, where $d = \dim(\mathrm{Conv}(S_{\mathcal{H}}))$. Conversely, the convex hull of a finite set of points is an $\mathcal{H}$-polytope.

As a consequence, a $\mathcal{V}$-polytope is an $\mathcal{H}$-polytope. The switching between these two presentations is a representation and conversion problem. The $\mathcal{H}$-representation to $\mathcal{V}$-representation is a vertex enumeration problem that can be done by the method of reverse search, lexicographic reverse search and double-description etc. Alternatively, $\mathcal{V}$-representation to $\mathcal{H}$-representation is the facet enumeration problem, which can be accomplished by using polarity and doing vertex enumeration.

## 2.3 The barycentric (BC) coordinate system

The system was introduced by August Ferdinand Möbius in 1827, which provides mass points to define a coordinate-free geometry [Warren et al., 2005]. The BC coordinates over simplices are a very common tool in many computations where the location of a point of a simplex (a triangle, tetrahedron, etc.) is specified as the center of mass, or barycenter. This yield a powerful and yet simple paradigm to interpolate data values on polyhedral domains. They represent interior points of the domain as an affine combination of a set of control points, defining an interpolation scheme for any function defined on a set of control points [Zhang et al., 2014].

### *2.3.1 Coordinates on simplices*

#### *2.3.1.1 On 2-simplex (triangle)*

Let us consider a triangle $ABC$ (Figure 3) with the vertices: $A = (1,0,0)$, $B = (0,1,0)$, $C = (0,0,1)$ and anywhere else within the triangle would be a combination of these weights.

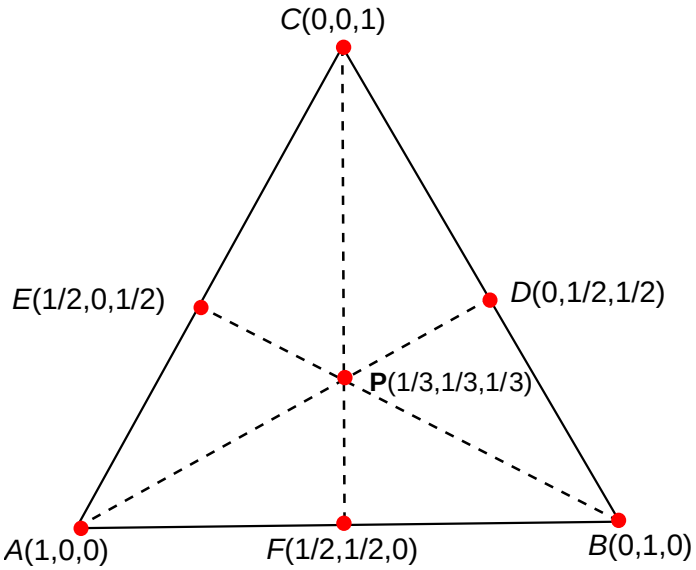



**Figure 3: The BC coordinates on the 2-simplex or on a equilateral triangle $ABC$.** The coordinates are $A = (1,0,0)$, $B = (0,1,0)$, $C = (0,0,1)$, respectively, and **P** is the barycenter having equal weights.

The barycenter of the triangle (**P**) is the point inside the triangle where there is a balance of the weights or the weights are contributing evenly to that point.

Assume that the weights at each vertex are equal to 1, then **P** is the mean of the weights (or vertices),

$$\mathbf{P} = \frac{1}{3}A + \frac{1}{3}B + \frac{1}{3}C \tag{2.8}$$

The BC coordinates for the barycenter then become $\left({}^{1}/_{3}, {}^{1}/_{3}, {}^{1}/_{3}\right)$ given in Eq. (2.8). The dashed lines in Figure 3 represent the BC axes each of which starts on a triangle edge, where the weight for the opposite vertex is 0. Then they extend through the barycenter of the triangle to the opposite vertex, where the weight for that vertex is 1 and the other coordinates in the base of each axis are ${}^{1}/_{2}$ for an equilateral triangle (Figure 3, the vertices $D, E, F$). The range of the values for each axis is (0,1) that are the weights of that vertex. So, a value less than 0 or greater than 1 will be outside of the triangle. The sum of the coordinates is equal to 1 as these coordinates represent the amounts (percentages) of each weight are being observed at that point.

#### *2.3.1.2 On n-simplex (n-polytope): generalized BC coordinates*

For $n$ vertices $v_i, i = 1,2, \dots, n$ of a simplex in an affine space $\mathbb{R}^n$, a point $\mathbf{P} \in \mathbb{R}^n$ is a barycenter of $v_i$s if,

$$\mathbf{P} = \frac{w_1 v_1 + \cdots + w_n v_n}{(v_1 + \cdots + v_n)} \tag{2.9}$$

The coefficients $(w_1, \dots, w_n)$ in Eq. (2.9) are called the BC coordinates of **P** with respect to $v_i$ if at least one of the $n$ coefficients exists. The coordinates of the vertices is given in Eq. (2.2). The points **P** lies in a convex hull if the coordinates of $v_i$'s are non-negative. Generally, the BC coordinates are not unique. But, for the restriction $\sum w_i = 1$, the coordinates are unique.

## 2.4 The algebraic geometry of contingency tables

Contingency tables are arrays of non-negative integers that arise from the cross-classification of a sample or a population of $N$ objects based on a set of categorical variables of interest [Bishop et al., 1975 & Lauritzen, 1996]. A table can be presented as a vector $n$ of non-negative integers where each integer indicates the number of times a given configuration of classifying criteria has been observed in the sample. The joint occurrence of the set of categorical variables can be presented for cell probabilities $p$ [Slavkovi´c & Fienberg, 2010]. The tables can be normalized to have entries which add to one. So, for the case of $r \times c$ two-way tables, all possible tables can be represented by the points within the $(rc-1)$-dimensional simplex in $rc$-dimensional space [Fienberg, 1968],

$$\Delta^{rc-1} = \left\{ (x_{11}, x_{12}, \dots, x_{1c}; \dots; x_{r1}, x_{r2}, \dots, x_{rc}) : x_{ij} \geq 0, \sum_{i,j} x_{ij} = 1 \right\} \qquad (2.10)$$

### *2.4.1 Geometry of contingency tables by simplex and the BC coordinates*

#### *2.4.1.1 For two-way tables*

***Case 1: For $\mathbf{2 \times 2}$ tables***

Let $X$ and $Y$ are two binary random variables that can be presented as a $2 \times 2$ table where $n_{ij}$ be the observed cell counts of the $2 \times 2$ table **n**. The row and column totals are denoted by $n_{i+}$ and $n_{+j}$, respectively, and the grand total designated $n_{++}$. The joint probability distribution of $X$ and $Y$ for a $2 \times 2$ table is represented with cell probabilities $p = (p_{ij})$, where $p_{ij} = Pr(X = i, Y = j)$, $i, j = 1, 2$, are non-negative and sum to one. The tabular illustration of these descriptions is given in Table 2.1.

**Table 2.1 Notations for a $2 \times 2$ table: Sample points (left) and parameter values (right).**

| | $Y_1$ | $Y_1$ | Total |
|---|---|---|---|
| $X_1$ | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| $X_2$ | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| Total | $n_{+1}$ | $n_{+2}$ | $n_{++}$ |

| | $Y_1$ | $Y_1$ | Total |
|---|---|---|---|
| $X_1$ | $p_{11}$ | $p_{12}$ | $p_{1+}$ |
| $X_2$ | $p_{21}$ | $p_{22}$ | $p_{2+}$ |
| Total | $p_{+1}$ | $p_{+2}$ | 1 |

The terminology of the statistical independence test for a $2 \times 2$ table can be presented in terms of the geometry of the tetrahedron ($\Delta^3$, Figure 4).

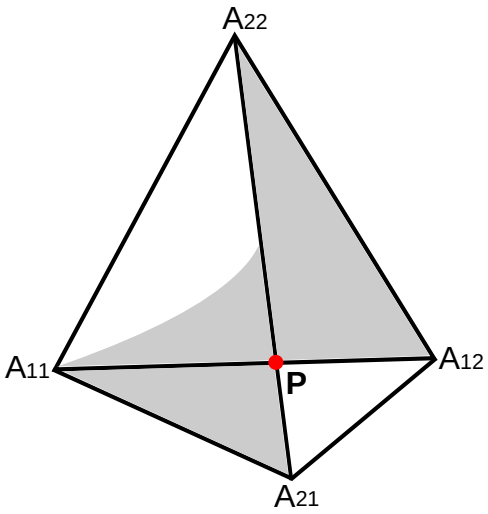



**Figure 4: The surface of independence for the $2 \times 2$ table.** The tetrahedron represents the set of all probability distributions $\mathbf{P} = (p_{11}, p_{12}, p_{21}, p_{22})$ for the $2 \times 2$ tables and the enclosed surface with gray color identifies the probability distributions satisfying the equation $p_{11}p_{22} = p_{12}p_{21}$, i.e., the toric variety for the model of independence.

The simplex in Eq. (2.10) can be represented by 4-dimensional BC coordinates as a tetrahedron (Figure 4) with vertices $A_{ij}$ for $i, j = 1,2$ where,

$$\begin{aligned} A_{11} &= (1,0,0,0) \\ A_{12} &= (0,1,0,0) \\ A_{21} &= (0,0,1,0) \\ A_{22} &= (0,0,0,1) \end{aligned} \tag{2.11}$$

The $\Delta^3$ is defined in a 4-dimensional reals space $\mathbb{R}^4$ with coordinates $\mathbf{P} = (p_{11}, p_{12}, p_{21}, p_{22})$. Geometrically, **P** is the barycenter and can be represented by the points within a 3-dimensional simplex as analogous to Eq. (2.10),

$$\mathbf{P} \in \Delta^3 = \left\{ (p_{11}, p_{12}, p_{21}, p_{22}) : p_{ij} \geq 0, \sum_{i,j} p_{ij} = 1 \right\} \tag{2.12}$$

The $\Delta^3$ is a full parameter space $(\Theta)$ when the observed counts $n = n_{ij}$ come from the multinomial distribution, $Multi(N, \mathbf{P})$. The marginal probability distributions for $X$ and $Y$ are,

$$\mathbf{P}_X = (p_{1+}, p_{2+}) = (s, 1-s) \tag{2.13}$$

$$\mathbf{P}_Y = (p_{+1}, p_{+2}) = (t, 1-t) \tag{2.14}$$

The lines $A_{11}A_{21}$ and $A_{12}A_{22}$ in the $\Delta^3$ (Figure 4) represent the set of all probability distributions,

$$\mathbf{P} = (s, 0, 1-s, 0) \tag{2.15}$$

$$\mathbf{P} = (0, s, 0, 1-s) \tag{2.16}$$

The joint distributions of the Eqs. (2.15) and (2.16) are equivalent to the marginal distribution given in Eq. (2.13).

Similarly, the lines $A_{11}A_{12}$ and $A_{21}A_{22}$ (Figure 4) represent the set of all probability distributions,

$$\mathbf{P} = (t, 1-t, 0, 0) \tag{2.17}$$

$$\mathbf{P} = (0, 0, t, 1-t) \tag{2.18}$$

Here, the marginal distribution in Eq. (2.14) is equivalent to the joint distributions of the equations (Eqs.) (2.17) and (2.18).

Let $\mathbb{R}^4_{\mathbf{c}}$ and $\mathbb{R}^4_{\mathbf{r}}$ are two four dimensional real spaces with coordinates, $\mathbf{c} = (c_{11}, c_{12}, c_{21}, c_{22})$ and $\mathbf{r} = (r_{11}, r_{12}, r_{21}, r_{22})$, respectively, and the conditional probability distributions $\mathbf{P}_{X|Y}$ and $\mathbf{P}_{Y|X}$ can be presented as $2 \times 2$ matrices of the coordinates in **c** and **r** as,

$$C = (c_{ij}) = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \tag{2.19}$$

$$R = (r_{ij}) = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} \tag{2.20}$$

For observed $Y = j$,

$$c_{ij} = Pr(X = i | Y = j) = p_{ij}/p_{+j} \tag{2.21}$$

For observed $X = i$,

$$r_{ij} = Pr(Y = j | X = i) = p_{ij}/p_{i+} \tag{2.22}$$

The conditional probabilities in the Eqs. (2.21) and (2.22) can be considered as 2-dimensional linear fractional transformations of either the cell counts or the cell probabilities. Any binary independence model (viz. log linear) for the vector **P** of cell probabilities can be fully specified by a binary design matrix $\mathbb{A}$ by imposing sets of linear constraints on the cell counts ($\sum n_{ij} = 1$) or the cell probabilities ($\sum p_{ij} = 1$) after fixing the marginal counts [Slavković & Fienberg, 2010]. For a single marginal $\mathbf{P}_X$, the geometric description of **P** can be given by an intersection of a plane, $\pi_X$ (Figure 5A) defined from,

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad t = \begin{pmatrix} s \\ 1-s \end{pmatrix} \tag{2.23}$$

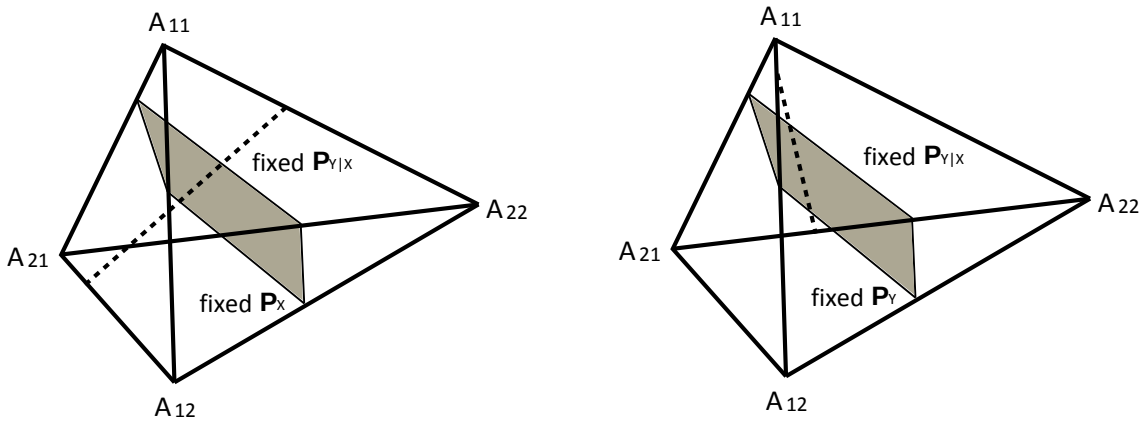



**Figure 5: The specification of the joint distribution of a $2 \times 2$ table by two sets of marginal and conditional parameters.** For a fixed marginal and the related conditional, the joint distribution of **P** is fully specified by the intersection line of the simplex of a plane. The distribution specification for the (A) fixed **r** with an intersection of a plane, $\pi_X$. (B) for fixed **c** with an intersection of a plane, $\pi_Y$.

Now, the vector **P** can be specified by a constraint matrix $\mathbb{A}$ and a vector $t$ which can be defined either of the following two ways,

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ r_{12} & -r_{11} & 0 & 0 \\ 0 & 0 & r_{22} & -r_{21} \end{pmatrix}, \quad t = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \text{ subject to } \sum p_{ij} = 1 \qquad (2.24)$$

or,

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ n_{12} & -n_{11} & 0 & 0 \\ 0 & 0 & n_{22} & -n_{21} \end{pmatrix}, \quad t = \begin{pmatrix} N \\ 0 \\ 0 \end{pmatrix} \text{ subject to } \sum n_{ij} = 1 \qquad (2.25)$$

Similarly, $\pi_Y$ can be defined for the fixed marginal $\mathbf{P}_Y$ (Figure 5B).

Hence, any contingency table with fixed margins and/or conditional probability values is a point in a convex polytope defined by a system of linear equations induced by observed marginal and conditionals.

For observed $2 \times 2$ tables, our goal is usually to make inferences about the joint distribution of the underlying categorical variables, i.e., finding estimates of and models for **P**. This can be uniquely identified by one marginal and the related conditional as,

$$Pr(X, Y) = Pr(X)Pr(Y|X) = Pr(Y)Pr(X|Y) \tag{2.26}$$

$$p_{ij} = p_{i+}r_{ij} = p_{+j}c_{ij} \tag{2.27}$$

This distribution can also be specified from the two conditionals $\mathbf{P}_{X|Y}$ and $\mathbf{P}_{Y|X}$. The geometric interpretation of a $2 \times 2$ table corresponds to fixing $\mathbf{P}_X$ and $\mathbf{P}_Y$, that is restricting the parameter space to the intersection of $\Delta^3$ with $\pi_X$ and $\pi_Y$ , respectively (Figure 5). The point **P** then lies in intersection of $\Delta^3$ with the segment $\pi_X \cap \pi_Y$ as,

$$C_{s,t} \coloneqq V(\langle p_{11} - p_{22} - (s + t - 1)\rangle) \tag{2.28}$$

***Case 2: For $r \times c$ tables***

For the general two-way tables having dimensions of $r \times c$, the simplex in Eq. (2.10) can be represented by $rc$-dimensional BC coordinates with vertices $A_{ij}$ for $i = 1,2, \dots, r$ and $j = 1,2, \dots, c$ as,

$$\begin{aligned} A_{11} &= (1,0,0,\dots,0,0,0) \\ A_{12} &= (0,1,0,\dots,0,0,0) \\ &\vdots \\ A_{r(c-1)} &= (0,0,0,\dots,0,1,0) \\ A_{rc} &= (0,0,0,\dots,0,0,1) \end{aligned} \tag{2.29}$$

The vertices of the simplex in Eq. (2.29) correspond to the respective $r \times c$ tables as,

$$\begin{matrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0, \end{matrix} \quad \begin{matrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1. \end{matrix} \tag{2.30}$$

Thus, there is a one-to-one correspondence between points in the simplex and population tables [Fienberg, 1968].

The barycenter in Eq. (2.12) is generalized for $r \times c$ tables,

$$\mathbf{P} \in \Delta^{rc-1} = \left\{ (p_{11}, p_{12}, \dots, p_{1c}; \dots; p_{r1}, \dots, p_{rc}) : p_{ij} \geq 0, \sum_{i,j} p_{ij} = 1 \right\} \tag{2.31}$$

### *2.4.1.2 The manifold of independence*

A linear space of $n$-dimensions can be considered as an $n$-flat [Sommerville, 1958], which is determined by the $(n+1)$ points and every $m$-flat $(m < n)$ is determined by the $(m+1)$ of these points lies entirely within the $n$-flat. For a $r \times c$ table, there exist $c(r-1)$-flats in the $(rc-1)$-dimensional space containing the $\Delta^{rc-1}$. There is $r$ distinct vertices of the simplex in the $c(r-1)$-flats, and all the points corresponds to the tables with their probability concentrated in one specific column. The vertices contained in the $j$-th of such $(r-1)$ are $A_{1j}, A_{2j}, \dots, A_{rj}$. Let us fix a general point in the first $(r-1)$-flat,

$$T_1 = (t_1, 0, \dots, 0;\ t_2, 0, \dots;\ t_r, 0, \dots, 0) \tag{2.32}$$

where

$$\sum_{i=1}^{r} t_i = 1 \text{ and } t_i \geq 0 \; \forall \; i \tag{2.33}$$

The analogous points in the remaining $(c-1)(r-1)$-flats are denoted by $T_2, T_3, \ldots, T_c$, respectively. The first $(c-1)$-flat containing $T_1, T_2, \ldots, T_c$ consists of all points as,

$$\mathbf{I} = (t_1 s_1, t_1 s_2, \ldots, t_1 s_c;\; t_2 s_1, t_2 s_2, \ldots, t_2 s_c; t_r s_1, t_r s_2, \ldots, t_r s_c) \tag{2.34}$$

The intersection of (2.34) with (2.10) consists of all such points that corresponds to the independent tables with $\sum_{j=1}^{c} s_j = 1$ and $s_j \geq 0 \; \forall \; j$. The family of non-intersecting $(c-1)$-flats contain all the points corresponds to the independent tables with varying $t_i$'s subject to the constraints (2.33). The manifold of independence is generated by these $(c-1)$-flats. Similarly, this manifold can be generated from the family of non-intersecting $(r-1)$-flats [Fienberg, 1968]. The tables corresponding to the points on any one of the family of nonintersecting $(c-1)$ or $(r-1)$-flat have the same row or column margins. Each of the $(c-1)$-flat meets with each $(r-1)$-flat in a single point corresponding to the independent table with the margins defined by the $(c-1)$-flat and the $(r-1)$-flat.

A manifold in $n$-space is said to have codimension $d$ iff there exists a 1-1 onto mapping taking the manifold into an $(n-d)$-flat with a non-zero $(n-d)$-dimensional volume. Thus the manifold of independence has codimension $(r-1)(c-1)$. For $2 \times 2$ tables the manifold of independence has codimension of 1, is a hyperbolic paraboloid, and the families of non-intersecting straight lines $(c-1)$-flats and the $(r-1)$-flats are simply two families of straight lines shown in Figure 4 [Fienberg, 1968].

#### *2.4.1.3 Tables with fixed margins*

Let us take a general point **P** given in Eq. (2.31) and a point **I** as shown in Eq. (2.34) on the manifold of independence where both of them are within the simplex. The vertices in the direction of the line **PI** are given as,

$$\mathbf{I} = (p_{11} - t_1 s_1,\; p_{12} - t_1 s_2, \ldots, p_{rc} - t_r s_c) \tag{2.35}$$

The line **PI** is orthogonal to the $(c-1)$-flat and the $(r-1)$-flat on the manifold of independence that pass through the point with all its coordinates equal to $1/rc$ and equivalent to $(r+c)$ linear constraints,

$$\sum_{i=1}^{r} p_{ij} = s_i \qquad \text{for } j = 1,2,\dots,c \tag{2.36}$$

$$\sum_{j=1}^{c} p_{ij} = t_i \qquad \text{for } i = 1,2,\dots,r \tag{2.37}$$

Since, $\sum_{i,j} p_{ij} = 1$, $\sum_{j=1}^{c} s_j = 1$ and $\sum_{i=1}^{r} t_i = 1$, so the number of constraints in Eqs. (2.36) and (2.37) reduce to $(r+c-2)$. Now, the locus of all the points corresponding to tables with fixed margins is the intersection of the simplex with the $(r-1)(c-1)$-flat orthogonal to the $(r-1)$-flat and the $(c-1)$-flat on the manifold of independence passing through the point with equal coordinates. This $(r-1)(c-1)$-flat intersects in exactly one point to the manifold of independence and to each manifold of constant intersection.
The $(r-1)(c-1)$ cross products of the table can be defined as,

$$\alpha_{ij} = \frac{p_{ij}p_{(i+1)(j+1)}}{p_{(i+1)j}\ p_{i(j+1)}} \qquad \text{for } i = 1,2,\dots,r-1;\ j = 1,2,\dots,c-1 \tag{2.38}$$

The method of constructing the manifold of independence is equivalent to saying that a general point **P** lies on the manifold iff all the $\alpha_{ij} = 1$ and the table expresses the non-independence iff $\alpha_{ij} \neq 1$. Hence, the manifolds of constant interaction can be constructed in the similar way of the manifold of independence.

Geometrically, an $r \times c$ table can be uniquely determined by its margin and $(r-1)(c-1)$ cross products (2.38) after setting two sets of margins and fixing a point in the correspondence $(r-1)(c-1)$-flat.

#### *2.4.1.4 For m-way tables*

Let us consider a 3-way table with dimension $r \times c \times d$, which can be represented by the points within the $(rcd-1)$-dimensional simplex with $rcd$ vertices. The locus of all points

corresponding to the independent tables is a manifold of independence generated by three families of non-intersecting flats having dimensions, $(r-1), (c-1), (d-1)$ with codimension,

$$[(r-1)(c-1)(d-1) + (r-1)(d-1) + (r-1)(c-1) + (c-1)(d-1)] \tag{2.39}$$

The tables having constant margins correspond to the points on a flat having dimensions (2.39) that is orthogonal to the generating flats on this manifold of independence passing through the point with coordinates all equal to $1/rcd$. In this case, there exists a manifold of codimension $(r-1)(c-1)(d-1)$ containing all the points corresponding to the tables with no third-order interaction and containing the manifold of independence as a sub-manifold.

***For m-dimensional multiway tables***

Let us consider $m$-categorical random variables, $X_1, \dots, X_m$, where each $X_i$ takes the values in the finite set of categories $[d_i] \equiv \{1, \dots, d_i\}$. Let $\mathcal{D} = \bigotimes_{i=1}^{m} [d_i]$, and $\mathbb{R}^{\mathcal{D}}$ be the vector space of $m$-dimensional arrays of format $d_1 \times \dots \times d_m$, with a total of $d = \prod_i d_i$ entries. The cross-classification of $n$ independent and identically distributed realizations of $(X_1, \dots, X_m)$ produce a random integer-valued array $\mathbf{n} \in \mathbb{R}^{\mathcal{D}}$, called a $m$-way contingency table, each of whose coordinate entry $\left(n_{i_1,\dots,i_m}\right)$ is the number of times the label combination, or cell is observed in the sample [Agresti, 2002; Bishop et al., 2007; Lauritzen, 1996]. A $m$-way table with dimension $\mathcal{D}$, which can be represented by the points within the $(\mathcal{D}-1)$-dimensional simplex with $\mathcal{D}$ vertices. The locus of all points corresponding to the independent tables is a manifold of independence generated by $m$ families of non-intersecting flats having dimensions, $(d_1 - 1) \dots (d_m - 1)$ with codimension,

$$(d_1 - 1) \dots (d_m - 1) + (d_1 - 1)(d_m - 1) + \dots + (d_{m-1} - 1)(d_m - 1) \tag{2.40}$$

The tables having constant margins correspond to the points on a flat having dimensions (2.40) that is orthogonal to the generating flats on this manifold of independence passing

through the point with coordinates all equal to $1/\mathcal{D}$. In this case, there exists a manifold of codimension $(d_1 - 1) \dots (d_m - 1)$ containing all the points corresponding to the tables with no $m$-order interaction and containing the manifold of independence as a sub-manifold. This can be obtained by the generalization of the arguments in the section 2.4.1.2.

## 2.5 Simplex-based presentation of tables geometry in $df$-dimensional space

The categorical variables and their combinations of a contingency table can be represented using the concept of simplices and rotations. Tables sharing the same marginal counts with an observed table can be embedded in $df$-dimensional space of the respective table as the tables are restricted by the marginal counts. The $df$-dimensional space is smaller than the dimension that the table actually has. Generally, the dimension of a table is defined by the number of the variables to be considered. Let us define the general form as a multi-way or $m$-way table where $m = 1, 2, 3, \cdots$. Here, tables are defined in the vectorized form to consider the geometry of contingency tables. Each of these tables can be represented by the simplex in an abstract way. Though the presentation is abstract but it has a geometric realization. The statistical procedures can be applied to that geometric space where the distribution can be handled as the function of the space.

### *2.5.1 Geometry of 2-way contingency tables*

Let us consider the following 2-way tables,

$$\mathrm{Y}_{2\times2} = \begin{pmatrix} \mathrm{Y}_{11} & \mathrm{Y}_{12} \\ \mathrm{Y}_{21} & \mathrm{Y}_{22} \end{pmatrix} \quad \mathrm{Y}_{2\times3} = \begin{pmatrix} \mathrm{Y}_{11} & \mathrm{Y}_{12} & \mathrm{Y}_{13} \\ \mathrm{Y}_{21} & \mathrm{Y}_{22} & \mathrm{Y}_{23} \end{pmatrix} \quad \mathrm{Y}_{2\times4} = \begin{pmatrix} \mathrm{Y}_{11} & \mathrm{Y}_{12} & \mathrm{Y}_{13} & \mathrm{Y}_{14} \\ \mathrm{Y}_{21} & \mathrm{Y}_{22} & \mathrm{Y}_{23} & \mathrm{Y}_{24} \end{pmatrix}$$

The three tables, $\mathrm{Y}_{2\times2}, \mathrm{Y}_{2\times3}$ and $\mathrm{Y}_{2\times4}$ can be realized as the 1-simplex (line segment), 2-simplex (right triangle) and 3-simplex (tetrahedron) in 1, 2 and 3-dimensional spaces as shown in Figure 1(A), (B) and (C), respectively. For the general 2-way contingency table having $r$ categories in the rows and $c$ categories in the columns,

$$\mathrm{Y}_{r\times c} = \begin{pmatrix} \mathrm{Y}_{11} & \cdots & \mathrm{Y}_{1c} \\ \vdots & \ddots & \vdots \\ \mathrm{Y}_{r1} & \cdots & \mathrm{Y}_{rc} \end{pmatrix}$$

can be realized as a $(r-1)(c-1)$-simplex in $(r-1)(c-1)$-dimensional space.

### *2.5.2 Geometry of multi-way contingency tables*

Let, Y be an $m$-dimensional multi-way contingency table with non-negative elements $\mathrm{Y}_i$, where $i \equiv (i_1,\ i_2, \dots, i_m)$, $1 \le i_j \le I_j$ and $I_j$ is the number of categories of $j$-th dimension. The index of $i$-th element for the vectorized form of the table Y is,

$$ind(i) \equiv 1 + \sum_{j=1}^{m} (i_j - 1) \prod_{k=1}^{j-1} I_k \tag{2.41}$$

Geometrically, contingency tables can be deployed in $M$-dimensional space with $M \equiv \prod_j I_j$. The tables can be located in the same space if the observation structures are the same. As the tables are restricted by the marginal counts, so a $m$-way table Y can be realized as a $\prod_j (I_j - 1)$ -simplex in $\prod_j (I_j - 1)$ -dimensional space. Here, $\prod_j (I_j - 1)$ is the $df$ of the $m$-way table Y. Hence, the contingency tables can be realized as a simplex in the corresponding $df$-dimensional space.

***Example: Realization of a $2 \times 3$ table in $df$-dimensional space***

Let us consider the following $2 \times 3$ table for the genotype counts for a single SNP with alleles "$A$" and "$a$", respectively,

**Table 2.2: A 2 × 3 table for genotype counts.**

| Phenotypes | Genotypes | | | Row totals |
|---|---|---|---|---|
| | $AA$ | $Aa$ | $aa$ | |
| Cases | 4 | 4 | 0 | 8 |
| Controls | 0 | 0 | 7 | 7 |
| Column totals | 4 | 4 | 7 | 15 |

All possible tables sharing the same marginal counts with the Table 2.2 can be constructed using the following six directional movement vectors,

(-1, +1 , 0)  (+1, -1 , 0)  (0, -1, +1)  (0, +1, -1)  (+1, 0, -1)  (-1, 0, +1)

All of these tables can be realized in the $df$-dimensional (2-dimensional) space within a triangle lattice presented in Figure 6. Each table is presented as a point on the lattice as shown by the black arrows on the right-hand side of the figure.

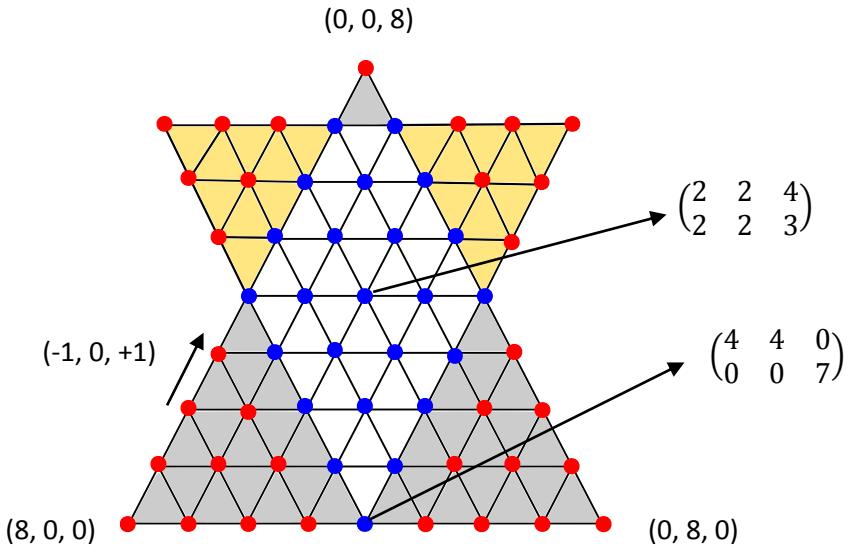



**Figure 6: A 2 × 3 table in a 2-simplex or in a triangle lattice in the $df$-dimensional space.**

The tables with non-negative integers are presented by blue dots and the red dots are for the tables with negative counts. Two consecutive triangles are observed in Figure 6, one is upward and the other is downward in the space where all the tables are realized and the downward one is smaller than the upward one as the row totals of the tables on this downward lattice are smaller.

Geometrically, a two-way table is a single observation table as it can be realized as a single observation in the $df$-dimensional space. For higher dimension, all the tables can be realized on a $df$-dimensional simplex in a $df$-dimensional space. This geometric presentation allows to combine un-ordered two or more categorical variables by combinations of respective categories that can be geometrically identified as the vectors of categorical points in $df$-dimensional space. Here, the values of each discrete variable are represented as a vertex of a simplex and the simplex is presented by the rotation matrix. The rotation lets the simplex to a standard position providing a geometrical meaning by a matrix corresponds to a simplex vertices.

### *2.5.3 Rotation matrices: coordinates of the vertices of simplices*

For a $m$-way table having $m$ variables, there is a set of $m$ rotation matrices defined from the number of the categories $I_j$ as,

$$\{R(I_1), \cdots, R(I_m)\} \tag{2.42}$$

The structure of the rotation matrix for one variable is,

$$R(n) \equiv \begin{pmatrix} r_1^T \\ r_2^T \\ \vdots \\ r_n^T \end{pmatrix} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \dots & r_{n,n} \end{pmatrix} \tag{2.43}$$

$$\text{where} \begin{cases} \forall i < n \quad r_i = \left(r_{i,j}\right) = \begin{cases} 0 & j < i \\ \sqrt{\frac{n-i}{n-i+1}} & j = i \\ -\sqrt{\frac{1}{(n-i)(n-i+1)}} & j > i \end{cases} \\ i = n \qquad r_n = \left(r_{n,j}\right) = \left(\frac{1}{\sqrt{n}}, \quad \frac{1}{\sqrt{n}}, \quad \cdots \quad , \frac{1}{\sqrt{n}}\right) \end{cases}$$

The matrix in Eq. (2.43) is a rotation matrix in a $n$-dimensional space where $n$ is the number of categories in $I_j$ for the $j$-th dimension.

The rotation matrix for the combinations of the $m$ variables is,

$$R \equiv \otimes \prod_{j=1}^{m} R\left(I_j\right) \tag{2.44}$$

Equation (2.44) defines the Kronecker product of $m$ rotation matrices of $m$-dimensional multiway contingency table that produces a matrix of dimension $M \times M$. The Kronecker product is an operation on two matrices of arbitrary size, resulting in a block matrix. It is a generalization of the outer product from vectors to matrices, and gives the matrix of the tensor product.

***Example: Rotation matrices for a* $\mathbf{2 \times 3}$ *table***

***For each variable***

For a $2 \times 3$ table, there are two variables having 2 categories in rows and 3 categories in columns that is the table has two categories in $I_j$ where $j = 2, 3$. The number of $n\ (= 2, 3)$ dimensional rotation matrices is 2 each for one variable with dimensions $2 \times 2$ and $3 \times 3$, respectively.

$$R(I_2) = \begin{pmatrix} \sqrt{\frac{n-i}{n-i+1}} & -\sqrt{\frac{1}{(n-i)(n-i+1)}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \end{pmatrix} \tag{2.45}$$

$$= \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

and,

$$R(I_3) = \begin{pmatrix} \sqrt{\frac{n-i}{n-i+1}} & -\sqrt{\frac{1}{(n-i)(n-i+1)}} & -\sqrt{\frac{1}{(n-i)(n-i+1)}} \\ 0 & \sqrt{\frac{n-i}{n-i+1}} & -\sqrt{\frac{1}{(n-i)(n-i+1)}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \end{pmatrix} \tag{2.46}$$

$$= \begin{pmatrix} \sqrt{2/3} & -1/\sqrt{2\times 3} & -1/\sqrt{2\times 3} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix}$$

***For combinations of variables***

$$R_{6\times 6} \equiv R(I_3) \otimes R(I_2)$$

$$= \begin{pmatrix} \sqrt{2/3} & -1/\sqrt{2\times 3} & -1/\sqrt{2\times 3} \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix} \otimes \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \tag{2.47}$$

$$= \begin{pmatrix} \sqrt{2/3}\,R(I_2) & -1/\sqrt{2\times 3} \times R(I_2) & -1/\sqrt{2\times 3} \times R(I_2) \\ 0 \times R(I_2) & 1/\sqrt{2} \times R(I_2) & -1/\sqrt{2} \times R(I_2) \\ 1/\sqrt{3} \times R(I_2) & 1/\sqrt{3} \times R(I_2) & 1/\sqrt{3} \times R(I_2) \end{pmatrix}$$

The numerical calculation yields,

$$R_{6\times 6} = \begin{pmatrix} 0.58 & -0.58 & -0.29 & 0.29 & -0.29 & 0.29 \\ 0.58 & 0.58 & -0.29 & -0.29 & -0.29 & -0.29 \\ 0 & 0 & 0.50 & -0.50 & -0.50 & 0.50 \\ 0 & 0 & 0.50 & -0.50 & -0.50 & -0.50 \\ 0.41 & -0.41 & 0.41 & 0.41 & 0.41 & -0.41 \\ 0.41 & 0.41 & 0.41 & 0.41 & 0.41 & 0.41 \end{pmatrix} \quad (2.48)$$

The rotation matrix in Eq. (2.48) is the universal form for any $2 \times 3$ table or any vector of length 6 regardless the cell counts of the table. In the R-program, Eq. (2.48) can be calculated by using the function, kronecker$\big(R(I_3), R(I_2)\big)$ or writing, $R(I_3)$%x%$R(I_2)$.

The dimension reduction of $R$ can be done by a suitable transformation to remove the zero and identical elements that produce for the tables sharing the same marginal counts. One of the transformation method called spherization is briefly discussed in the Chapter 6.

## 2.6 Concluding remarks

In this chapter, the algebraic and geometric outlines of contingency tables are introduced. The geometric relation of the presentation of tables are based on the concepts of simplices and polytopes along with the barycentric coordinates. Based on the dimension of the tables, the order of the simplices are subjected to change. So, the simplicial definition of a contingency table with the vertices defines a point called a barycenter where the weights or the contribution of all the vertices are the same. Then, the joint distribution of this barycenter defines the distribution of the corresponding table that can be completely determined from fixed conditional probabilities and marginal counts. The manifold of independence for the corresponding test is defined for the tables with fixed marginal counts as the intersection line of the simplex of a plane. On the other hand, this geometric demonstration can also be done in the table's $df$-dimensional space with the additional

concept of the rotation to translate the contents of the simplices into a rotation matrix. This rotational information facilitates significant dimension reduction for the space defined by the combination of variables without the loss of information.

# Chapter 3

# The distribution of MAX-test statistic

## 3.1 Introduction

The purpose of this chapter is to briefly describe the MAX-test and its asymptotic distribution according to González et al. [2008]. Also, the demonstration of this test for a particular genetic scenario using multiple additive models is given here. The MAX-test for multiple $df = 1$ tests of additive models along with underlying distribution is also presented here. The GWAS use single-marker analysis where tests for one SNP is performed at a time. For this test, the trend test derived from an additive genetic model is often used. A robust test, MAX, based on the maximum of three trend test statistics derived for recessive, additive, and dominant models, has been proposed as an alternative for GWAS [Li et al., 2008]. The association is assessed by using the largest test statistic from dominant, recessive and additive models [Milne et al., 2006; Pooley et al., 2006; Cargill et al., 2007]. This statistic is known as the MAX-statistic [Freidlin et al., 2002]. The previous study results showed the advantages of using the MAX-statistic over the Cochran-Armitage trend test (CATT), derived from the additive model, to prioritize SNPs or to detect disease-associated SNPs [Zheng et al. 2009]. González et al. [2008] showed that the use of the MAX-statistic is a powerful approach that provides safeguard against model uncertainty. The distribution of the MAX-statistic should be used to calculate $p$-values to solve the computational burden of the permutation testing approach where an extensive number of permutations is required to get the accurate estimate of the distribution of the statistic for those SNPs that are strongly associated with the disease and the $p$-values are in the tail of the distribution [González et al., 2008].

The organization of the present chapter is as follows. The MAX-test is discussed in section 3.2. The distribution of the MAX-test statistic and a generalized form of MAX-test for multiple additive models are provided in sections 3.3 and 3.4, respectively. Finally, section 3.5 contains some concluding remarks.

## 3.2 Maximum test (MAX-test)

The MAX-test for additive model is based on the CATT (Cochran, 1954; Armitage, 1955). Assume a case control study (Table 3.1) with a diallelic locus whose major and minor alleles are "*a"* and "*A"*, respectively.

**Table 3.1: Genotype distribution among cases and controls for single SNP.**

| Phenotypes | Genotypes | | | Total |
|---|---|---|---|---|
| | $aa$ | $aA$ | $AA$ | |
| Cases | $r_0$ | $r_1$ | $r_2$ | $R$ |
| Controls | $s_0$ | $s_1$ | $s_2$ | $S$ |
| Total | $n_0$ | $n_1$ | $n_2$ | $N$ |

Let, $p = (p_0,\ p_1,\ p_2)$ and $q = (q_0,\ q_1,\ q_2)$ are the sample proportions and $r = (r_0,\ r_1,\ r_2)$, $s = (s_0,\ s_1,\ s_2)$ are the numbers of cases and controls, respectively, in genotype categories $G_i$, where $i \in \{0, 1, 2\}$ reflects the number of *"A"* alleles a subject has. Thus, $G_0$, $G_1$ and $G_2$ correspond to genotypes "*aa"*, "*Aa"* and "*AA"*, respectively. For the weight vector $x_i$, $i = 0, 1, 2$ having constraints $x_0 \le x_1 \le x_2$ and $x_0 < x_2$, the CATT statistic is,

$$T = \frac{\left[\frac{1}{N}\sum_{i=0}^{2} x_i(Sr_i - Rs_i)\right]^2}{\frac{RS}{N^3}\left[N\sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2\right]} \tag{3.1}$$

where, $R$, $S$ and $n_i$ are the marginal sums; $N = R + S = \sum_{i=0}^{2} n_i$ is the total sample size. The distribution of both $r$ and $s$ is trinomial with parameters $(R, p_i)$ and $(S, q_i)$ where,

$p_i$ and $q_i$ are the sample proportions for cases and controls, respectively. Now, the test statistic for additive model can be written from equation Eq. (3.1) in terms of $p_i$ and $q_i$ and weight vector $x_i = (-1, 0, 1)$,

$$\begin{aligned}\chi_A^2 &= T(p,q) \\ &= \frac{(p_0 - p_2 - q_0 - q_2)^2\, RS(R+S)}{-[(p_0 - p_2)R + (q_0 - q_2)S]^2 + (R+S)[(p_0 - p_2)R + (q_0 - q_2)S]}\end{aligned} \tag{3.2}$$

Eq. (3.2) follows $\chi^2$-distribution with $df = 1$, where, $p = (p_0, p_2)$ and $q = (q_0, q_2)$.

The $\chi^2$-statistic for the dominant and recessive model can be derived from the likelihood ratio test for a given SNP proposed by Cordell and Clayton [2002]. The test statistic for a general contingency table is,

$$G = 2\sum_i O_i \, log\left({}^{O_i}/_{E_i}\right) \tag{3.3}$$

where the sum is computed over all cells, $i$, and $O$ denotes the observed number of cases and $E$ the expected number of cases under the null hypothesis $(H_0)$ of no association. For a general contingency table $2 \times k$ using the similar notations of Table 3.1, the Eq. (3.3) can be expressed as,

$$\begin{aligned}G(p,q) &= 2\left\{\sum_{i=0}^{k} r_i \log\left(\frac{r_i N}{n_i R}\right) + s_i \log\left(\frac{s_i N}{n_i S}\right)\right\} \\ &= 2\left\{\sum_{i=0}^{k} p_i R \log\left(\frac{2p_i}{p_i + q_i}\right) + q_i S \log\left(\frac{2q_i}{p_i + q_i}\right)\right\}\end{aligned} \tag{3.4}$$

where $k$ denotes the number of column categories of the independent variable, $p = (p_0,\ p_1,\ p_2, \dots, p_k)$ and $q = (q_0,\ q_1,\ q_2, \dots, q_k)$ are the sample proportions for cases and controls, respectively. In particular, for the case of a codominant model given in Table 3.1, $p = (p_0,\ p_1,\ p_2)$ and $q = (q_0,\ q_1,\ q_2)$. In this case, the $G$ statistic follows a $\chi^2$ -

distribution with 2 $df$. The $\chi^2$-statistic for the dominant ($\chi_D^2$) and the recessive ($\chi_R^2$) models can be expressed using the $G$ statistic as,

$$\chi_D^2 = G(p,q) = 2\left(p_1 R \log\left[\frac{2p_1}{p_1+q_1}\right] + q_1 S \log\left[\frac{2q_1}{p_1+q_1}\right] + (p_2+p_3) R \log\left[\frac{2(p_2+p_3)}{p_2+p_3+q_2+q_3}\right] + (q_2+q_3) R \log\left[\frac{2(q_2+q_3)}{p_2+p_3+q_2+q_3}\right]\right) \quad (3.5)$$

$$\chi_R^2 = G(p,q) = 2\left((p_1+p_2) R \log\left[\frac{2(p_1+p_2)}{p_1+p_2+q_1+q_2}\right] + (q_1+q_2) S \log\left[\frac{2(q_1+q_2)}{p_1+p_2+q_1+q_2}\right] + p_3 R \log\left[\frac{2p_3}{p_3+q_3}\right] + q_3 S \log\left[\frac{2q_3}{p_3+q_3}\right]\right) \quad (3.6)$$

For the three models, the MAX-statistic is,

$$\chi_{max}^2 = \max\{\chi_A^2, \chi_D^2, \chi_R^2\} \quad (3.7)$$

## 3.3 The asymptotic distribution of MAX-test statistic

Let us consider the Eq. (3.7) as,

$$T_{max} = (-1)^s \max\left\{\sqrt{\chi_A^2}, \sqrt{\chi_D^2}, \sqrt{\chi_R^2}\right\} \quad (3.8)$$

where the positive square roots are taken for the Eq. (3.8), and $s = 1$ when $(r_2/r_0) < (s_2/s_0)$ (i.e., the sample odds ratio for "$AA$" against "$aa$" is less than one), 0 otherwise. As the square root is a monotonic function, the maximum $\chi^2$-test will concur with the

maximum square-root test. Define $T_A = (-1)^s \max\left\{\sqrt{\chi_A^2}\right\}$, $T_D = (-1)^s \max\left\{\sqrt{\chi_D^2}\right\}$ and $T_R = (-1)^s \max\left\{\sqrt{\chi_R^2}\right\}$. Under the independence it is noticeable that,

$$\begin{aligned} & Pr\{|T_{max}| \leq \chi_{max}\} \\ & = Pr\{|T_A| \leq \chi_{max} \cap |T_D| \leq \chi_{max} \cap |T_R| \leq \chi_{max}\} \\ & = Pr\left\{\sqrt{\chi_1^2} \leq \chi_{max}\right\}^3 \end{aligned} \tag{3.9}$$

The joint distribution of the three test statistics can be derived using the delta method [Agresti, 2002] to compute the asymptotic $p$-values. In a typical case-control setting, we can consider that the data have been generated from two independent multinomial distributions. One for cases with sample proportions $p = (p_0,\ p_1,\ p_2)$ and other for controls with sample proportions $q = (q_0,\ q_1,\ q_2)$. For the case of a codominant model given in Table 3.1, assume $r$ follow a trinomial distribution with parameters $R$ and $p$, and $s$ follow a trinomial distribution with parameters $S$ and $q$. Let $\Xi_p = (\sigma_{p_{ij}})$ is a covariance matrix of $p$ for $\sigma_{p_{ij}} = p_j(1 - p_j)$ , $\sigma_{p_{jk}} = p_j p_k$ for $j \neq k$ and $\Xi_q = (\sigma_{q_{ij}})$ is a covariance matrix of $q$ for $\sigma_{q_{ij}} = q_j(1 - q_j)$ , $\sigma_{q_{jk}} = q_j q_k$ for $j \neq k$. The $\Xi_p$ and $\Xi_q$ matrices have the following forms,

$$\Xi_p = [\mathrm{diag}(p) - pp'] \tag{3.10}$$

$$\Xi_q = [\mathrm{diag}(q) - qq'] \tag{3.11}$$

, respectively. Therefore, the full covariance matrix, $\Xi$, of the observed proportions in the contingency table will be,

$$\Xi = \begin{pmatrix} \Xi_p & 0 \\ 0 & \Xi_q \end{pmatrix} \tag{3.12}$$

If $\pi$ is the vector of sample proportions, $\pi = (p, q) = (p_0,\ p_1,\ p_2,\ q_0,\ q_1,\ q_2)$, the joint distribution of $\pi$ is asymptotically multivariate normal [Agresti, 2002]. The delta method generalizes further to a vector of functions of an asymptotically normal random vector. Therefore, as $T_{max}$ is a vector of functions of $\pi$, which is asymptotically normal, we obtain that $T_{max}$ follows a trivariate normal distribution with asymptotic variance given by,

$$\phi(H_i)' \, \Xi \, \phi(H_i) \tag{3.13}$$

and the covariances for each pair as,

$$\phi(H_i)' \, \Xi \, \phi(H_j) \tag{3.14}$$

where $\phi(H_i) = \frac{\delta H_i}{\delta \pi_c}$, $c = 1, 2, \ldots\ , 2k$. $\pi_c$ indicates the different components of the vector $\pi$, $H_1 = T_A, H_2 = T_D\ , H_3 = T_R$ and $k$ is the number of columns considered in the $2 \times k$ table (for Table 3.1, $k = 3$). Each of the test statistics $T_A$, $T_D$, $T_R$ has mean zero under no association, the probability in Eq. (3.9) can be estimated from the trivariate normal distribution, $N_3$, as follows,

$$\int_{-m}^{m} \int_{-m}^{m} \int_{-m}^{m} N_3(z; 0, \Sigma) \tag{3.15}$$

where $m = \max\left\{\sqrt{\chi_A^2}, \sqrt{\chi_D^2}\ , \sqrt{\chi_R^2}\right\}$ and $z = (z_1, z_2, z_3)$ are the multivariate normal variables with covariance matrix,

$$\Sigma = \begin{pmatrix} 1 & \rho_{T_A T_D} & \rho_{T_A T_R} \\ \rho_{T_D T_A} & 1 & \rho_{T_D T_R} \\ \rho_{T_R T_A} & \rho_{T_R T_D} & 1 \end{pmatrix} \tag{3.16}$$

The correlations in Eq. (3.16) can be computed using Eq. (3.13) and Eq. (3.14) as follows,

$$\rho_{T_A T_D} = \frac{\phi(T_A)'\Xi\phi(T_D)}{[\phi(T_A)'\Xi\phi(T_A)]^{1/2}[\phi(T_D)'\Xi\phi(T_D)]^{1/2}} \tag{3.17}$$

Similarly, the quantities $\rho_{T_A T_R}$ and $\rho_{T_D T_R}$ can be computed.

## 3.4 The MAX-test for $n$-additive models

For a SNP subset having $n$ SNPs, the number of additive tests equal to $n$ ($n_{test}$). The contingency table of combinatorial genotypes has dimension of $2 \times 3^n$ (Table 3.2), where $k = 3^n$ is the number of columns considered in the table, and the genotype categories $G_i$, where $i \in \{0, 1, 2, \dots, k\}$.

**Table 3.2: Genotype distribution among cases and controls for $n$ SNPs.**

| Phenotypes | Genotypes | | | | Total |
|---|---|---|---|---|---|
| | $G_0$ | $G_1$ | $\cdots$ | $G_k$ | |
| Cases | $r_0$ | $r_1$ | $\cdots$ | $r_k$ | $R$ |
| Controls | $s_0$ | $s_1$ | $\cdots$ | $s_k$ | $S$ |
| Total | $n_0$ | $n_1$ | $\cdots$ | $n_k$ | $N$ |

For the weight vector $x_i$, $i = 0, 1, 2, \dots, k$ having constraints $x_0 \leq x_1 \leq \cdots \leq x_k$ and $x_0 < x_k$. For $n_{test}$, the MAX-statistic can be expressed as,

$$\chi^2_{\max} = \max\{\chi^2_{A_i}\}, i = 1,2, \dots, n_{test} \tag{3.18}$$

where each $\chi^2_{A_i}$ is the function of $p = (p_0,\ p_1,\ p_2, \dots, p_k)$ and $q = (q_0,\ q_1,\ q_2, \dots, q_k)$ derived from Eq. (3.1) assuming a multinomial distribution of $r = (r_0,\ r_1,\ r_2, \dots, r_k)$ and $s = (s_0,\ s_1,\ s_2, \dots, s_k)$, respectively.

To derive the distribution of $\chi^2_{max}$, let us consider the Eq. (3.18) as,

$$T_{max} = (-1)^s \max\left\{\sqrt{\chi^2_{A_i}}\right\}, i = 1,2,\dots,n_{test} \tag{3.19}$$

where $s = 1$ when $(r_k/r_0) < (s_k/s_0)$, 0 otherwise. Due to the monotonicity of Eq. (3.19), the maximum $\chi^2$-test will concur with the maximum square-root test. From the assumption of independence,

$$\begin{aligned} &Pr\{|T_{max}| \le \chi_{max}\} \\ &= Pr\left\{|T_{A_1}| \le \chi_{max} \cap |T_{A_2}| \le \chi_{max} \cap \dots \cap \left|T_{A_{n_{test}}}\right| \le \chi_{max}\right\} \\ &= Pr\left\{\sqrt{\chi^2_1} \le \chi_{max}\right\}^{n_{test}} \end{aligned} \tag{3.20}$$

where, $T_{A_i} = (-1)^s \max\left\{\sqrt{\chi^2_{A_i}}\right\}\ \forall\ i$.

To derive the joint distribution of the $n_{test}$ test statistics, let us consider the data from two independent, multinomial distributions for cases and controls with sample proportions $p$ and $q$, respectively. In the case of a codominant model given in Table 3.2, assume $r$ follow a multinomial distribution with parameters $R$ and $p$, and $s$ follow a multinomial distribution with parameters $S$ and $q$. The covariance matrix, $\Xi$, of the observed proportions in the contingency table has the same form given in Eq. (3.12) where the covariance matrices $\Xi_p$ and $\Xi_q$ are generalized for the $k$ sample proportions for each $p$ and $q$ keeping structures are the same as in Eqs. (3.10) and (3.11), respectively.

Now, $T_{max}$ is a vector of functions of $\pi$ where $\pi = (p_0,\ p_1, \dots, p_k,\ q_0,\ q_1, \dots, q_k)$, having an asymptotically multivariate normal distribution [Agresti, 2002]. Hence, $T_{max}$ follows a $k$-variate normal distribution with asymptotic variance and the covariances for each pair $\phi(H_i)'\ \Xi\ \phi(H_i)$ and $\phi(H_i)'\ \Xi\ \phi(H_j)$, respectively, where $\phi(H_i) = \frac{\delta H_i}{\delta \pi_c}$, $c = 1, 2, \dots\ , 2k$. $\pi_c$ indicates the different components of the vector $\pi$, $H_1 = T_{A_1}, H_2 = T_{A_2}, \dots\ , H_n = T_{A_{n_{test}}}$. Under the assumption of no association, the probability in Eq. (3.20) can be estimated from the $k$-variate normal distribution, $N_k$,

$$\int_{-m}^{m}\int_{-m}^{m}\dots\int_{-m}^{m} N_k(z;0,\Sigma) \tag{3.21}$$

where $m = \max\left\{\sqrt{\chi^2_{A_i}}\right\}$ and $z = (z_1, z_2, \dots, z_k)$ are the multivariate normal variables with covariance matrix,

$$\Sigma = \begin{pmatrix} 1 & \rho_{T_{A_1}T_{A_2}} & \cdots & \rho_{T_{A_1}T_{A_k}} \\ \rho_{T_{A_2}T_{A_1}} & 1 & \cdots & \rho_{T_{A_2}T_{A_k}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{T_{A_k}T_{A_1}} & \rho_{T_{A_k}T_{A_2}} & \cdots & 1 \end{pmatrix} \tag{3.22}$$

The correlations in Eq. (3.22) are computed using the similar ways as in Eq. (3.17) for all possible pairs of $T_{A_i}$s.

The probabilities defined in Eqs. (3.15) or (3.21) can be computed using numerical integration as described in Genz [1992]. The Delta method may fail near singularity points that may happen with sparse data or with rare mutations. In this case, the theorem given by Puig [1998], which is an entropy based method considering the geometry of tests can be used to avoid problems with singularity. For GWAS multiple testing, both the Eqs. (3.15) and (3.21) will estimate the cumulative probability that is considered as the $\alpha$-value of the test. The TND approach using the minimax tilting by Botev [2017] can be used to estimate this probability that use the simulation from the truncated multivariate normal distribution in high dimensions by solving the recurrent problem in statistical computing using approximate Markov chain Monte Carlo sampling. The minimax tilting method is used for exact independently and identically distributed data simulation from the TND, which provides both a method for simulation and an efficient estimator to hitherto intractable Gaussian integrals that has a rare vanishing relative error (VRE) asymptotic property.

## 3.5 Concluding remarks

The MAX-test is useful in the case-control genetic association studies. All the computational burden for the conventional permutation procedure could be reduced by one

evaluation of the three or multiple models if the distribution of the MAX-statistic is known. Also, the geometric definition of the underlying distribution of test's contour lines provides a clear and useful insight about the shape of the defined distribution where the shape of the distribution of the tests can be defined by the truncated hyperplanes in a given space. Thus the estimated $p$-values are equivalent to the probabilities defined by the corresponding TND, which is an integral problem over the space where the test is defined. The exponential tilting with IS and using the optimization concept of the minimax tilting approach can be applied with a greater precision as compared to the simple Monte Carlo estimation.

# Chapter 4

# Some modified MAX-tests

## 4.1 Introduction

This chapter surveys the literature on some existing methodologies related to the development and the modification of the MAX-tests. To estimate $p$-values in GWAS, the trend $\chi^2$ $(df = 1)$ test for additive mode and CATT that fit to the additive model are most powerful to detect the markers in additive mode. But, these tests may not be robust when the additive assumption is not appropriate for the true underlying disease model. A robust test, MAX, based on the maximum of three trend test statistics derived for recessive, additive, and dominant models, has been proposed as an alternative for GWAS. The MAX-test for the additive model is commonly used for the $2 \times 3$ table of individual SNP test [Yamada & Okada, 2008] because of providing safeguard against model uncertainty [González et al., 2008] and having reasonable power to detect both additive and dominant effects [Bush et al., 2012]. Under $H_0$ of no association, the MAX-statistic does not follow the standard normal distribution asymptotically. Thus a computationally intensive resampling-based procedure is required to estimate it's $p$ -value [Li et al., 2008; Knijnenburg et al., 2009], which is computationally challenging for the analysis of GWAS. Because it is not straightforward to obtain $p$ -value for the combination of these $df = 1$ tests, Gonzalez et al. [2008] proposed to estimate the $p$-values for the maximum value among $\chi^2$ $(df = 1)$ for recessive and dominant modes and CATT for additive mode by formulating asymptotic distribution of the MAX-statistic with Delta method. There are so many literatures on the modification and development of the MAX-test provided by the researchers on over the decade from different viewpoints, some of them are discussed in this review chapter.

The organization of the present chapter is as follows. Section 4.2 discusses about the two robust procedures that are commonly used in multiple testing to set out the type of test based on the mode of inheritance of the underlying model. In section 4.3, some existing modifications of the MAX-test are discussed in details. Finally, section 4.4 contains some concluding remarks.

## 4.2 Robust procedures: Maximin efficiency robust tests (MERT) and MAX-test

Let $\{T_i\colon\ i \in I\}$ are the optimal test statistics obtained from a collection of alternative models $\{M_i,\ i \in I\}$, where $I$ can be either a finite set or an interval. Under the $H_0$, assume each of the $T_i \xrightarrow{asymptotic} N\big(\mathrm{E}(T_i), \mathrm{var}(T_i)\big)$ that is $Z_i = [T_i - \mathrm{E}(T_i)]/\{\mathrm{var}(T_i)\}^{1/2}$, converges to $N(0,1)$ where $\mathrm{E}(T_i)$ and $\mathrm{var}(T_i)$ are the mean and the variance of $T_i$ under $H_0$. Suppose the joint distribution of $Z_i$ and $Z_j$ are also normal for any $i, j \in I$ with correlation $\rho_{ij}$. The optimal test $Z_i$ would be used if $M_i$ is the true model. But, when the true $M_i$ is unknown and the $Z_j$ is used, assume the Pitman asymptotic relative efficiency of $Z_j$ relative to $Z_i$ is $e\big(Z_j, Z_i\big) = \rho_{ij}^2$ for $i, j \in I$ [Zheng et al., 2006].

When each model in the family is scientifically reasonable even if the true model is unknown, the minimum asymptotic relative efficiency compared to the optimum test for each model, $Z_i$ and $Z_j$ is, $\inf_{i\in I} e\big(Z_j, Z_i\big)$ for $j \in I$. One robust test is to choose the optimal test $Z_l$ from the family $\{Z_i\colon\ i \in I\}$ which maximizes the minimum asymptotic relative efficiency, that is,

$$\inf_{i\in I} e(Z_l, Z_i) = \sup_{j\in I} \inf_{i\in I} e\big(Z_j, Z_i\big) \tag{4.1}$$

From the definition given in Eq. (4.1), $Z_l$ is the most robust test in $\{Z_i\colon\ i \in I\}$ whose distribution converges in distribution to a standard normal random variable under the $H_0$.

For a family of models, all the corresponding optimal tests are obtained and all the consistent tests with an asymptotically normal distribution can be used. Let $C$ be the all these tests for the family of models. The original family of test statistics can be expanded to $C$. The purpose is to find a test $Z$ from $C$, rather than from the original family $\{Z_i\colon\ i \in I\}$, such that,

$$\inf_{i\in I} e(Z, Z_i) = \sup_{Z\in C} \inf_{i\in I} e(Z, Z_i) \tag{4.2}$$

The test $Z$ satisfying Eq. (4.2) is called the MERT [Gastwirth, 1966]. When the family $C$ is restricted to the convex linear combinations of $\{Z_i\colon\ i \in I\}$, the resulting robust test is denoted as $Z_{\mathrm{MERT}}$. Since $\{Z_i\colon\ i \in I\} \subset C$,

$$\sup_{Z\in C} \inf_{i\in I} e(Z, Z_i) \geq \sup_{j\in I} \inf_{i\in I} e\big(Z_j, Z_i\big) \tag{4.3}$$

Gastwirth [1966] proved that $Z_{\mathrm{MERT}}$ uniquely exists and can be written as a closed convex combination of optimal tests $Z_i$ in the family $\{Z_i\colon\ i \in I\}$.

A MAX-statistic is the natural non-linear robust statistic that is the maximum over the extreme pair $(Z_s, Z_t)$ or the triple $(Z_s, Z_u, Z_t)$ for the entire family [Freidlin et al., 2002],

$$Z_{\mathrm{MAX2}} = \max(Z_s, Z_t) \text{ or } Z_{\mathrm{MAX3}} = \max(Z_s, Z_u, Z_t) \tag{4.4}$$

In general, the maximum over the entire family is,

$$Z_{\mathrm{MAX}} = \max_{i\in I} Z_i \text{ or } Z_{\mathrm{MAX}} = \max_{i\in C} Z_i \tag{4.5}$$

When the range of models is wide, $Z_{\mathrm{MAX}}$ should have greater efficiency robustness than $Z_{\mathrm{MERT}}$ although $Z_{\mathrm{MERT}}$ is easier to compute and to use than $Z_{\mathrm{MAX}}$. The selection of the robust test depends on the minimum correlation $\rho_{st}$ of the entire family of optimal tests. Results from Freidlin et al. [2002] showed that when $\rho_{st} \geq 0.75$, MERT and MAX2 (or MAX3) have similar power and hence the simpler MERT can be used.

## 4.3 Existing modifications on MAX-tests

As there is no single best test for all situations (Kuo & Feingold, 2010), researchers have developed the MAX-tests by considering different aspects (Freidlin et al., 2002; González et al., 2008; Li et al., 2008; Yamada & Okada, 2008; Hothorn LA & Hothorn T, 2009; So & Sham, 2011; Zang & Fung, 2011; Loley et al., 2013; Buzdugan et al., 2016), which are considered more robust than the CATT and more powerful than Pearson's $\chi^2$ test, according to the literature and findings of the aforesaid research papers. The detailed methodological modifications of González et al. [2008] are described in the Chapter 3. Following sections are presenting the modified theories of MAX-tests by some other researchers.

### *4.3.1 A Comparison of MAX-test and the MERT [Freidlin et al., 2002]*

To provide the formulas for calculating the required sample size in order to achieve a prespecified power that account for the need to estimate the variance of the test statistic was the aim of this study. The theoretical variance $(\sigma_0^2)$ is usually unknown and is replaced by a consistent estimate under the $H_0$. When the $H_a$ is true, the estimated standard deviation converges to $\sigma^*$, which may differ from $\sigma_0$. Therefore, power and sample size calculations should be based on $\sigma^*$. Here, the given formulas was used to approximate the sample size needed to achieve a desired power. The demonstration indicated that the use of the recessive scores for the recessive model instead of the test optimal for additive model typically leads to noticeably smaller required sample sizes. When the allele frequency was high there was a substantial difference between the optimal tests for dominant and additive models if the underlying model of inheritance was dominant.

Let us consider the data from a case-control study that are represented in Table 3.1 with the similar description as in the section 3.2 (Chapter 3), where "$A$" is a high risk candidate allele and "$a$" is any of the other alleles. The $H_0$ of no association can be written as,

$$H_0: p_i = q_i \quad \text{for } i = 0, 1, 2 \tag{4.6}$$

Let $f_0, f_1$ and $f_2$ are the penetrances of the genotypes "$aa$", "$aA$" and "$AA$", respectively and $K$ denote the disease prevalence and $\gamma_i = f_i/f_0$ and $\delta_i = (1-f_i)/(1-f_0)$ for $i = 1, 2$ be genotype relative risks. The population genotype probabilities will be denoted by $g_0 = P(aa), g_1 = P(aA), g_2 = P(AA)$ and $K = \sum_i f_i g_i$. Now, the $p_i$ and $q_i$ can be expressed as,

$$p_i = \frac{f_i g_i}{K} = \frac{\gamma_i g_i}{\sum_i \gamma_i g_i} \text{ and } q_i = \frac{(1-f_i)g_i}{1-K} = \frac{\delta_i g_i}{\sum_i \delta_i g_i} \tag{4.7}$$

where $\gamma_0 = \delta_0 = 1$.

Hence, the $H_0$ in Eq. (4.6) is equivalent to,

$$H_0: \gamma_1 = \gamma_2 = 1 \quad \text{i.e. } f_0 = f_1 = f_2 \tag{4.8}$$

and the $H_a$,

$$\begin{gathered} H_a: \gamma_2 > \gamma_1 \geq 1 \\ \text{or} \\ H_a: \gamma_2 \geq \gamma_1 > 1 \end{gathered} \tag{4.9}$$

Define $x = (x_0, x_1, x_2)$ be the a set of scores for the genotypes "$aa$", "$aA$" and "$AA$". Also, define, $x = (0, 1, 2)$, $x = (0, 1, 1)$ and $x = (0, 0, 1)$ for the additive $(\gamma_2 = 2\gamma_1 - 1)$, dominant $(\gamma_1 = \gamma_2)$ and recessive $(\gamma_1 = 1)$ models, respectively. For given score $x$ the CATT statistics can be written as,

$$Z_T^* = \frac{U}{\left[\text{var}_{H_0}(U)\right]^{1/2}} \tag{4.10}$$

where $U = \frac{1}{N}\sum_{i=0}^{2} x_i(Sr_i - Rs_i)$ and

$$\mathrm{var}_{H_0}(U) = N\sigma_0^2 = \frac{RS}{N}\left[\sum_{i=0}^{2} x_i^2 q_i - \left(\sum_{i=0}^{2} x_i q_i\right)^2\right] \tag{4.11}$$

The $q_i$ in Eq. (4.11) may not be known and $\sigma_0$ is estimated from the data by representing the $q_i$ by $n_i/N$. This yields,

$$\widehat{\mathrm{var}}_{H_0}(U) = N\hat{\sigma}_0^2 = \frac{RS}{N^3}\left[N\sum_{i=0}^{2} x_i^2 n_i - \left(\sum_{i=0}^{2} x_i q_i\right)^2\right] \tag{4.12}$$

Let us rewrite the Eq. (4.10) as,

$$Z_T = \frac{U}{\left[\widehat{\mathrm{Var}}_{H_0}(U)\right]^{1/2}} \tag{4.13}$$

Reject $H_0$ if $|Z_T| > z_{1-\alpha/2}$, where $z_q$ is the $q$-th percentile of a standard normal distribution and $\alpha$ is the level of significance.

***Relation between power and sample size***

Define $\mathrm{E}_{H_a}(U) = N\mu_a$ and $\mathrm{var}_{H_a}(U) = N\sigma_a^2$. According to Slager and Schaid [2001],

$$\mu_a = \frac{RS}{N^2}\sum_{i=0}^{2} x_i(p_i - q_i) \tag{4.14}$$

$$\begin{aligned}\sigma_a^2 = \frac{RS^2}{N^3}&\left[\sum_{i=0}^{2} x_i^2 p_i - \left(\sum_{i=0}^{2} x_i p_i\right)^2\right] \\ &+ \frac{R^2 S}{N^3}\left[\sum_{i=0}^{2} x_i^2 q_i - \left(\sum_{i=0}^{2} x_i q_i\right)^2\right]\end{aligned} \tag{4.15}$$

Under the $H_0$ in Eq. (4.6), $\sigma_a^2$ reduces to $\sigma_0^2$ by the Eq. (4.11). If $\sigma_0^2$ is known, The corresponding power and sample size for the trend tests are,

$$\begin{aligned}\text{Power}^* &= P_{H_a}\left(|Z_T^* > |z_{1-\alpha/2}\right) \\ &= \Phi\left(\frac{-z_{1-\alpha/2}\sigma_0 - N^{1/2}\mu_a}{\sigma_a}\right) + 1 \\ &\quad - \Phi\left(\frac{z_{1-\alpha/2}\sigma_0 - N^{1/2}\mu_a}{\sigma_a}\right)\end{aligned} \tag{4.16}$$

$$N^* = \left(\frac{z_{1-\alpha/2}\sigma_0 - z_{1-\beta}\sigma_a}{\mu_a}\right)^2 \tag{4.17}$$

where $\Phi$ is the cumulative distribution function $(cdf)$ of the standard normal distribution. The values $p_i$ and $q_i$ in Eqs. (4.16) and (4.17) can be calculated using Eq. (4.7). Since, $\sigma_0^2$ is often unknown, so the statistic $Z_T$ in Eq. (4.13) with the estimated variance, $\hat{\sigma}_0^2$ is applied to the data in practice, and $\hat{\sigma}_0^2$ converges to $(\tilde{\sigma}_0^2 + \mu_a^2)^{1/2}$ under the $H_a$ in probability, where

$$\begin{aligned}\tilde{\sigma}_a^2 = \frac{R^2 S}{N^3}&\left[\sum_{i=0}^{2} x_i^2 p_i - \left(\sum_{i=0}^{2} x_i p_i\right)^2\right] \\ &+ \frac{RS^2}{N^3}\left[\sum_{i=0}^{2} x_i^2 q_i - \left(\sum_{i=0}^{2} x_i q_i\right)^2\right]\end{aligned} \tag{4.18}$$

Hence,

$$\begin{aligned}\text{Power} &= Pr_{H_a}\left(|Z_T > |z_{1-\alpha/2}\right) \\ &= Pr_{H_a}\left(\frac{U}{N^{1/2}\hat{\sigma}_0} < -z_{1-\alpha/2}\right) + Pr_{H_a}\left(\frac{U}{N^{1/2}\hat{\sigma}_0} > z_{1-\alpha/2}\right) \\ &= Pr_{H_a}\left(\frac{U}{N^{1/2}(\tilde{\sigma}_0^2 + \mu_a^2)^{1/2}} < -z_{1-\alpha/2}\right)\end{aligned} \tag{4.19}$$

$$+Pr_{H_a}\left(\frac{U}{N^{1/2}(\tilde{\sigma}_0^2+\mu_a^2)^{1/2}} > z_{1-\alpha/2}\right)$$

$$= \Phi\left(\frac{-z_{1-\alpha/2}(\tilde{\sigma}_0^2+\mu_a^2)^{1/2} - N^{1/2}\mu_a}{\sigma_a}\right) + 1$$

$$-\Phi\left(\frac{z_{1-\alpha/2}(\tilde{\sigma}_0^2+\mu_a^2)^{1/2} - N^{1/2}\mu_a}{\sigma_a}\right)$$

and the required sample size,

$$N = \left(\frac{z_{1-\alpha/2}(\tilde{\sigma}_0^2+\mu_a^2)^{1/2} - z_{1-\beta}\sigma_a}{\mu_a}\right)^2 \tag{4.20}$$

When $R = S$, $\tilde{\sigma}_a^2 = \sigma_a^2$, and using the Taylor series expansion Eq.(4.20) can be approximated by,

$$N \approx \frac{\left[\left(z_{1-\alpha/2} - z_{1-\beta}\right)\sigma_a\right]^2}{\mu_a^2} + z_{1-\alpha/2}\left(z_{1-\alpha/2} - z_{1-\beta}\right) \tag{4.21}$$

The investigation of this study showed that the sample size obtained using Eqs. (4.20) and (4.21) are consistent rather than that obtained using Eq. (4.17).

The summary of the study is as follows: If the mode of inheritance is known, the optimum test for that model is preferable. If not known, then a robust test should be used. Usually, the MAX is the superior to the simpler MERT. But, in case when the plausible set of models is smaller, viz. a recessive model, the simpler method would be the choice. The robust procedures provides more protection against an incorrect choice of model than using the additive scores. To achieve the desired level of type I and type II errors sample sizes should be calculated under an assumed alternative and use the corresponding optimal test. But, when the specification of the underlying model is not known, a robust test should be used. The sample sizes determined by using the optimal test for a particular model are too small if a different model holds. Hence, the sample size should be increased to allow for the loss of power incurred by model uncertainty. The use of an efficient robust method helps to

minimize the worst loss of power or needed increase in sample size over the range of plausible models.

### *4.3.2 A rhombus formula to approximate the p-value of MAX-test [Li et al., 2008]*

This method provides a simple approximation for the $p$-value of the MAX-test with or without adjusting for the covariates. The new method avoids resampling steps and thus makes the MAX-test readily applicable to GWAS. The probability was calculated for a symmetric rejection region about the origin. A tight upper bound was also given for the $Pr(T_{max} > t)$ where $t$ is a given value of $T_{max}$. Here, $T_{max}$ is calculated either from MAX-test statistic based on the trend tests,

$$Z_{max} = (|Z_R|, |Z_A|, |Z_D|) \tag{4.22}$$

or MAX-test statistic based on the Wald test,

$$W_{max} = (|W_R|, |W_A|, |W_D|) \tag{4.23}$$

Let there are $k$ test statistics, $T_1, T_2, \cdots, T_k$, each of which is used to test the $H_0$, under which these $k$ test statistics approximately follow the standard normal distribution whose density and probability functions are denoted by $\phi(x)$ and $\Phi(t)$, respectively. Assume that the correlation coefficient $\mathrm{cor}(T_i, T_j)$, for $i, j \in \{1,2, \cdots, k\}$, is either known or can be estimated consistently. For an observed value of $T_{max} = t$, Efron [1997] developed the upper bound for a one-sided rejection region. The rhombus formula based on the techniques used in Efron's one is given by,

$$
\begin{aligned}
Pr(T_{max} > t) \leq (k-2)[\Phi(t) - \Phi(-t) - 1] - \frac{4\phi(t)(k-1)}{t} \\
+ \frac{4\phi(t)}{t} \sum_{i=1}^{k-1} \left\{ 2\Phi\left(\frac{tL_{i(i+1)}}{2}\right) + e^{-\frac{t^2 L_{i(i+1)}^2}{8}} \left[ \Phi\left(\frac{t\left(\pi - L_{i(i+1)}\right)}{2}\right) \right.\right. \\
\left.\left. - \Phi\left(\frac{tL_{i(i+1)}}{2}\right) \right] \right\} I\left\{0 \leq L_{i(i+1)} \leq \frac{\pi}{2}\right\} \\
+ \frac{4\phi(t)}{t} \sum_{i=1}^{k-1} \left\{ 2\Phi\left(\frac{t\left(\pi - L_{i(i+1)}\right)}{2}\right) + e^{-\frac{t^2\left[\pi - L_{i(i+1)}\right]^2}{8}} \right. \\
\left. - \Phi\left(\frac{t\left(\pi - L_{i(i+1)}\right)}{2}\right) \right] \right\} I\left\{\frac{\pi}{2} \leq L_{i(i+1)} \leq \pi\right\}
\end{aligned}
\tag{4.24}
$$

where $L_{ij} = \arccos\left(\text{cor}\left(T_i, T_j\right)\right)$ and $I\{\cdot\}$ is an indicator function. The estimated upper bound for the $p$-value in the Eq. (4.24) depends on how these $k$ test statistics are indexed, but $Pr(T_{max} > t)$ is independent of the index. The rhombus formula in Eq. (4.24) provides a theoretical upper bound for the true $p$-value if $(T_1, T_2, \cdots, T_k)$ follows a joint normal distribution with a known correlation matrix. In real applications, $(T_1, T_2, \cdots, T_k)$ is asymptotically normal and a tighter upper bound can be computed by comparing the upper bound evaluated under all possible orderings of the $k$ test statistics and choosing the smallest value as the estimation for the $p$-value. This strategy is applicable for $Z_{max}$ and $W_{max}$ with $k = 3$ and thus the true $p$-value of the MAX-test is not necessarily less than the bound calculated by the rhombus formula. The numerical evaluation of this method shows that the rhombus formula tended to overestimate the true $p$-values.

### *4.3.3 The optimal dose-effect mode trend test [Yamada & Okada, 2008]*

This study proposed a new test based on the relations among the $\chi^2_{df=2}$ (dominant), $\chi^2_{df=1}$ (recessive) and the trend (additive) tests which tests the $H_0$ that the disease-susceptible allele has a dose-effect somewhere between the recessive and dominant modes and this corresponds to the optimal dose-effect for the observed data. The theoretical development focused on the constrained range of maximum likelihood method [Wang & Sheffield, 2005] where the mean genetic effect of the heterozygous genotype is between the two homozygous genotypes. The contingency table-based $\chi^2$ statistic was used instead of

likelihood-ratio statistic and the $p$-value was calculated based on the exact occurrence probability.

Let us consider the Table 3.1 (Chapter 3), a $2 \times 3$ observed table $(\tau)$ resulting from a case control association test for three genotypes of a SNP with three genotypes "$aa$", "$aA$" and "$AA$". The $\chi^2_{df=2}$ statistic for $\tau$ is,

$$\chi^2_{df=2} = \sum_{i,j} \frac{d_{ij}^2}{e_{ij}} \tag{4.25}$$

and the general trend $\left(\chi^2_{df=1}\right)$ is determined as [van Bell et al., 2004],

$$\Upsilon^2(w) = \frac{n_{..}^2}{n_{1.}\, n_{2.}} \frac{\left(\sum_j \left(d_{1j}\left(w_j - W\right)\right)\right)^2}{\left(\sum_j \left(n_{.j}\left(w_j - W\right)\right)\right)^2} \tag{4.26}$$

where $d_{ij} = n_{ij} - e_{ij}$, $n_{ij} = \{n_{10}, n_{11}, n_{12}, n_{20}, n_{21}, n_{22}\} = \{r_0, r_1, r_2, s_0, s_1, s_2\}$, $e_{ij}$ are the expected frequencies under the $H_0$, $N = n_{..}$, $R = n_{1.}, S = n_{2.}$, $w = \{w_0, w_1, w_2\}$ $(|w| \neq 0)$ is a weight vector for the three genotypes and $W = \sum_j \left(\frac{w_j n_{.j}}{N}\right)$. Since $\Upsilon^2(w)$ cannot be defined from the Eq. (4.26) when $w_0 = w_1 = w_2$ because of $W = w_0 = w_1 = w_2$. So, the meaningful weight vectors can be divided into two groups,

$$w_{g_1} = \{w_0, w_1, w_2 \neq w_0\} \tag{4.27}$$

$$w_{g_2} = \{w_0, w_1 \neq w_0, w_2 = w_0\} \tag{4.28}$$

In terms of the weight vectors, the Eq. (4.26) can be rewritten as,

$$\Upsilon^2(w) = \Upsilon^2\left(w' = \left\{0, r = \frac{w_1 - w_0}{w_2 - w_0}, 1\right\}\right) \text{ for weight vector } w_{g_1} \tag{4.29}$$

$$\Upsilon^2(w) = \Upsilon^2(w' = \{0, w_1 - w_0, 0\}) = \Upsilon^2(w'' = \{0, 1, 0\}) \quad \text{for weight vector } w_{g_2} \tag{4.30}$$

The $\Upsilon^2(w)$ can be written as a function of $r \in \mathbf{R} = \{r| -\infty \leq r \leq \infty\}$,

$$\begin{cases} \Upsilon^2(r) = \Upsilon^2(w = \{0, r, 1\}) & \{r| -\infty < r < \infty\} \\ \Upsilon^2(r) = \Upsilon^2(w = \{0,1,0\}) & \{r|r = \pm\infty\} \end{cases} \tag{4.31}$$

The weight vectors of additive, dominant and recessive modes for the allele ‘‘$A$’’ are,

$$w_A = \{0, 0.5, 1\},\ w_D = \{0, 1, 1\} \text{ and } w_R = \{0, 0, 1\} \tag{4.32}$$

So, the trend statistics for the additive, dominant and recessive modes are given as $\Upsilon^2(0.5)$, $\Upsilon^2(1)$ and $\Upsilon^2(0)$, respectively. The $\chi^2_{df=2}$in the form of $\Upsilon^2(r)$ is,

$$\chi^2_{df=2} = \max\big(\Upsilon^2(r)\big); r \in \mathbf{R} = \{r| -\infty \leq r \leq \infty\} \tag{4.33}$$

The statistic in Eq. (4.33) takes the maximum value when,

$$r = r_0 = \frac{\frac{n_{11}}{n_{.1}} - \frac{n_{10}}{n_{.0}}}{\frac{n_{12}}{n_{.2}} - \frac{n_{10}}{n_{.0}}} \tag{4.34}$$

Therefore,

$$\chi^2_{df=2} = \Upsilon^2(r_0) \tag{4.35}$$

The statistic $\Upsilon^2(0.5)$ is almost identical to the CATT statistic,

$$\chi^2_{\text{CATT}} = \frac{n_{..} - 1}{n_{..}\ \Upsilon^2(0.5)} \tag{4.36}$$

Here, $\Upsilon^2(0.5)$ is treated as the representative one as the difference between $\Upsilon^2(0.5)$ and $\chi^2_{\text{CATT}}$ is trivial and the $\Upsilon^2(1)$ and $\Upsilon^2(0)$ are equal to the $\chi^2$ statistics for the corresponding $2 \times 2$ tables for the dominant and the recessive models, respectively.

For any subset $\mathbf{r} \in \mathbf{R}$ (a rational number field), the new best-fit mode trend statistic $\left(W^2(r)\right)$ is,

$$W^2(r) = \max\left(\Upsilon^2(r)\right),\ r \in \mathbf{r} \tag{4.37}$$

Using the Eq. (4.37), the trend statistics for the additive, recessive and dominant modes and $\chi^2_{df=2}$ become the best-fit mode trend statistics when $\mathbf{r} = \{0.5\}, \mathbf{r} = \{1\}, \mathbf{r} = \{0\}$ and $\mathbf{r} = \{r| -\infty \leq r \leq \infty\}$, respectively, and they can be expressed in terms of $W^2$ as,

$$\begin{aligned} \Upsilon^2(0.5) = \Upsilon^2_A = W^2(\mathbf{r}_A = \{r|r = 0.5\}) \\ \Upsilon^2(1) = \Upsilon^2_D = W^2(\mathbf{r}_D = \{r|r = 1\}) \\ \Upsilon^2(0) = \Upsilon^2_R = W^2(\mathbf{r}_R = \{r|r = 0\}) \\ \Upsilon^2(r_0) = \chi^2_{df=2} = W^2\left(\mathbf{r}_{df=2} = \{r| -\infty \leq r \leq \infty\}\right) \end{aligned} \tag{4.38}$$

For $\hat{\mathbf{r}}$, the best-fit mode trend statistic, $W^2(\hat{\mathbf{r}})$ is regarded as the optimal dose-effect mode trend statistic. Since, $\Upsilon^2(r)$ takes the maximum value when $r = r_0$ and has two extrema at most, the $W^2(\hat{\mathbf{r}})$ can be simplified as,

$$W^2(\hat{\mathbf{r}}) = \begin{cases} \Upsilon^2(r_0) & \text{when } r_0 \in \hat{\mathbf{r}} \\ \max\begin{cases}\Upsilon^2(0) \\ \Upsilon^2(1)\end{cases} & \text{when } r_0 \notin \hat{\mathbf{r}} \end{cases} \tag{4.39}$$

The probability density function $(pdf)$ of each of the test statistics defined in the Eq. (4.38) are unknown. Since, the $pdf$s of $\Upsilon^2(0.5)$, $\Upsilon^2(1)$, $\Upsilon^2(0)$ and $\Upsilon^2(r_0)$ are known, the $p$-values for the conventional $\chi^2$ tests are,

$$\begin{aligned} &Pr(\Upsilon^2(0.5)|df = 1) \\ &Pr(\Upsilon^2(1)|df = 1) \\ &Pr(\Upsilon^2(0)|df = 1) \\ &Pr(\Upsilon^2(r_0)|df = 2) \end{aligned} \tag{4.40}$$

The statistical significance of $W^2(\hat{\mathbf{r}})$ was based on the exact occurrence probabilities of ,

$$T = \{t(i,j)\} \tag{4.41}$$

where $i, j \in I$, represents the counts of $n_{10}$ and $n_{12}$ that is a set of tables that shares the same marginal counts with the observed table $\tau$. Let $W^2(\hat{\mathbf{r}})(\tau)$ and $Pex\big(W^2(\hat{\mathbf{r}})(\tau)\big)$ denote the optimal dose-effect mode trend statistic for a table $\tau$ and its statistical significance, respectively. $Pex\big(W^2(\hat{\mathbf{r}})(\tau)\big)$ is defined as,

$$Pex\big(W^2(\hat{\mathbf{r}})(\tau)\big) = \sum_{t\in T} Pr(t)\ \delta(t|\tau) \tag{4.42}$$

$$\delta(t|\tau) = \begin{cases} 1 & \text{when } W^2(\hat{\mathbf{r}})(t) \le W^2(\hat{\mathbf{r}})(\tau) \\ 0 & \text{otherwise} \end{cases} \tag{4.43}$$

where $Pr(t)$ is the exact probability of occurrence of table $t$, so that,

$$Pr(t) = \prod_i n_{i.}! \prod_j n_{.j}! / n_{..}! \prod_{i,j} n_{ij}! \tag{4.44}$$

The distribution of the new test statistic was smoother than MAX-test because the proposed method assigned equal weights to all modes between the recessive and dominant modes. This smoothness is an indicator of the superiority over the MAX-test where the MAX-test make an exception for three particular modes and this exceptional treatment corresponded to the ridges in the plots. The lack of this exceptional treatment for a particular node is an advantage when the researcher's aim is to test the $H_0$ that the disease susceptible allele has a dose-effect somewhere between the recessive and dominant modes.

### *4.3.4 A conditional MAX-test of scores-specific linear-by-linear association tests [Hothorn LA & Hothorn T, 2009]*

This is a reformulated CATT for proportions where a conditional maximum test of scores-specific linear-by-linear association was proposed. Here, the asymptotic expectation and covariance were derived in a closed form for the maximum type, sum and quadratic test

statistics and the limiting distribution was known. The computation of association studies can be performed using the software provided by Strasser and Weber [1999]. Both the limiting distribution and approximations of the exact conditional distribution can easily be computed using standard software packages. This idea was extended to the area of application to stratified designs, studies involving more than two groups and the simultaneous analysis of multiple loci by means of multiplicity adjusted $p$-values for the underlying multiple CATT. This study embed the MAX-test as suggested by Freidlin et al. [2002] into the flexible framework for conditional independence tests introduced by Strasser and Weber [1999]. The reformulation of the MAX-test into a general framework for conditional inference procedures and the derivation of its limiting distribution and propose extensions to stratified designs, more than two groups and multiple loci were the main purpose of this research.

Let us consider genotype distribution for a case control study as in Table 3.1 where "*A*" is the high risk candidate allele and "*a*" is any of the other alleles of a simple bi-allelic marker. Here, the interest is in a comparison of the genotype distributions, i.e. the penetrances $f_j = P(\text{case}/j)$ for $j \in \{aa,\ aA,\ AA\}$:

$$H_0: f_{aa} = f_{aA} = f_{AA} \text{ versus } H_a: f_{aa} \leq f_{aA} \leq f_{AA} \tag{4.45}$$

where at least one of the inequalities in the $H_a$ is strict.

The CATT statistic with scores $\xi = (\xi_{aa}, \xi_{aA}, \xi_{AA})$ is essentially,

$$\text{CATT}(\xi) = \sum_{j \in \{aa,aA,AA\}} \xi_j r_j \tag{4.46}$$

where $r_j$ are the weights for the cases as presented in Table 4.1.

**Table 4.1: Genotype distribution reformulated.**

| $i$ | $Y_i$ | $X_i$ | $w_i$ | $h(Y_i)$ | $g_A(X_i)$ | $g_D(X_i)$ | $g_R(X_i)$ |
|---|---|---|---|---|---|---|---|
| 1 | Case | $aa$ | $r_{aa}$ | 1 | 0 | 0 | 0 |
| 2 | Case | $Aa$ | $r_{Aa}$ | 1 | 1 | 1 | 0 |
| 3 | Case | $AA$ | $r_{AA}$ | 1 | 2 | 1 | 1 |
| 4 | Control | $aa$ | $s_{aa}$ | 0 | 0 | 0 | 0 |
| 5 | Control | $Aa$ | $s_{Aa}$ | 0 | 1 | 1 | 0 |
| 6 | Control | $AA$ | $s_{AA}$ | 0 | 2 | 1 | 1 |

Let $Y_i$ denote the case and control status and $X_i$ the genotype for all cells $i = 1,2,\cdots,n = 6$. The weights $w_i$ represent the number of observations in each cell with total number of observations $N = \sum_i w_i$. The influence function $h$ provides a zero-one dummy coding of the groups (cases = 1 and controls = 0). Moreover, three transformations $g$ of the genotype are under test: $g_D$ assigns scores $\xi_D$ to genotypes $(aa,\ aA,\ AA)$, $g_A$ assigns scores $\xi_A$ and $g_R$ implements scores $\xi_R$, as shown in Table 4.1.

The interest was to test the $H_0$ of independence of grouping $Y$ and genotype $X$, i.e. the equality of the distribution $\mathcal{D}(Y)$ to the conditional distribution $\mathcal{D}(Y|X)$ of $Y$ given $X$,

$$H_0: \mathcal{D}(Y|X) = \mathcal{D}(Y) \tag{4.47}$$

against ordered alternatives. First, a three-dimensional statistic $T$ was defined, each dimension being associated with one of the scores $g_A, g_D, g_R$. Each statistic was defined by the sum of the scores multiplied by the weights associated with cases, i.e. was equivalent to the CATT statistic in Eq. (4.46),

$$T = (T_A, T_A, T_R) = \sum_{i=1}^{n} w_i g(X_i) h(Y_i) \ \in \mathbb{R}^3 \tag{4.48}$$

with $g(X_i) = (g_A(X_i), g_D(X_i), g_R(X_i))$. Thus, the three-dimensional linear statistic $T$ is the vector of the unstandardized CATT statistics $(\mathrm{CATT}(\xi_A), \mathrm{CATT}(\xi_D), \mathrm{CATT}(\xi_R))$ for the dominant, additive, and recessive model.

The distribution of $T$ depends on the joint distribution of $Y$ and $X$, which is unknown under almost all practical circumstances. The conditional expectation of the influence function $h$ is [Strasser & Weber, 1999],

$$\mathrm{E}(h) = N^{-1} \sum_i w_i h(Y_i) \in \mathbb{R} \tag{4.49}$$

with corresponding variance,

$$\mathrm{var}(h) = N^{-1} \sum_i w_i (h(Y_i) - \mathrm{E}(h))^2 \tag{4.50}$$

The conditional expectation of the linear statistic T is,

$$\mu = \mathrm{E}(T) = \mathrm{E}(h) \sum_{i=1}^{n} w_i g(X_i) \tag{4.51}$$

and the covariance,

$$\begin{aligned} \Sigma &= \mathrm{var}(T) \\ &= \frac{N}{N-1} \mathrm{var}(h) \left( \sum_i w_i (g(X_i) g(X_i)^{\mathrm{T}}) \right) \\ &\quad - \frac{1}{N-1} \mathrm{var}(h) \left( \sum_i w_i g(X_i) \right) \left( \sum_i w_i g(X_i) \right)^{\mathrm{T}} \end{aligned} \tag{4.52}$$

The three-dimensional expectation $\mu$ and the three diagonal elements of the covariance matrix $\Sigma$ contain the mean and the variances for the additive, dominant, and recessive CATT statistics under $H_0$, as given in Eqs. (4.46) and (4.48), respectively.

Under $H_0$, the test statistics and their distribution under the conditions described in the $H_0$ can be defined based on the three-dimensional statistic $T$ with its expectation $\mu$ and covariance matrix Σ. As $N \to \infty$, the limiting distribution of the three-dimensional statistic $T$ is a three-dimensional normal distribution with expectation $\mu$ and covariance Σ [Strasser and Weber, 1999]. Thus, the asymptotic distribution of a maximum type statistic,

$$c_{\max}(T, \mu, \Sigma) = \max \left| \frac{T - \mu}{\mathrm{diag}(\Sigma)^{1/2}} \right| \tag{4.53}$$

can be evaluated by computing three-dimensional normal probabilities. Here, $\mathrm{diag}(\Sigma)^{1/2}$ are the conditional standard deviations of the elements of $T$. Alternatively, either the sum (or average) statistic,

$$c_{\mathrm{sum}} = \frac{1^{\mathrm{T}} T - 1^{\mathrm{T}} \mu}{1^{\mathrm{T}} \Sigma 1}, 1 = (1, 1, 1)^{\mathrm{T}} \tag{4.54}$$

or a quadratic form based on the Moore–Penrose inverse $\Sigma^{+}$ of the conditional covariance matrix Σ, i.e.

$$c_{\mathrm{quad}}(T, \mu, \Sigma) = (T - \mu)^{\mathrm{T}} \Sigma^{+} (T - \mu) \tag{4.55}$$

can be used. Eqs. (4.53) and (4.55) follow a standard normal or a $\chi^2$ distribution with 2-$df$, respectively.

Under any circumstances, the exact conditional distribution can be approximated by conditional Monte-Carlo methods, which is especially attractive for small sample sizes when asymptotic distributions are not expected to work well [Hothorn LA & Hothorn T, 2009].

### *4.3.5 An analytic approach to robust association tests under the covariates [So & Sham, 2011]*

This research developed an analytical approach of the most robust MAX-test for both the quantitative or binary traits as well as covariates. They considered the challenge of multiple testing to prevent the inflation of type I error rate and modified the existing MAX-test based on the score statistic. The comparison of theoretically calculated $p$-values well matched with the resampling ones. For computation, the researchers provided an R-package "RobustSNP" which is available at,
http://sites.google.com/site/honcheongso/software/robustsnp.

According to Lin [2005a], all the commonly employed statistics are related to the score statistic and can be expressed as,

$$T_j = U_j' V_j^{-1} U_j \tag{4.56}$$

where the subscript $j$ refers to the $j$-th hypothesis that is to be tested and,

$$U_j = \sum_{i=1}^{n} U_{ji} \tag{4.57}$$

$$V_j = \sum_{i=1}^{n} U_{ji}\, U_j' \tag{4.58}$$

where $U_{ji}$ is the score function calculated from data from the $i$-th subject only and $n$ refers to the sample size.

When the $j$-th hypothesis is truly null, $U_j$ is approximately normally distributed with mean 0 and covariance matrix $V_j$ in large samples. Hence, Eq. (4.56) follows an approximately $\chi^2$ distribution with $df$ equal to the dimension of $U_j$.

If a total of $m$ hypotheses to be tested and all of them are truly null, with large samples, $(U_1, U_2, \cdots, U_m)$ follows approximately a multivariate normal distribution with mean vector 0 and the covariance between $U_j$ and $U_k$ of any two hypothesis tests $j$ and $k$ is,

$$V_{jk} = \sum_{i=1}^{n} U_{ji}\, U'_{ki} \tag{4.59}$$

The Eq. (4.59) was used as a base of this research to correct for testing of multiple genetic models. Here, the score statistic for each of the three genetic models (dominant, recessive and additive) were constructed and the covariance matrix of the three statistics were calculated under the $H_0$ by the Eq. (4.59) where the appropriate significance level was obtained by trivariate integration.

When covariates are present, $U_{ji}$ in the Eq. (4.59) represent the $i$-th subject's efficient score function for $\beta_j$, the parameter of interest [Bickel et al. 1993; Lin 2005a, b]. Then,

$$U_{ji} = U_{\beta_j,i} - V_{\beta_j,\alpha_j} V_{\alpha_j\alpha_j}^{-1} U_{\alpha_j,i} \tag{4.60}$$

where $U_{\beta_j,i}$ and $U_{\alpha_j,i}$ are the score function for the $i$-th subject for parameters $\beta_j$ and $\alpha_j$; $\alpha_j$ being the nuisance parameters. $V_{\beta_j,\alpha_j}$ and $V_{\alpha_j,\alpha_j}$ are sub-matrices of the limiting Fisher information matrix of $\beta_j$ and $\alpha_j$,

$$V_{\beta_j,\alpha_j} = \mathrm{cov}\left(U_{\beta_j}, U_{\alpha_j}\right) \tag{4.61}$$

$$V_{\alpha_j,\alpha_j} = \mathrm{var}\left(U_{\alpha_j}\right) \tag{4.62}$$

Eqs. (4.61) and (4.62) are the covariances and variances, respectively. For testing the effect of a genetic marker under different genetic models, with or without covariates, a single test was considered for simplicity and hence the subscript $j$ will be dropped. Let $y_i$ be the measured outcome, $X_{gi}$ be the coding of the genotype and $X_{ei}$ be a vector of environmental covariates including 1 as the first element (for the intercept) for the $i$-th subject, $X_{gi}$ is

coded differently under different genetic models by denoting the three genotypes of a marker by $aa$, $Aa$ and $AA$, they will be coded as (0, 1, 2), (0, 1, 1) and (0,0,1) under additive, dominant and recessive models, respectively, where "$A$" is assumed to be the risk allele.

Assume the outcome y and the predictor variables $(X_{gi}, X_{ei})$ are related through a generalized linear model **(**GLM),

$$\eta_i = {X_{ei}}'\alpha + {X_{gi}}'\beta = Z_i'\gamma \tag{4.63}$$

where $Z_i = ({X_{ei}}', {X_{gi}}')$ and $\gamma = (\alpha, \beta)$. Here, $\alpha$ and $\beta$ reflect the effects of the environmental covariate and genetic marker on the outcome respectively. $\eta$ is related to the actual outcome $y$ through the link function $f$ so that,

$$\mathrm{E}(y_i|Z_i) = f^{-1}\eta_i \tag{4.64}$$

The likelihood of the observed outcome $y_i$ given covariates $Z_i$ for the $i$-th subject is,

$$L(y_i|Z_i) = \exp\left[\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right] \tag{4.65}$$

where $a$, $b$ and $c$ are known functions and $\phi$ is the dispersion parameter.

The score function for genetic markers for the parameter $\beta$, with adjustment for environmental covariates, can be written as,

$$U_\beta = \sum_{i=1}^{n} \frac{\delta \log L_i}{\delta \beta} = \sum_{i=1}^{n} \frac{y_i - \tilde{y}_i}{a(\phi)} X_{gi} \tag{4.66}$$

The fitted values $\tilde{y}_i$ need not to be calculated as the score test is constructed under the $H_0$: $\beta = 0$, and to be calculated when a large number of SNPs is tested. The contribution from the $i$-th subject is,

$$U_{\beta,i} = \frac{\delta \log L_i}{\delta \beta} = \frac{y_i - \widetilde{y_i}}{a(\phi)} X_{gi} \tag{4.67}$$

Similarly,

$$U_{\alpha,i} = \frac{\delta \log L_i}{\delta \alpha} = \frac{y_i - \widetilde{y_i}}{a(\phi)} X_{ei} \tag{4.68}$$

The variance and covariance of the score functions of $\alpha$ and $\beta$ are,

$$V_{\alpha\alpha} = \sum_{i=1}^{n} \frac{b''(\eta_i)}{a(\phi)} X_{ei} X_{ei}' \tag{4.69}$$

$$V_{\alpha\beta} = \sum_{i=1}^{n} \frac{b''(\eta_i)}{a(\phi)} X_{ei} X_{gi}' \tag{4.70}$$

$$V_{\beta\beta} = \sum_{i=1}^{n} \frac{b''(\eta_i)}{a(\phi)} X_{gi} X_{gi}' \tag{4.71}$$

Based on the above results, the $i$-th subject's contribution to the efficient score function can be calculated by,

$$U_i = U_{\beta,i} - V_{\alpha\beta} V_{\alpha\alpha}^{-1} U_{\alpha,i} \tag{4.72}$$

The forms of $a(\phi)$, $b''(\eta_i)$ and $\widetilde{y_i}$ for linear, logistic and Poisson regressions are given by Schaid et al. [2002]. The efficient score functions were calculated for each subject and for each genetic model. Since each test is 1 $df$, the $z$ statistic was used,

$$z = {U_j}\Big/{\sqrt{V_j}} \tag{4.73}$$

For two $z$-statistic from two genetic models, $z_j$ and $z_k$, the covariance between them is,

$$\begin{aligned}\operatorname{cov}(z_j, z_k) &= \frac{\operatorname{cov}(U_j, U_k)}{\sqrt{V_j}\sqrt{V_k}} \\ &= \frac{\sum_{i=1}^{n} U_{ji} U_{ki}{}'}{\sqrt{\sum_{i=1}^{n} U_{ji} U_{ji}{}'}\ \sqrt{\sum_{i=1}^{n} U_{ki} U_{ki}{}'}}\end{aligned} \tag{4.74}$$

Let the observed maximum $z$ statistic be $c$ and the maximum $z$ under the complete $H_0$ be $z_{null,max}$. Hence,

$$\begin{aligned}p_{\text{corrected}} &= 1 - Pr(|z_{null,max}| \le c) \\ &= 1 - \int_{-c}^{c}\int_{-c}^{c}\int_{-c}^{c} \varphi_3(\mathbf{z}; \mathbf{0}, \boldsymbol{\Sigma}) d\mathbf{z}\end{aligned} \tag{4.75}$$

where $\varphi_3$ is the trivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$. The integral is computed by numerical methods [Genz, 1992] implemented in the R-package "mvtnorm".

### *4.3.6 A robust MAX3-test based on the Mantel-Haenszel test (MHT) [Zang & Fung, 2011]*

This test avoided estimating the nuisance parameters induced by the covariates and thus valid under both large and small numbers of partial tables while still enjoys the property of efficiency robustness. The asymptotic distribution of the test under the $H_0$ was also derived that makes the calculation of $p$-values easier. This study proved that the new test can be equally derived through a conditional likelihood. So, the original MAX3 based on the trend tests or the matching trend tests can be treated as a special case and generally incorporated into the newly proposed test.

Let "$A$" is the risk allele for the genotype distribution of a SNP given in the Table 3.1. Based on the values taken by the covariates, the individuals in a case-control genetic association study are stratified into $K$ partial tables indexed by $C_k$ $(k = 1,2,\cdots,K)$. The genotype counts,$(G_0, G_1, G_2) = (aa, aA, AA)$ in cases and controls are given in the Table 4.2 for the partial table.

**Table 4.2: Genotype distribution for partial table $C_k$.**

| Phenotypes | Genotypes | | | Total |
|---|---|---|---|---|
| | $aa$ | $aA$ | $AA$ | |
| Cases | $n_{10k}$ | $n_{11k}$ | $n_{12k}$ | $n_{1+k}$ |
| Controls | $n_{20k}$ | $n_{21k}$ | $s_{22k}$ | $n_{2+k}$ |
| Total | $n_{+0k}$ | $n_{+1k}$ | $n_{+2k}$ | $n_{++k}$ |

Here, $n_{1+k} = \sum_{i=0}^{2} n_{1ik}$ , $n_{2+k=} \sum_{i=0}^{2} n_{2ik}$ , $n_{+ik=} n_{1ik} + n_{2ik}$ $(i = 0,1,2)$ and $n_{++k} = n_{1+k} + n_{2+k}$. In the partial table $C_k$, let penetrance is,

$$f_{ik} = P(\text{case}|G_i, G_k) \tag{4.76}$$

The disease prevalence is,

$$d_k = P(\text{case}|C_k) = \sum_{i=0}^{2} f_{ik} P(G_i, |G_k) \tag{4.77}$$

If the genetic model is known, a generalized MHT [Mantel, 1963] can be used to test for association based on the $2 \times 3 \times k$ contingency table. Specifically, under $H_0$ and conditional on the marginal totals of the partial table $C_k$, $(n_{11k}, n_{12k})$ follows a multivariate hypergeometric distribution,

$$P(n_{11k} = t_1, n_{12k} = t_2) = \frac{\left(\prod_{i=1}^{2} n_{i+k}\,!\right)\left(\prod_{j=0}^{2} n_{+jk}\,!\right)}{n_{++k}!\,(n_{1+k} - t_1 - t_2)!\,t_1!\,t_2!\,(n_{+0k} - n_{1+k} + t_1 + t_2)!\,q_1!\,q_2!} \tag{4.78}$$

where $t_1 = 0, \cdots, \min\{n_{+1k}, n_{1+k}\}$ and $t_2 = 0, \cdots, \min\{n_{+2k}, n_{1+k} - t_1\}$, $q_1 = (n_{+1k} - t_1)$, $q_2 = (n_{+2k} - t_1)$.

Now, the test statistic of the generalized MHT [Mantel, 1963] is,

$$Z(x) = \frac{\sum_{k=1}^{K}\left(T_k(x) - \mu_k(x)\right)}{\sqrt{\sum_{k=1}^{K} V_k(x)}} \tag{4.79}$$

where $T_k(x) = n_{12k} + x\, n_{11k}$ , $x$ is a pre-specified value between 0 and 1 and

$$\mu_k(x) = \mathrm{E}\left(T_k(x)\right) = \frac{n_{1+k}(n_{+2k} + x n_{+1k})}{n_{++k}} \tag{4.80}$$

$$\begin{aligned} & V_k(x) \\ & = \mathrm{var}\left(T_k(x)\right) \\ & = \frac{[n_{1+k} - n_{1+k}^2/n_{++k}][(x^2 n_{+1k} + n_{+2k}) - (x n_{+1k} + n_{+2k})^2/n_{++k}]}{n_{++k} - 1} \end{aligned} \tag{4.81}$$

The Eq. (4.79) has an approximate standard normal distribution under $H_0$.

An alternative expression of this test was derived using the conditional likelihood as described in subsection 4.3.4. Let $X_i$ $(i = 0,1,2)$ denote the scores for the genotypes which take three values 0, $x$, and 1 for the genotypes $G_0, G_1, G_2$, respectively where $x = 0, 0.5$ and 1 for the recessive, additive, and dominant models. The conditional likelihood was defined based on the marginal totals of each partial table,

$$\begin{aligned} & L(x) \\ & = \frac{\prod_{k=1}^{K}\prod_{i=0}^{2}[a/(1+a)]^{n_{2ik}}}{\prod_{k=1}^{K}\sum_{n_{11k}^*, n_{12k}^* \in W_k}\prod_{i=0}^{2}[a/(1+a)]^{n_{1ik}^*}[1/(1+a)]^{n_{2ik}^*}} \\ & = \frac{\prod_{k=1}^{K}\exp(\beta\sum_{i=0}^{2} X_i n_{1ik})}{\prod_{k=1}^{K}\sum_{n_{11k}^*, n_{12k}^* \in W_k}\exp(\beta\sum_{i=0}^{2} X_i n_{1ik}^*)} \end{aligned} \tag{4.82}$$

where $a = \exp(\alpha_k + \beta X_i)$ , $W_k$ is the set of all possible values of $n_{11k}$ and $n_{12k}$ maintaining the marginal totals of the partial table $C_k$ and $n_{10k}^* = n_{1+k} - n_{11k}^* - n_{12k}^*$ and $n_{2ik}^* = n_{+ik} - n_{1ik}^*$ $(i = 0,1,2)$. Then, the score function can be derived as,

$$U(x) = \left(\frac{\delta \log L(x)}{\delta \beta}\right)_{\beta=0} = \sum_{k=1}^{K}\left(n_{12k} + xn_{11k} - \frac{n_{1+k}(n_{+2k} + xn_{+1k})}{n_{++k}}\right) \tag{4.83}$$

and the observed information is,

$$I(x) = -\left(\frac{\delta^2 \log L(x)}{\delta^2 \beta}\right)_{\beta=0} = \sum_{k=1}^{K} \text{var}(n_{12k}, xn_{11k}) = \frac{[n_{1+k} - n_{1+k}^2/n_{++k}][(x^2 n_{+1k} + n_{+2k}) - (xn_{+1k} + n_{+2k})^2/n_{++k}]}{n_{++k} - 1} \tag{4.84}$$

The robust statistic based on the MHT $\left(Z_{max_{\text{MH}}}\right)$ by adjusting the covariate effects for the three genetic models viz. recessive, additive and dominant was,

$$Z_{max_{\text{MH}}} = \max\{|Z(0)|, |Z(0.5)|, |Z(1)|\} \tag{4.85}$$

where $Z(x)$ is from the Eq. (4.79) which is optimal for the recessive, additive and dominant models when $x = 0, 0.5$ and $1$, respectively. The variance-covariance matrix of the statistic (4.85) is,

$$V_{max_{\text{MH}}} = \text{cov}((Z_0, Z_{0.5}, Z_1)') = \begin{pmatrix} 1 & \rho_{0,0.5} & \rho_{0,1} \\ \rho_{0,0.5} & 1 & \rho_{0.5,1} \\ \rho_{0,1} & \rho_{0.5,1} & 1 \end{pmatrix} \tag{4.86}$$

Defining $n_k = (n_{11k}, n_{12k})'$ and with simple algebra under multivariate hypergeometric distribution expressed in Eq. (4.78),

$$\mathrm{cov}(n_k) = \begin{pmatrix} \rho_{n_k}^{(1,1)} & \rho_{n_k}^{(1,2)} \\ \rho_{n_k}^{(1,2)} & \rho_{n_k}^{(2,2)} \end{pmatrix} \tag{4.87}$$

where $\rho_{n_k}^{(1,1)} = \frac{n_{1+k}n_{+1k}(n_{++k}-n_{1+k})(n_{++k}-n_{+1k})}{n_{++k}^2(n_{++k}-1)}$, $\rho_{n_k}^{(2,2)} = \frac{n_{1+k}n_{+2k}(n_{++k}-n_{1+k})(n_{++k}-n_{+2k})}{n_{++k}^2(n_{++k}-1)}$

$\rho_{n_k}^{(1,2)} = -\frac{n_{1+k}n_{+1k}n_{+2k}(n_{++k}-n_{1+k})}{n_{++k}^2(n_{++k}-1)}$.

Now, using Eq. (4.78),

$$\begin{aligned} \rho_{x_i x_j} &= \mathrm{cov}\left(\frac{U(x_i)}{\sqrt{I(x_i)}}, \frac{U(x_j)}{\sqrt{I(x_j)}}\right) \\ &= \frac{\sum_{k=}^{K}\left[\rho_{n_k}^{(2,2)} + (x_i + x_j)\rho_{n_k}^{(1,2)} + x_i x_j \rho_{n_k}^{(1,1)}\right]}{\sqrt{I(x_i)I(x_j)}} \end{aligned} \tag{4.88}$$

where $x_i, x_j = 0, 0.5, 1$ with $i \neq j$ and the closed forms of $I(x_i)$, $\rho_{n_k}^{(1,1)}$, $\rho_{n_k}^{(1,2)}$, $\rho_{n_k}^{(2,2)}$ can be found in the Eqs. (4.84) and (4.87), respectively. $V_{max_{\mathrm{MH}}}$ is a singular matrix that is $Z_0, Z_{0.5}, Z_1$ are linear dependent [Zang et al., 2010a]. Hence,

$$\begin{aligned} \omega_0 &= \left(\rho_{0,0.5} - \rho_{0,1}\rho_{0.5,1}\right)/\left(1 - {\rho_{0,1}}^2\right) \\ \omega_1 &= \left(\rho_{0.5,1} - \rho_{0,1}\rho_{0,0.5}\right)/\left(1 - {\rho_{0,1}}^2\right) \end{aligned} \tag{4.89}$$

Now, the following relations can be defined,

$$Z_{0.5} = \omega_0 Z_0 + \omega_1 Z_1 \tag{4.90}$$

The null distribution of $Z_{\mathrm{MAX3_{MH}}}$ can be derived from the joint density function $f(Z_0, Z_1, \Sigma)$ of $(Z_0, Z_1)'$ where

$$\Sigma = \begin{pmatrix} 1 & \rho_{0,1} \\ \rho_{0,1} & 1 \end{pmatrix} \tag{4.91}$$

Eq. (4.91) defines the variance-covariance matrix of $(Z_0, Z_1)'$.

$$\begin{aligned} Pr\left(Z_{max_{\mathrm{MH}}} < t\right) &= Pr(|Z_0| < t, |\omega_0 Z_0 + \omega_1 Z_1| < t, |Z_1| < t) \\ &= 2 \int_0^{t(1-\omega_1)/\omega_0} \int_{-t}^{t} f(z_0, z_1, \Sigma)\, dz_1\, dz_0 \\ &\quad + 2 \int_{t(1-\omega_1)/\omega_0}^{t} \int_{-t}^{(t-\omega_0 z_0)/\omega_1} f(z_0, z_1, \Sigma) \end{aligned} \tag{4.92}$$

For an observed value $t^*$, the $p$-value of $Z_{max_{\mathrm{MH}}}$ can be obtained as,

$$P\left(Z_{max_{\mathrm{MH}}} < t^*\right) = 1 - Pr\left(Z_{max_{\mathrm{MH}}} < t^*\right) \tag{4.93}$$

### *4.3.7 The MAX-test derived from the Delta method and the generalized linear model (GLM) [Loley et al., 2013]*

This study derived the asymptotic distribution of the MAX-test using the GLM together with the Delta method which is applicable to the binary, quantitative, and survival traits. Also, it can be used for unrelated individuals, family-based studies, and matched pairs. The approach provided both the point and the interval estimates. The R-codes are available at http://www.nature.com/ejhg. By the simulation study, this research showed that the asymptotic MAX-test framework meets type I error levels well, has good power, and good model selection properties for minor allele frequencies$\geq 0.3$. The MAX-test approach derived in this work can be used for statistical testing, the calculation of confidence intervals, and the selection of the most plausible genetic model using the minimum $p$-value. The interval estimate is the new addition of the approach over the previous modifications.

They embed the MAX-test approach into the GLM and derived its limiting distribution. The approach can handle all common response variables, including binary (disease), count (symptom score), quantitative intermediate phenotypes (metabolite, hormone, protein, and gene expression), and even survival endpoints. It naturally allows for the inclusion of covariates representing environmental factors or adjustments for population stratification.

The core of the new method is to first estimate a GLM with three dummy variables according to the three possible genotypes but without an intercept. Linear contrasts were used in the second step to form the genetic models of interest, and the asymptotic distribution which is a multivariate normal distribution, was derived by the Delta method that is similar to the method by Gonza´lez et al [2008]. Since, the MAX-test is the maximum over a multivariate normal distribution so the model-specific confidence intervals can be derived using the multivariate normal distribution, and $p$-values adjusted for the testing of three tests at a single genetic marker were obtained. The procedure of Hothorn LA and Hothorn T [2009] was used for genetic model selection where the minimum of the $p$-values were used to decide which genetic model is the most plausible one.

A sample of $n$ independent subjects were considered with trait values $y_1, y_2, \cdots, y_n$ and the $q \times 1$ vector of covariates of an individual is denoted by $z_i$. Typical covariates may be environmental variables or adjustments for population stratification and the genetic marker of interest is a SNP, with alleles "$a$" and "$A$" where "$A$" is the risk allele. Three dummy variables were used to code the three genotypes with $x_{i,0}, x_{i,1}$ and $x_{i,2}$ corresponding to the genotypes $aa$, $Aa$ and $AA$, respectively, i.e. $x_{i,j} = 1$ if the subject $i$ has $j$ $(j = 0,1,2)$ alleles, and 0 otherwise. The considered GLM function was,

$$\mathrm{E}(y_i) = g\left(\beta_0 x_{i,0} + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \gamma' z_i\right) \tag{4.94}$$

where $g^{-1}$ is a link function, $\beta = (\beta_0, \beta_1, \beta_2)'$ are the genetic parameters of interest, and $\gamma$ is the $q \times 1$ vector of covariates. Model in the Eq. (4.94) does not include the regression constant. For a disease phenotype and a quantitative trait, the logistic link function and the

identity link function, respectively, are the natural link functions and, therefore, the standard choices. For survival endpoints, Eq. (4.94) was adapted to the Cox model.

All parameters of Eq. (4.94) were estimated by maximum likelihood estimation and are jointly asymptotically normally distributed. More specifically, the vector $\hat{\beta}$ is asymptotically normally distributed with mean $\beta$ and a $3 \times 3$ covariance matrix, $\Sigma = \text{var}\left((\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)'\right)$.

Specific genetic models can be obtained by introducing restrictions in terms of linear contrasts on the parameter vector $\beta$. For a standardized linear contrast, the sum over all coefficients needs to be 0, and the sum of all positive-valued coefficients should be 1. For the simplification, the additive model with penetrances 0, $\frac{1}{2}$, and 1 for carriers of 0, 1 and 2 "*A*" alleles was considered. The linear contrast $(-1,\ 0,\ +1)$ was obtained by subtracting $\frac{1}{2}$ from the penetrances and multiplying these values by 2. Similarly, $\left(-\frac{1}{2}, -\frac{1}{2},\ +1\right)$ and $\left(-1, +\frac{1}{2},\ +\frac{1}{2}\right)$ were the contrasts for the recessive and dominant models, respectively. Now, the new parameter vector was,

$$\hat{\xi} = \xi\begin{pmatrix}\beta\\ \gamma\end{pmatrix} = \begin{pmatrix} -\hat{\beta}_0 + \frac{1}{2}\hat{\beta}_1 + \frac{1}{2}\hat{\beta}_2 \\ -\hat{\beta}_0 + \hat{\beta}_2 \\ -\frac{1}{2}\hat{\beta}_0 - \frac{1}{2}\hat{\beta}_1 + \hat{\beta}_2 \\ \hat{\gamma} \end{pmatrix} \tag{4.95}$$

$\hat{\xi}$ is asymptotically normally distributed with mean $\xi\begin{pmatrix}\beta\\ \gamma\end{pmatrix}$ and the covariance matrix can be derived using the Delta method. The MAX-statistic is given by,

$$T = \max(T_A, T_D, T_R) = \max_{j=1}^{3} \frac{\left|\hat{\xi}_j - \xi_j^0\right|}{\sqrt{\left[\widehat{\Omega}\right]_{ij}}} \tag{4.96}$$

where $[\widehat{\Omega}]_{ij}$ denotes the variance estimate of $\hat{\xi}_j$. Thus, the MAX-test was obtained by taking the maximum of the first three components, or, phrased differently, by taking the maximum of the contrast tests for $T_A, T_D, T_R$, which were the test statistics corresponding to the additive, dominant and recessive modes of inheritance, respectively. The $H_0$ tested here was,

$$H_0: \xi_1^0 = \xi_2^0 = \xi_3^0 = 0 \tag{4.97}$$

Since, $\xi$ is asymptotically normal, $p$-values of the MAX-test statistic $T$, which consists of three components, can be obtained from a three-dimensional normal distribution. Specifically,

$$Pr(T \leq t) = Pr(\max(T_A, T_D, T_R) \leq t) \tag{4.98}$$

The three-dimensional normal distribution can also be used for the calculation of confidence intervals with or without adjustment for the joint analysis of three different modes of inheritance.

### *4.3.8 Estimation of p-values using the GLMs [Buzdugan et al., 2016]*

This study proposed a procedure in which all the SNPs were analyzed in a multiple GLM, and its use were shown for extremely high-dimensional datasets. The method yielded $p$-values for assessing significance of single SNPs or groups of SNPs while controlling for all other SNPs and the family wise error rate. Hence, the method tested whether or not a SNP carries any additional information about the phenotype beyond that available by all the other SNPs. This rules out spurious correlations between phenotypes and SNPs that can arise from marginal methods because the 'spuriously correlated' SNP merely happens to be correlated with the 'truly causal' SNP. In addition, the method offered a data driven approach to identifying and refining groups of SNPs that jointly contain informative signals about the phenotype. Due to use the hierarchical structure of the procedure, the number of computations can be massively reduced. If the final clusters or groups are relatively high up in the hierarchy of the cluster tree, a few hypothesis tests needed to compute relatively.

Let each of the $n$ samples is indexed with $i \in \{1,2,\cdots,n\}$ and a response variable $Y_i$ for the $i$-th sample point encodes the status of a phenotype of interest. For example, the binary status of a disease with $Y_i \in \{0,1\}$, the continuous value of a survival time with $Y_i \in \mathbb{R}^+$or the continuous degree of an exposure concentration with $Y_i \in \mathbb{R}$. The regressor $X_i$ is a $p \times 1$ vector which encodes the SNP profile for the $i$-th sample point: $X_{ij} \in \{0,1,2\}$ is the value of the $j$-th SNP for sample point $i$, taking three possible values corresponding to the number of minor alleles per person. The logistic regression model for the binary response is,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \sum_{j=1}^{q} \beta_j X_{i,j} \tag{4.99}$$

where $Y_i \sim \text{Bernouli}(\pi_i)$, $\pi_i = P(Y_i = 1|X_i, \beta) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$, $(i = 1,2,\cdots,n)$. Here, $\pi$ represents the probability of individual $i$ having a case status given its SNPs $X_i$ and there is no additional noise term where the stochastic nature of the model comes from the probability $\pi_i$.

The ultimate goal was to infer statistical significance of a single SNP or of a possibly large group of correlated SNPs for a given phenotype. More precisely, to obtain the corrected $p$-values for the multiple testing.

For a single SNP $j$,

$$H_{0,j} : \beta_j = 0 \quad \text{versus} \quad H_{a,j} : \beta_j \neq 0 \tag{4.1A}$$

or for a group $G \subseteq \{1,\cdots,q\}$ of SNPs,

$$\begin{gathered} H_{0,G} : \beta_j = 0 \quad \text{for all } j \in G \\ \text{versus } H_{a,G} : \text{at least for one } j \in G \text{ we have that } \beta_j \neq 0 \end{gathered} \tag{4.2A}$$

But, in the higher dimensional case, the conventional statistical inference procedures fail.

Recent progress based on new methods such as multiple sample splitting, has allowed to obtain statistical significance measures for regression parameters $\beta_j$ [Bühlmann, 2013; Meinshausen et al., 2009; Zhang CH & Zhang S, 2014] or groups thereof [Mandozzi & Bühlmann, 2015]. This study followed the Mandozzi & Bühlmann [2015] for splitting the sample. The entire statistical procedure is schematically summarized in Figure 7.

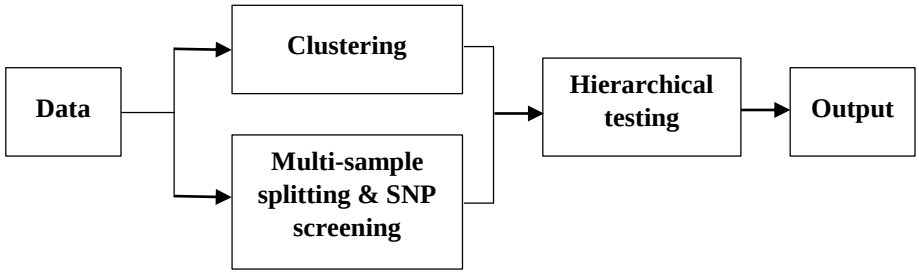



**Figure 7: Schematic overview of the sample splitting process used in the Buzdugan method.** "Clustering" refers to the step of hierarchically clustering the SNPs. SNPs on different chromosomes were clustered separately, after which the 22 clusters were joined into one final cluster containing all SNPs. "Multi-Sample Splitting and SNP Screening" stands for the SNP selection in steps 1 and 2 of the method described in construction of $p$-values section. These selected SNPs were used to compute the $p$-values. Finally, the last step of the method was hierarchical testing that used the selected SNPs to test groups of SNPs and eventually single SNPs. This testing was done hierarchically, on the cluster previously constructed. The output of the method consisted of significant groups, or single SNPs, along with their $p$-values, that were adjusted for multiple testing.

In view of the high-dimensional nature of GWAS, it is rather unlikely to detect single SNPs which are significant when controlling for all other SNPs. Thus, it is a priori more likely to detect (large) significant groups of SNPs with respect to the group hypotheses $H_{0,G}$ in a regression model.

### *Clustering*

Hierarchical clustering with average linkage [Jain & Dubes, 1988] was used which can be represented as a cluster tree, denoted by $\mathcal{T}$. This approach is similar to the construction of haplotype maps [Barrett et al., 2005]. A distance measure between SNPs was defined by $(1 - r^2)$, where $r^2$ refered to the pairwise $r$-squared between two SNPs [Hill & Robertson, 1968]. The measure of LD was taken within each chromosome separately and separate

cluster trees were constructed for each chromosome. Then these trees were combined into one tree $\mathcal{T}$ which contained all the SNPs in the study, as shown in Figure 8.

Let a cluster is denoted by generic letter $G$ that encodes a subset of $\{1, \cdots, q\}$ of single SNP.

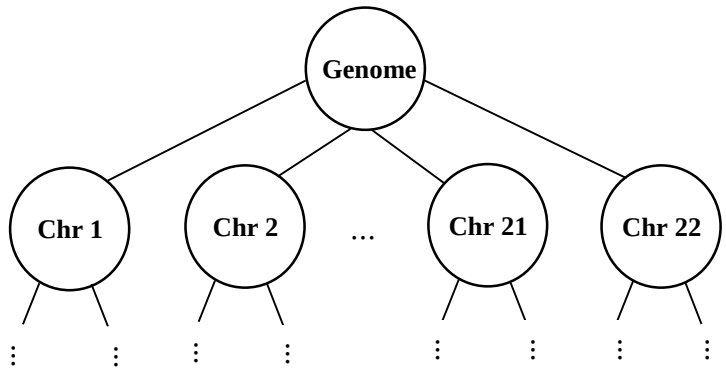



**Figure 8: The final cluster tree of the Buzdugan method.** The SNPs were first partitioned into chromosomes, and then a cluster tree was built for each chromosome separately using hierarchical clustering with average linkage.

***Hierarchical inference***

To test a $H_0$ for a group $H_{0,G}$ in Eq. (4.2A) or for a single SNP $H_{0,j}$ in Eq. (4.1A) the steps were as follows:

1. Test the global hypothesis $H_{0,G_{\text{global}}}$ where $G_{\text{global}} = \{1, \cdots, q\}$: that is, test whether all SNPs have corresponding (generalized) regression coefficients equal to zero or alternatively, whether there is at least one SNP which has a non-zero regression coefficient. If $H_{0,G_{\text{global}}}$ is rejected, the step 2 will be carried out.

2. Test the hypotheses $H_{0,G_1}, \cdots, H_{0,G_{22}}$ where $G_k$ contains all the SNPs on chromosome $k$. For those chromosomes $k$ where $H_{0,G_k}$ can be rejected, the next step is followed.

3. Test hierarchically the groups $G$ which correspond to chromosomes $k$ where $H_{0,G_k}$was previously rejected. First consider the largest groups and then proceed hierarchically (down the cluster tree) to smaller groups until a hypothesis $H_{0,G}$ cannot be rejected anymore or the level of single SNPs is reached.

4. The output was a collection of groups $G_{\text{final},1}, \cdots, G_{\text{final},m}$ where $G_{\text{final},k}$ was rejected $(k = 1, \cdots, m)$ and all subgroups of $G_{\text{final},k}$ $(k = 1, \cdots, m)$ downwards in the cluster tree were not significant anymore.

This is a hierarchical testing procedure belongs to the scheme of sequential multiple hypothesis testing where the multiple testing adjustment was resolution dependent. The hypothesis tests must be performed at different significance levels to control the familywise error rate.

***Construction of p-values***

Due to the high-dimensionality with $q \gg n$, obtaining a $p$-value for the hypotheses $H_{0,j}$ or $H_{0,G}$ in Eqs. (4.1A) or (4.2A) was a non-trivial problem. Based on the multiple sample splitting approach from Meinshausen et al. [2009], and following the method from Mandozzi & Bühlmann [2015], the idea was as follows. For $b = 1, \cdots, B$ repetitions:

1. The $n$ samples were randomly partitioned into two parts, say $N_{\text{in}}^{(b)}$ and $N_{\text{out}}^{(b)}$.

2. By a variable selection procedure viz. logistic or Lasso [Friedman et al., 2010; Tibshirani, 1996], the regressors (SNPs) based on data from the first half-sample $N_{\text{in}}^{(b)}$ were selected. The selected regressors were denoted by $\hat{S}^{(b)} \subseteq \{1, \cdots, q\}$. Because a Lasso estimated model has cardinality smaller or equal to $\min(n, q)$, the number of selected variables $\left|\hat{S}^{(b)}\right| < n/2$ will be smaller than half of the sample size, the first $n/6$ SNPs that enter the Lasso path were selected. This ensured that there were enough regressors for computing $p$-values.

3. Based on data from the second half-sample $N_{\text{out}}^{(b)}$, the classical $p$-value constructions was used in a linear or GLM with the selected SNPs from $\hat{S}^{(b)}$ in the step 2. The construction of a $p$-value of a cluster $G$ was done in the following manner: the hierarchy $\mathcal{T}$ constructed in the hierarchical clustering with $\hat{S}^{(b)}$ were intersected and an induced hierarchy with root node $\hat{S}^{(b)}$ was obtained. The testing was then applied on this induced hierarchy. Finally, the $p$-value to the entire cluster $G$ was assigned, although only the variables in $G \cap \hat{S}^{(b)}$ were used.

$$p^{G,(b)} = \begin{cases} p_{\text{out}}^{G \cap \hat{S}^{(b)}} & based\ on\ Y_{N_{out}^{(b)}}, X_{N_{out}^{(b)}}, if\ \ G \cap \hat{S}^{(b)} \neq \emptyset \\ 1, & if\ \ G \cap \hat{S}^{(b)} \neq \emptyset \end{cases} \tag{4.3A}$$

where $p_{\text{out}}^{G'}$ is the $p$-value for $H_{0,G'}$ based on data from $N_{\text{out}}^{(b)}$ $(G' \subseteq \{1, \cdots, q\})$. For a cluster $\in \mathcal{T}$ , the multiplicity adjusted $p$-value is,

$$p_{adj}^{G,(b)} = min\left(p^{G,(b)} \frac{\left|\hat{S}^{(b)}\right|}{\left|G \cap \hat{S}^{(b)}\right|}, 1\right) \tag{4.4A}$$

If $G \cap \hat{S}^{(b)} \neq \emptyset$ and $p_{\text{adj}}^{G,(b)} = 1$ otherwise.

Steps 1-3 were repeated for $B$ times and the $B$ $p$-values were aggregated. The aggregated $p$-value of any cluster $G$ was computed by considering its empirical quantile,

$$p^G = min\left\{1, \quad (1 - log\gamma_{min}) \inf_{\gamma \in (\gamma_{min}, 1)} \mathcal{Q}^G(\gamma)\right\} \tag{4.5A}$$

where $\mathcal{Q}^G(\gamma) = \min\left\{1, q_\gamma\left(\left\{p_{adj}^{G,(b)}/\gamma\,; b = 1, \cdots, B\right\}\right)\right\}$ , $\gamma \in (0,1)$ , $\gamma_{min} = 0.05$ and $q_\gamma(\cdot)$ is the empirical $\gamma$-quantile function. Finally, the hierarchically adjusted $p$-value of a cluster $G$ was,

$$p_{\text{ḥ}}^G = \max_{D \in \mathcal{T}: G \subseteq D} p^G \tag{4.6A}$$

The multi-sample splitting method was computationally fast since Lasso in step 2 was rather cheap to perform and step 3 requires classical $p$-value computations in low-dimensional models with fewer than $n$ regressors only. In terms of type I error control, this method has been found very reliable in extensive simulations relative to other methods.

## 4.4 Concluding remarks

When the underlying genetic model is unknown, the MAX-test has been shown to be robust against genetic model misspecification which takes the maximum of the trend tests over a family of scientifically plausible genetic models. In practice, if we know a genetic model, we should choose the trend test corresponding to that model. Also, if we know the scientifically plausible model is restricted to the genetic models between the recessive model and the dominant model, MAX-test should be used. On the other hand, if the genetic model outside of the above range is likely, viz. the over dominant model, we should consider Pearson's test [Zheng et al. 2009]. The MAX-statistic is very easy to calculate without complex computer programs, but, an extensive permutations are required to estimate the $p$-values specially when the $p$-values are too small. Hence, a number of modifications comes out from different researcher from different aspects. One of them is the derivation of the distribution of the MAX-test statistic that can solve the computational issue to a greater extent.

# Chapter 5

# Importance sampling and minimax tilting

## 5.1 Introduction

The importance sampling (IS) that have greater significance to improve the estimation accuracy in computer simulation studies along with the minimax tilting approach is presented in this chapter. Simulation is a very useful tool for computing integrals, because most integrals can be written as an expectation called Monte Carlo Integration. In this process, the quantile transform is done after drawing a sample from a distribution with the $cdf$. Using the IS, an alternative distribution is drawn whose support is concentrated in the truncation region. The minimax tilting method is useful for exact independently and identically distributed data simulation from the truncated multivariate normal distribution, which provides both a method for simulation and an efficient estimator to hitherto intractable Gaussian integrals. The derived estimator from the minimax exponential tilting has comparatively smaller estimation errors [Botev, 2017]. Geometrically, the contour lines of a MAX-test constitutes a convex polytope in a $df$-dimensional space where the $\alpha$-value is the probability defined by the TND, which can be estimated by the minimax tilting approach of the Botev's method [Botev, 2017].

The organization of the present chapter is as follows. The procedures about the exponential tilting and IS procedures along with applications in the case of integral approximation are contained in section 5.2. The minimax tilting approach is introduced in section 5.3. Finally, section 5.4 contains some concluding remarks.

## 5.2 Exponential tilting and importance sampling (IS)

Exponential tilting is a method to change the shape of a baseline density to incorporate additional information about the underlying random variable. Such additional information often comes in the form of moments [Metaxoglou, 2015]. The predictive density from a model plays the role of the baseline density. In Monte Carlo estimation, the exponential tilting or twisting is a distribution shifting technique commonly used in rare event simulation, and rejection and IS in particular. Particularly, it is the process to reduce the errors by improving the focus.

IS is a general technique for estimating properties of a particular distribution only generating the samples from a different distribution than the distribution of interest. This is a direct method to calculate integrals or expectations. In the computation of an expectation or other integral, the integrand has a very small value on a dominant fraction of the whole integration volume. If the points are chosen evenly in the integration volume, the small minority of the points close to the peak give the dominant contribution to the integral.

Let us consider the general evaluation of an integral of the form,

$$\mathrm{E}[h(X)] = \int h(x)f(x)dx \tag{5.1}$$

where $f(x)$ is a $pdf$ from where an infinite number of random variables can be generated. A sample of size $n$, $(x_1, x_2, \dots, x_n)$ is generated from the density $f(x)$ by the Monte Carlo method. The integral (5.1) is then approximated with,

$$\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(x_i) \tag{5.2}$$

If $X_i \sim F(X)$ where $F(X)$ is the $cdf$ of $f(x)$. For many problems, $F(X)$ is difficult to sample from and/or variance of $h(x)$ is large. In that case, a related and easily sampled $pdf$, $g(x)$ can be used to facilitate the sampling and to reduce the variance. The logic

underlying IS lies in a simple rearrangement of terms in the target integral and multiplying by 1,

$$\int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = \int h(x)w(x)g(x)dx \tag{5.3}$$

Here, $g(x)$ is another density function whose support is the same as that of $f(x)$. That is, the sample space corresponding to $f(x)$ is the same as the sample space corresponding to $g(x)$ (at least over the range of integration). $w(x)$ is called the importance function where a good importance function will be larger when the integrand is large and small otherwise.

### *5.2.1 Example: Improve the integral approximation*

Consider the function,

$$h(x) = 10e^{-2|x-5|} \tag{5.4}$$

Suppose, we want to calculate $E[h(X)]$ where $X \sim \text{Uniform}(0,1)$. That is, we want to compute the integral as,

$$\int_0^{10} e^{-2|x-5|}dx \tag{5.5}$$

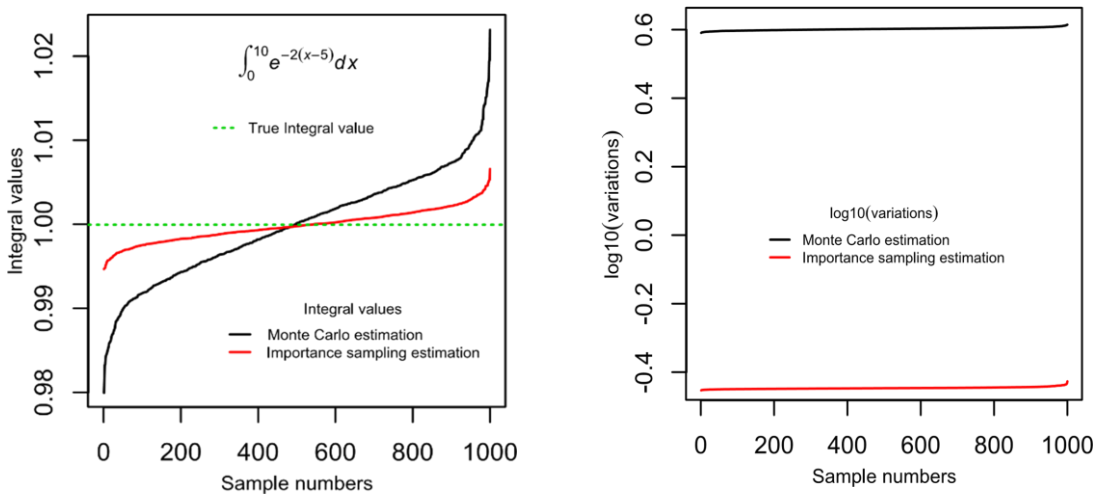
The true value of the integral is 1 (approx.). First, the simple way to do this is, generate $X_i$ from the uniform density within the range (0,10) and then compute the sample mean $h(x)$, which is equivalent to IS with importance function $w(x) = f(x)$. This Second, the integral (5.5) can be rewritten in terms of a Gaussian function as,

$$\int_0^{10} 10e^{-2|x-5|}\frac{1/10}{\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}}\frac{1}{\sqrt{2\pi}}e^{-(x-5)^2/2}dx \tag{5.6}$$

Eq. (5.6) defines the expectation as, $\mathrm{E}[h(x)w(x)]$, where $x \sim N(5,1)$. So, $f(x) = {}^{1}/_{10}$, $g(x)$ is the $N(5,1)$ density, and $w(x) = \frac{\sqrt{2\pi}\, e^{(x-5)^2/2}}{10}$ is the importance function. Hence, the integral (5.6) becomes,

$$\int_0^{10} e^{-2|x-5|} \sqrt{2\pi}\, e^{(x-5)^2/2}\, \frac{1}{\sqrt{2\pi}} e^{-(x-5)^2/2} dx \qquad (5.7)$$

In Eq. (5.7), the expectation is being calculated for the quantity, $e^{-2|x-5|} \sqrt{2\pi}\, e^{(x-5)^2/2}$ and $\frac{1}{\sqrt{2\pi}} e^{-(x-5)^2/2} dx$ is the density being integrated against $N(5,1)$. The first and the second method are the Monte Carlo and the IS estimations, respectively. The comparison of these two aforesaid methods is presented in Figure 9 where the estimated values of the integral (5.5) and the respective variations of the two methods are separately plotted for a repeated estimation of 1,000 times.




(A) (B)

**Figure 9: The estimation accuracy of IS and Monte Carlo sampling for estimating an integral.** (A) The plot of estimated integral values from the both methods in repeated samples. (B) The variations plot of the estimations in log-scale.

The Figure 9A and 9B are presenting the results of the estimation of Eq. (5.5) without and with IS. In Figure 9A, the two solid lines with black, red colors are presenting the estimates

from the simple Monte Carlo and IS procedures, respectively. The green dotted line in the same figure is presenting the true value of the integral. From the Figure 9, it was observed that the integral values with the IS are closer to the true value 1 (Figure 9A) having relative smaller variations as compared to that of without IS (Figure 9B). Particularly, the variations of IS are $1/10$ times of the simple Monte Carlo integral approximation (Figure 9B, the red line). Because, the function $h(x)$ in this case is peaked at 5, and decays quickly elsewhere. Therefore, under the uniform distribution, many of the points are contributing very little to this expectation whereas a Gaussian function ($ce^{-x^2}$) gives more precision with smaller variance at the same pick.

## 5.3 Minimax tilting

The goal of this method is to solve a minimax (saddle point) optimization problem to find an exponential tilting of a suitable IS. This is an efficient way of optimization because it exploits log-concavity properties of the normal distribution. The method constructs an unbiased estimator with a tight deterministic bound on its relative error and a concomitant exact stochastic confidence interval. Botev [2017] develops a new estimator, which is a modified version of the most celebrated Genz estimator [1992] using minimax tilting. The new estimator ensures that it enjoys theoretically better variance properties than the Genz estimator. Specifically, minimax tilting yields an estimator with the VRE property in an appropriate asymptotic tail regime [Kroese et al., 2011]. The Monte Carlo estimators rarely have the valuable VRE property [L'Ecuyer et al., 2010] within the light-tailed exponential family. Botev [2017] considers the multivariate normal law to develop this estimator, which can be easily generalized to other multivariate elliptic distributions for higher dimensions. The $d$-dimensional multivariate normal law under linear restrictions is,

$$f(\mathbf{z}) = \frac{1}{l} e^{\frac{1}{2}Z^T Z} \, \mathbb{I}\,(\mathbf{l} \le A\mathbf{z} \le \mathbf{u}) \tag{5.8}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathbf{z} = (z_1, \ldots, z_d)^T$, $\mathbf{u}, \mathbf{l} \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times d}$, $\text{rank}(A) = m \le d$ and $l = P(\mathbf{l} \le AZ \le \mathbf{u})$ is the probability that a random variable $Z \sim N(0,\ I_d)$ falls

in the $\mathcal{H}$-ploytope defined by the linear inequalities. The $l$ can be estimated by the simulation from the truncated multivariate normal density $f(\boldsymbol{z})$.

Let us define the following LQ-decomposition,

$$A = LQ^T \tag{5.9}$$

where $L$ is a $m \times d$ lower triangular matrix of non-negative entries in the main diagonal and $Q^T = Q^{-1}$ is a $d \times d$ orthogonal matrix. A simple change of the variable $\mathbf{x} \leftarrow Q^T\mathbf{z}$ yields,

$$l = Pr(\mathbf{l} \le AZ \le \mathbf{u}) = \int \phi(\mathbf{x}; 0, I)d\mathbf{x} \tag{5.10}$$

The bound of the integration (5.10) is, $\mathbf{l} \le A\mathbf{x} \le \mathbf{u}$, $\phi(\mathbf{x}; 0, I)$ is the $pdf$ of the $N(\mu, \Sigma)$ distribution. Here, $m = d$ is considered so that $A$ is a full rank matrix. The separation of variables (SOV) estimator of Genz [1992] can be defined from Eq. (5.9) with the bounds $\mathbf{u}, \mathbf{l}$, and the uniform sequence $U_k \overset{IID}{\sim} U(0,1)$, $k = 1,2, \ldots, d-1$. For each $k$, $X_k$'s are generated from a conditional simulation with,

$$\tilde{l}_k(X_1, \ldots, X_{k-1}) \le X_k \le \tilde{u}_k(X_1, \ldots, X_{k-1}) \tag{5.11}$$

and from the inverse transform method by setting,

$$X_k = \Phi^{-1}[\Phi(\tilde{l}_k) + U_k\{\Phi(\tilde{u}_k) - \Phi(\tilde{l}_k)\}] \tag{5.12}$$

where $\Phi(\cdot)$ is the cumulative density function from $N(0,1)$. The decomposition in Eq. (5.11) motivates the SOV estimator of $l$ as,

$$\hat{l} = \frac{\phi(\mathbf{X}; 0, I)}{g(\mathbf{X})}, \qquad \mathbf{X} \sim g(\mathbf{x}) \tag{5.13}$$

where $g$ is an IS density over the set, $\mathscr{b} = \{\mathbf{x}: \mathbf{l} \leq A\mathbf{x} \leq \mathbf{u}\}$ which defines the region where the decomposition of (5.11) was done by Genz [1992]. Then the SOV forms as,

$$g(\mathbf{x}) = g_1(x_1)g_2(x_2/x_1) \dots g_d(x_d/x_1, \dots, x_{d-1}), \qquad \mathbf{x} \in \mathscr{b} \tag{5.14}$$

Now, for all $k = 1,2, \dots, d$ the estimator is obtained as,

$$g_k(x_k/x_1, \dots, x_{k-1}) \propto \phi(x_k; 0,1)\, \mathbb{I}\left(\tilde{l}_k \leq x_k \leq \tilde{u}_k\right) \tag{5.15}$$

From here we get the sample estimate,

$$l^{\circ} = \prod_{k=1}^{d}[\Phi\{\tilde{u}_k(X_1, \dots, X_{k-1})\} - \Phi\{\tilde{l}_k(X_1, \dots, X_{k-1})\}] \tag{5.16}$$

Eq. (5.16) is estimated for $n$ times to obtain independently and identically distributed (IID) samples, $l_1^{\circ}, \dots, l_n^{\circ}$ that constitutes the unbiased point estimator,

$$\bar{l} = \frac{(l_1^{\circ} + \cdots + l_n^{\circ})}{n} \tag{5.17}$$

and 95% confidence interval (approx.), $(\bar{l} \pm 1.96S/\sqrt{n})$, where $S$ is the sample standard deviation of $l_1^{\circ}, \dots, l_n^{\circ}$. This SOV approach allows to simulate from $f(\mathbf{z})$ exactly by using an instrumental density $g(x)$ from the exponential tilting in the accept-reject simulation [Kroese et al., 2011]. For a given light-tailed $pdf$ $h(x)$ on $\mathbb{R}$, the exponentially tilted version is, $h_\mu(y) = e^{\{\mu y - K(\mu)\}}h(y)$, where $K(\mu) = \ln[\mathbb{E}\{e^{\mu X}\}] < \infty$ is the cumulant generating function for an open set $\mu$. The tilted version of Eq. (5.15) is,

$$g_k(x_k/x_1, \dots, x_{k-1}) = \frac{\phi(x_k; \mu_k, 1)\, \mathbb{I}\left(\tilde{l}_k \leq x_k \leq \tilde{u}_k\right)}{\Phi(\tilde{u}_k - \mu_k) - \Phi\left(\tilde{l}_k - \mu_k\right)} \tag{5.18}$$

With the simplification the estimator (5.13) becomes [Botev, 2017],

$$\hat{l} = e^{\phi(\mathbf{x};\, \mu)} \tag{5.19}$$

The minimax tilting is an optimization (saddle point) problem is defined as,

$$\inf_{\mu} \max_{\mathbf{x} \in \mathcal{B}} \psi(\mathbf{x}; \mu) \tag{5.20}$$

with a unique solution given by the concave optimization problem,

$$(\mathbf{x}^*, \mu^*) = \arg\max_{\mathbf{x}, \mu} \psi(\mathbf{x}; \mu) \tag{5.21}$$

The Eq. (5.21) minimizes with respect to $\mu$.

## 5.4 Concluding remarks

The minimax tilting is an effective method for exact simulation from the truncated multivariate normal distribution, which can also be used to estimate normal probability integrals. From the theoretical point of view, the method improves the performance of the existing estimators viz. SOV estimator and can achieve the most efficient VRE property for the estimation or approximations in the tail of a distribution. In the MAX-tests, the smallest $p$-values are the cutoffs or significance thresholds in the context of multiple testing. Such too small $p$-values can be considered as the rare events as these probabilities are defined at the tails of the respective distributions. The estimation of these probabilities can be done with the simulation studies whose accuracy depends on a huge number of permutations that are a highly computational issue. The idea of minimax tilting is an obvious solution here, which can be extended to other multivariate densities related to the normal distribution.

# Chapter 6

# Spherization

## 6.1 Introduction

Spherization combines the idea of algebraic geometry of a contingency table [Slavković & Fienberg, 2010] and $\chi^2$-statistic [Hirosawa et al., 2010; Yamada & Okada, 2009] to develop the $MAX_{st}$ from the spherical geometry. This is a linear algebraic transformation to handle multiway tables and their $df = 1$ tests in the corresponding $df$-dimensional space, in which the tables are in the standard $df$-variate normal distribution and the test vectors are the unit vectors. The method is based on vectorization of the table arrays, the Kronecker product of simplex-based rotation matrices representing the combinations of variables, and an eigenvalue decomposition. The distribution of the multiway tables is a multiple multinomial distribution for which the support is a simplex. The spherization lets the vertices of the simplex represent the values of each of the discrete variables, and simplex can then be expressed as a rotation matrix. The Kronecker product of these rotation matrices allows to handle multiway tables geometrically. This principle can be applied to transfer multiway tables that share same marginal counts to standard multivariate normal distributions in $df$-dimensional space.

The organization of the present chapter is as follows. The spherical geometry of the $\chi^2$-test statistic in the context of the proportion trend test is discussed in section 6.2. Section 6.3 presents the geometric features of convex polytopes in the transformed $df$-dimensional space. Finally, section 6.4 contains some concluding remarks.

## 6.2 Geometry of $\chi^2$-test

Let us consider the simplex-based and rotational geometry of the vectorization of a $m$-way contingency table Y, having number of categories of $I_j$ in $j$-th dimension as described in the section 2.5.2 in Chapter 2. From that discussion, it was concluded that tables sharing the same marginal counts can be located in the same $df$-dimenaional space. The similar notations as in section 2.5.2 will be considered in the following subsections to describe the proposed transformation method called spherization.

### *6.2.1 Dimension reduction of contingency tables*

For the $\chi^2$-test, the tables sharing same marginal counts can be located in the same $df \equiv \prod_j (I_j - 1)$ dimensional space (Figure 6) which is smaller than $M \equiv \prod_j I_j$ dimensional space because of the restriction of similar marginal counts.

The dimension reduction is performed by shifting in parallel and rotating them. First, the center of the distribution of tables is moved to the expected table and the differential vector is defined as,

$$d \equiv o - e \tag{6.1}$$

where $o$ and $e$ are the observed and expected frequency vectors for arbitrary tables in $M$-dimensional space, respectively.

Next, the Kronecker product of the rotation matrices, $R$ [Eq. (2.44)] is defined as described in section 2.5.3 in Chapter 2 depending on $I_j$, $j = 1, 2, \cdots, m$.

Rotation of the equation (6.1) by $R$ is the rotation of a simplex in the $I_j$ dimension so that the vertices of it are placed in a hyperplane, $[Rd]_{ji} = \frac{1}{\sqrt{n}}$, where $Rd$ is a rotated differential vector and $n$ is the number of categories in the $j$-th dimension. This rotation produces many zero elements ($|Rd|_0$) that are common for all the tables. Specifically, $|Rd|_0 \leq df$ where $|\cdot|_0$ is the number of non-zero elements. The positions of $|\cdot|_0$ for tables sharing same

marginal counts are consistent with each other and can be known from $I_j$s. Then an index vector $v$ is defined to specify the positions of $|Rd|_0$ and $|\cdot|_0$ elements,

$$v \equiv \otimes \prod_{j=1}^{m} u(I_j) \tag{6.2}$$

where $u$ is an $n$-dimensional vector, the last element of which is 0, and the others are 1.

A matrix $V$ is defined from $v$, which removes zero elements from $Rd$ by specifying the columns for which $[v]_i = 0$. For example, for a $2 \times 3$ table,

$$v = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ and } V = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Multiplying $Rd$ by $V$, the dimension of the vectors of tables are reduced without loss of information.

### *6.2.2 Spherization of the $\chi^2$-statistic*

Using vectorized form of Y, the $\chi^2$-statistic in $df$-dimensional space is,

$$\chi^2_{df} = d^T E^{-1} d \tag{6.3}$$

where $E$ is a diagonal matrix having $e$ in the main diagonals.

For dimension reduction the equation (6.3) can be written as,

$$\chi^2_{df} = d^T R^T R E^{-1} R^T R \, d \tag{6.4}$$

$$= d^T R^T V^T V \, R \, E^{-1} R^T V^T V R d \tag{6.5}$$

where $VRd$ is a $df$-dimensional vector, and $V\ R\ E^{-1}R^TV^T$ is a $(df \times df)$-dimensional symmetric matrix.

Using LU decomposition, Eq. (6.5) can be transformed into,

$$\chi^2_{df} = d^T R^T V^T X^T XVR\ d \tag{6.6}$$

$$= g^T g \tag{6.7}$$

where $X$ is an LU decomposed matrix of $V\ R\ E^{-1}R^TV^T$ and $g = XVR\ d$.

Hence, Eq. (6.7) is a squared Euclidean distance from the center to the rotated hyperplane in $df$-dimensional space and $g$ is assumed to be a sample from $df$-dimensional multivariate normal distribution.

### *6.2.3 Geometry of the proportion trend test*

The geometry of the $\chi^2$-statistic can be applied to deploy the proportion trend test in $df$-dimensional space. Let us consider a $2 \times 3$ contingency table for the description. To define the proportion trend test, a weight vector has to be defined. For example, (2 1 0) is the weight vector, where the weights are defined for each column. For generalization, the weights are defined for each cell of the contingency table as, $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. For the single weight vector $w$ or one test, the test statistic $\chi_w$ is the function of a weighted sum of the differential table,

$$\chi_w = f(w^T d) \tag{6.8}$$

To deploy Eq. (6.8) in $df$-dimensional space, we use the deformation of the definition of $g$ in Eq. (6.7),

$$g \equiv \mathcal{P}d \tag{6.9}$$
$$= XVRd \tag{6.10}$$
$$VRd = X^{-1}g \tag{6.11}$$
$$V^T VRd = V^T X^{-1} g \tag{6.12}$$
$$Rd = V^T X^{-1} g \tag{6.13}$$
$$d = R^T V^T X^{-1} g \tag{6.14}$$
$$= \mathcal{P}^{\dagger} g \tag{6.15}$$

where $\mathcal{P}^{\dagger} = R^T V^T X^{-1}$ is a sparse version of the pseudo-inverse of $\mathcal{P}$.

The deformation from the Eq. (6.12) to Eq. (6.13) only holds for cases in which the dimension reduction to $df$-dimensional space of spherization is applicable. Then, from Eq. (6.8),

$$\chi_w = f(w^T \mathcal{P}^{\dagger} g) \tag{6.16}$$

The test vector defined by the weight vector and rotation is,

$$\tau^T = \frac{w^T \mathcal{P}^{\dagger}}{|w^T \mathcal{P}^{\dagger}|} \tag{6.17}$$

Therefore, the proportion trend test is projected in $df$-dimensional space with a direction of the inner relation of the test as,

$$\tau^T g = |g| cos\theta \tag{6.18}$$

Eq. (6.18) defines the inner product, which is the relation between observations and weight, where $\theta$ is the angle between observation and weight. If the observation is in the direction of $\theta = 0$ so that $(\tau^T g)^2 = |g|^2 = \chi^2_{df}$, then the $\chi^2$-statistic for a vectorized weight is,

$$\chi^2_w = |\tau^T g|^2 \tag{6.19}$$

where $\chi_w^2$ follows a $\chi_1^2$-distribution, and the $p$-value is the proportion of tables having $\chi^2$-values greater than or equal to the observations per the number of all possible tables.

For one observation and one weight, one test vector is deployed into $df$-dimensional space. For one observation and multiple weights, the multiple test vectors are deployed into the same $df$-dimensional space as in the case of the $\text{MAX}_{\text{st}}$.

We can draw pairs of truncating hyperplanes in $df$-dimensional space depending on a set of test vectors and $\chi_w$. The hyperplanes are defined as vertical to the test vector, and the difference from the origin to the hyperplane is $\chi_w$. Because the distribution of tables is transformed into a standard multivariate normal distribution in spherized space, the $p$-value is the cumulative standard normal probability out of the convex polytope, truncated by the pair of hyperplanes.

## 6.3 Geometric features of convex polytopes in spherized space

The geometric configuration of the convex polytope defines the multiple testing $p$-values. Figure 10A shows the case of $df = 2$ with $df = 1$ test for a single test.

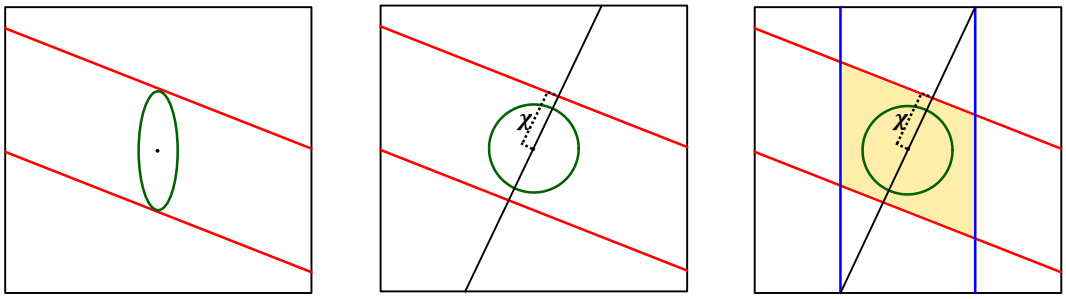



**Figure 10: The spherical geometry of tests in $df$-dimensional space.** (A) The contour hyperplanes of the distribution of tables with the same $\chi^2$-values for a single test where ellipse for the $df = 2$ test and a pair of parallel hyperplanes for the $df = 1$ test, respectively. (B) and (C) present the spherization in $df$-dimensional space. (B) The single test: The transformation of parallel hyperplanes into tangent hyperplanes and ellipse into sphere (circle). The black line crossing

the center indicates the direction of the test. The $\chi^2$-value is the squared Euclidean distance from the center to the hyperplane having distance $\chi$. (C) Multiple testing: Each pair of parallel hyperplanes of the same color presents a contour hyperplane of the tables having the same $\chi^2$-values from the individual $df = 1$ test. Multiple pairs of these parallel hyperplanes form a convex polytope (light-yellow colored area). The spherical space defined by the convex polytope produces the TND. The $p$-value is the integral of the probability defined by the TND out of the polytope.

All the possible tables sharing the same marginal counts with an observed table and having non-negative cell counts can be realized as the points on a $df$-dimensional simplex in the $df$-dimensional space (Figure 6). Hence, tables with the same $\chi^2$-values are drawn with a high dimensional ellipsoid contour for the $df = 2$ test and a pair of truncating parallel hyperplanes for the $df = 1$ test. The ellipsoid contour is not easy to handle from the distributional standpoint since all of the tables on the ellipsoid are not equidistant from the origin (Figure 10A). These parallel hyperplanes can be transformed into tangent hyperplanes and a high-dimensional ellipse into an $n$-sphere (generalization of the ordinary sphere into spaces of $df$-dimension) by spherization, in which all of the tables are placed on the transformed $df$-dimensional $n$-sphere. The $\chi^2$-value is calculated as the squared Euclidean distance from the expected table to the hyperplane with distance $\chi$ and the $p$-value is the proportion of tables outside of the convex polytope (Figure 10B). The definition of the hyperplane for a single test is,

$$-\chi_w \leq \tau^T Z \leq \chi_w \tag{6.20}$$

where $Z$ is the axis of $df$-dimensional space.

In multiple testing, multiple pairs of parallel hyperplanes form a convex polytope (Figure 10C, light yellow-colored area), in which each pair corresponds to the tables with same $\chi^2$-values from the individual $df = 1$ test. The definition of the convex polytope in equation (6.20) is generalized for $n_{test}$ as,

$$-\chi_{max} 1_{n_{test}} \leq AZ \leq \chi_{max} 1_{n_{test}} \tag{6.21}$$

where $\chi_{max} \equiv \max_k \chi_{w_k}$; $k = 1,2, \ldots, n_{test}$, $A$ is a $(n_{test} \times df)$ matrix with rows that are test vectors, $\chi_{max}$ is the largest value of the test statistics among $n_{test}$, and $1_{n_{test}}$ is a $(n_{test} \times 1)$ vector, all of the elements of which are 1.

If the maximum value among a set of test statistic values is representative in cases of multiple testing, the tables on the surface of the convex polytope have the same representative $\chi^2$-values. Hence, the probability of observing tables with representative $\chi^2$-values greater than or equal to the observed table's $\chi^2$-value is the integral of the $pdf$ of a multivariate normal distribution over the space out of the convex polytope. Therefore, the $p$-value for $n_{test}$ is,

$$p = 1 - \int N(Z|0, I)\ \delta\left(-\chi_{max} 1_{n_{test}} \leq AZ \leq \chi_{max} 1_{n_{test}}\right) dZ \tag{6.22}$$

where $\delta$ is the delta function.

Tables in spherized space are considered to be observed in standard normal distributions as having a center corresponding to the expected table in the context of $H_0$ testing. Because of the relationship between $n$-dimensional normal distributions and $\chi^2$-distributions, we define $\chi$ as the distance from the center to the point corresponding to the observed table, and we also obtain the spherized space (Figure 10B, circle) with a multivariate normal distribution demarcated by a convex polytope, producing TND (Figure 10C).

## 6.4 Concluding remarks

In this chapter, we introduced a geometric transformation to handle the space of the distribution defined by the contour lines of tables and tests. The proposed method is based on the algebraic geometry of rotation and scaling of the multiway contingency tables that rotates the tables to a $df = 1$ dimensional space where the distribution of the tables follows a standard $df$-variate normal distribution in the transformed space. As the multivariate normal distribution is symmetric in all the directions, the transformed space also symmetric

and that provides the simplest way to handle multiple $df = 1$ tests mutually in higher dimensions.

# Chapter 7

# Methodology

## 7.1 Introduction

This chapter aims to present the proposed Sph-Btv approach and its evaluation steps with a real GWAS data. The new method is a combination of spherization (Chapter 6) and the TND estimation technique [Botev, 2017] based on the minimax tilting optimization process in IS for estimating the cumulative probabilities from TND. The Sph-Btv approach first uses spherization to handle the space of the distribution defined from the table's and the test's contour lines and finally provides a reformed symmetric space having $df$-variate TND. Next, the probability defined by the multivariate TND is the $p$-value of multiple testing for a given SNP subset, which is estimated by the Botev [2017] method. The method was evaluated by a real GWAS data from the Nagahama Study. The SNP subsets located in the gene loci over the genome were considered for the evaluation. The $p$-values for each of the SNP subsets were estimated for the multiple cutoff $\chi^2$-values, including the common GWAS cutoff of $p_{sig} = 5 \times 10^{-08}$.

The organization of the present chapter is as follows. Section 7.2 derives the inner product matrix of the Sph-Btv approach and the detailed derivation of the proposed method is given in section 7.3. The description of the real genotype data is introduced in section 7.4. Section 7.5 contains the experimental design of the evaluation process of the new method using real genotype data. The source of the data set and the R-codes used in this research are given in section 7.6. Finally, section 7.7 contains some concluding remarks.

## 7.2 The inner product matrix of the spherization and Botev's approach (Sph-Btv)

The probability of a $df$-dimensional vector $Z$ falls outside of a convex polytope can be defined from the linear restriction of $d$-dimensional multivariate normal law [Botev, 2017] as,

$$1 - Pr(l \leq AZ \leq u) = 1 - \int_{l}^{u} \phi(Z; 0,\ I_d)\, dZ \tag{7.1}$$

where $l$ and $u$ are the lower and upper truncation limits, respectively; $A$ is a full rank matrix; $Z \sim N(0,\ I_d)$ having a $d$-dimensional multivariate normal distribution with mean vector 0 and variance-covariance matrix $I_d$; and $\phi(Z; 0,\ I_d)$ is the $pdf$ of $N(0,\ I_d)$.

Since $A$ is a full rank matrix, and $Z \sim N(\mu = 0,\ \Sigma_Z = I_d)$, we can simply estimate Eq. (7.1) as [Botev, 2017],

$$1 - Pr(l - A\mu \leq A\mathbb{Z} \leq u - A\mu) \tag{7.2}$$

$$= 1 - Pr(l \leq A\mathbb{Z} \leq u) \tag{7.3}$$

$$= 1 - \int_{l}^{u} \phi(\mathbb{Z}; 0,\ A\ \Sigma_Z\ A^T)\, d\mathbb{Z} \tag{7.4}$$

where $\mathbb{Z} \sim N(0,\ A\ \Sigma_Z\ A^T)$, $\Delta_{n_{test} \times n_{test}} = A\ \Sigma_Z\ A^T = A\ A^T$ is an inner product matrix.

## 7.3 The Sph-Btv approach

In our approach, the $p$-values were estimated by Botev's approach [Botev, 2017] after the transformation of the space with spherization. The definition of the convex polytope for $n_{test}$ can be defined from the Eq. (7.3),

$$l \leq A\mathbb{Z} \leq u \tag{7.5}$$

where $A_{n_{test}\times df}$ is the $df$-dimensional matrix, the rows of which are the test vectors defined by the test models ($W_{n_{test}\times k}$) and rotation matrix ($\mathcal{R}_{k\times df}$) of the table; $k$ is the number of column categories of the table. $\mathbb{Z}\sim N(0,\ \Delta)$ is a $df$-dimensional multivariate normal vector; and $l$ and $u$ are two ($n_{test} \times 1$) vectors of the intercepts of tangent contour hyperplanes of tests.

Since our method's test vector is a unit vector, and we define the distance from the origin to the hyperplane as $\chi$, where $\chi$ is the maximum value of the test statistics among $n_{test}$ from the proportion trend test. Therefore, the elements of vectors $l$ and $u$ are $-\chi$ and $\chi$, respectively. The $p$-value for the Sph-Btv approach is the probability that the $df$-dimensional vector $\mathbb{Z}$ falls outside of a convex polytope defined by the linear inequalities in Eq. (7.5),

$$\alpha = 1 - Pr\left(l_\chi \leq A\mathbb{Z} \leq u_\chi\right) = 1 - \int_{l_\chi}^{u_\chi} \phi(\mathbb{Z};\ 0,\ \Delta)\ d\mathbb{Z} \tag{7.6}$$

where $l_\chi = -\chi_{max}1_{n_{test}}$ and $u_\chi = \chi_{max}1_{n_{test}}$ are the lower and upper truncation limits, respectively.

The $\Delta$ in Eq. (7.6) is an inner product matrix of the Sph-Btv approach, which is a positive definite variance-covariance matrix. The upper or lower triangle of $\Delta$ consists of $q$ ($= n_{test}(n_{test}-1)/2$) elements, each of which is the pairwise inner product of two test vectors. These $q$-elements determine the size of the convex polytope, and they consist of full information of the $pdf$ of $MAX_{st}$ and provide information about $p$ by considering the gene-wise LD structures.

## 7.4 Real genotype data

The genotype data from the Nagahama Study (Yoshimura et al. 2012; Higasa et al. 2016) of 1,813 samples having 996,339 SNPs were considered. The phenotypes were randomly

generated using equal probabilities for cases (= 907) and controls (= 906). The real phenotype data were not used in this study because, we were interested to extract the features of our method with a GWAS data considering the character of its LD structures per gene, where the case-control-wise null hypotheses throughout. The SNP subsets ($N$ subsets) located at each of 14,941 gene loci were considered to calculate the gene-based multiple testing corrected $p$-values.

### *7.4.1 Genotyping*

A total of 1,813 individuals were genotyped with a HumanOmni2.5 BeadChip Kit (1,761 individuals) or HumanOmni5M-exome BeadChip (52 individuals) from Illumina Inc.

### *7.4.2 Quality control*

The quality control of raw genotype data was performed on each array as: samples were excluded with call-rates $< 0.99$, and SNPs were excluded with minor allele frequency $< 0.01$, Hardy–Weinberg equilibrium testing $p$-value $< 1.0 \times 10^{-6}$ or call-rate $< 0.99$. In the results, 1,150,672 SNPs and 1,656,875 SNPs remained for the HumanOmni2.5 and HumanOmni5M-exomes, respectively. No samples were excluded from either array. We then merged the genotypes and retained 1,045,731 SNPs that remained in both arrays for the subsequent imputation procedure.

### *7.4.3 Imputation*

The quality-controlled genotype data as a template were compared with the reference genotype data set, the 1000 Genome p3v5 having 2,504 samples, of which 104 samples were Japanese. However, the low-quality markers were excluded from the quality control, but a screening was performed before haplotyping to exclude the SNPs with highly deviated allele frequencies from the reference genotype data. We excluded 30,617 SNPs with allele frequencies is 0.2 greater or less than the 1000 Genome East Asian population. As a result, genotypes of the 1,813 samples and 1,015,114 markers were used for the imputation. The ShapeIT2 [Delaneau et al., 2013] and Minimac3 [Das et al., 2016] methods were used for haplotyping and imputation, respectively. A total of 48,935,186M SNPs were obtained after

imputation. We selected the only genotyped markers having $\hat{r}^2 > 0.8$ as the imputation quality.

## 7.5 Experimental design

### *7.5.1 Characterization of SNP subsets*

The real genotype data consisted of SNP subsets from a total of 14,941 genes. A total of 1,229 genes were excluded from the screening of subsets for which the covariance was singular, according to the estimation procedure of the Botev method [Botev, 2017]. Among 13,712 genes, 1,590 genes had 1 SNP, and 129 genes had identical SNP pairs, i.e., all pairwise $r^2 = 1$ having the essential SNP number of 1. The conventional $\chi^2$-test was performed to evaluate these 1,719 genes with 1 SNP, and 11,993 genes were evaluated by the Botev method. Various $\chi$-values were selected from (0.01, 8.5) by increasing by 0.1 along with the GWAS cutoff of 5.45, in which the $\chi^2$=29.72, and $p_{sig} = 5 \times 10^{-08}$ for evaluating the $\alpha$-values of the two genes *FES* and *NCS1*. Among them eight representatives $\chi$-values: 1.41, 2.51, 3.91, 5.45, 5.91, 6.71, 7.51, 8.31 with approximately evenly spaced were selected for evaluating the overall heterogeneity for all genes. The $p$-values were calculated of each gene for multiple cutoff $\chi^2$-values. For each subset, the calculation was repeated 1,000 times, and 1,000 Monte Carlo samples were used for each repetition.

## 7.6 Data availability

The genotype data from the Nagahama Study reported in this paper is available in the DNA Data Bank of Japan (DDBJ) via Japanese Genotype-phenotype Archive (JGA) under the accession number JGAS00000000012 (https://ddbj.nig.ac.jp/jga/viewer/view/study/JGAS00000000012). The R-codes to calculate $p$-values by the Sph-Btv method are available free on-line via a GitHub repository: https://github.com/statgenetJimu/Spherization/blob/master/pvalueCalculation_Botev.R. The algorithm and the codes are also provided in the Appendix 4 and Appendix 5, respectively.

## 7.7 Concluding remarks

This chapter provides the mathematical formulation of the proposed Sph-Btv method and the experimental framework of the result evaluation. The new method combines the transformation technique of spherization and the estimation process of TND probabilities using the optimization twisting for the distribution shifting in case of too small probability values. The new estimator is expected to have a relatively small variance of estimation.

# Chapter 8

# Results and discussion

## 8.1 Introduction

This chapter presents the evaluation results of the proposed method. For a real GWAS data (Chapter 7), the method was applied for the SNP subsets located in the gene loci over the genome. The $p$-values for each of the SNP subsets were estimated for the multiple cutoff $\chi^2$ -values, including the common GWAS cutoff of $p_{sig} = 5 \times 10^{-08}$. A detailed discussion and interpretation of the obtained outputs are also given here.

The organization of the present chapter is as follows. The results from the method evaluation for a real GWAS data about the heterogeneity of $p$-values for various cutoffs are presented in section 8.2. The detail discussion and the interpretation of the results of the analysis is provided in section 8.3. Finally, section 8.4 contains some concluding remarks.

## 8.2 Heterogeneity of $p$-values at various cutoff in a GWAS data set

To evaluate the gene-based heterogeneity using our proposed method, we considered the SNP subsets located in the gene locus over the genome. The effect of the number of SNPs per gene and the individual LD structures in $p$-values is presented in Figure 11.

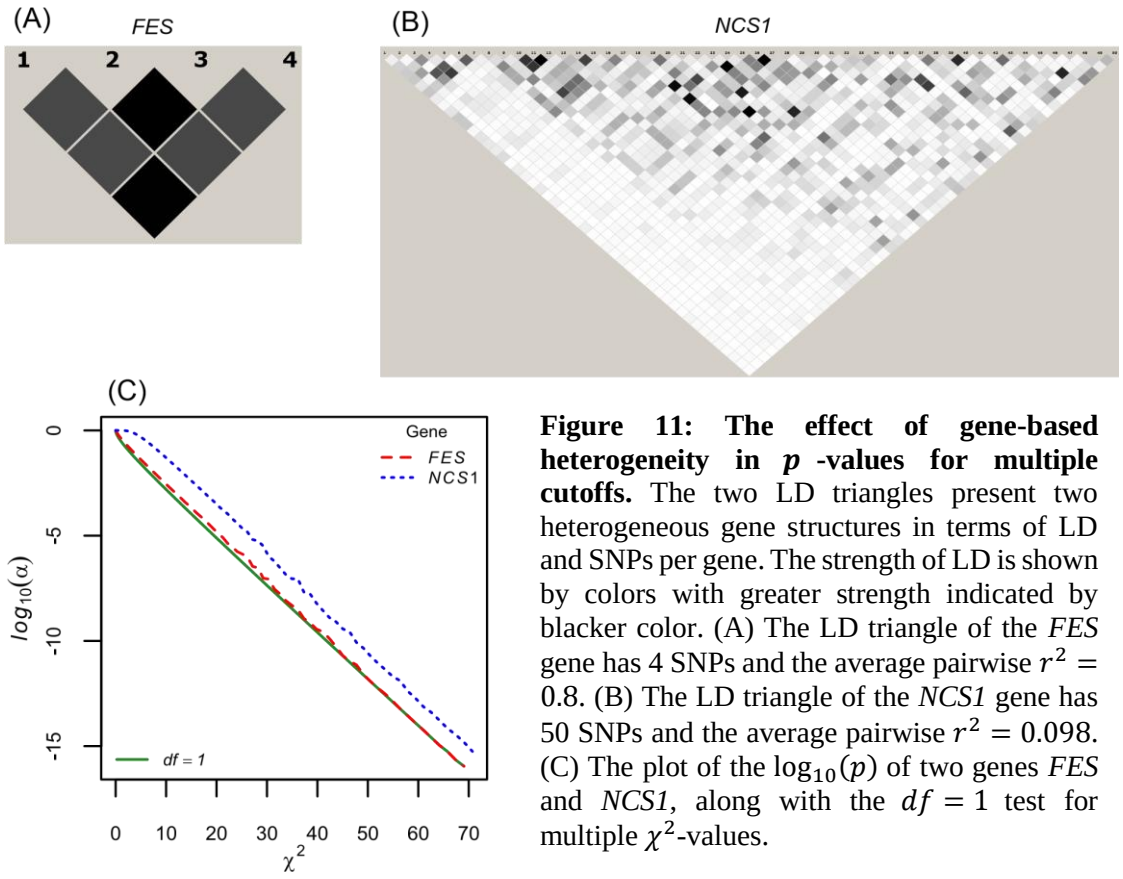



**Figure 11: The effect of gene-based heterogeneity in $p$ -values for multiple cutoffs.** The two LD triangles present two heterogeneous gene structures in terms of LD and SNPs per gene. The strength of LD is shown by colors with greater strength indicated by blacker color. (A) The LD triangle of the *FES* gene has 4 SNPs and the average pairwise $r^2 = 0.8$. (B) The LD triangle of the *NCS1* gene has 50 SNPs and the average pairwise $r^2 = 0.098$. (C) The plot of the $\log_{10}(p)$ of two genes *FES* and *NCS1*, along with the $df = 1$ test for multiple $\chi^2$-values.

For example, the gene *FES* in Figure 11A has relatively small number of SNPs with quite strong pairwise LD structures, compared to the gene *NCS1* in Figure 11B. These dissimilar physical features of the two genes lead them to produce quite heterogeneous values of $p$ and strongly motivate the values to move either toward or away from the $df = 1$ testing (Figure 11C). The $p$-values are highly deviated from the $df = 1$ test for the gene *NCS1*, whereas the gene *FES*, is closer to the values of $df = 1$, particularly the $p$-values are almost an exact approximation for the higher $\chi^2$-values (Figure 11C, red dotted line at the rightmost end). This variation in $p$-values is an important indicator about the variation in type I error rate per gene at a uniform confidence level, which may affect the false detection of a gene. For example, if we set the GWAS cutoff of $p_{sig} = 5 \times 10^{-08}$ ($\chi^2 = 5.45^2$) for all the $\chi^2$-values along the horizontal axis of Figure 11C, the type I error rate for a gene becomes significantly greater as we move from the left-hand corner to the right-hand corner

along the horizontal line. Moreover, SNPs per gene and genome-wide gene's LD structures are not uniform. The genes with more SNPs and weak LD structures tend to produce higher $p$-values than genes with relatively smaller numbers of SNPs and stronger LD (Figure 11). So, if we set the uniform cutoff to calculate the type I error rate for all the genes over the genome, then it is very likely that the type I error rate of the gene in Figure 11A will be higher than that of the gene in Figure 11B.

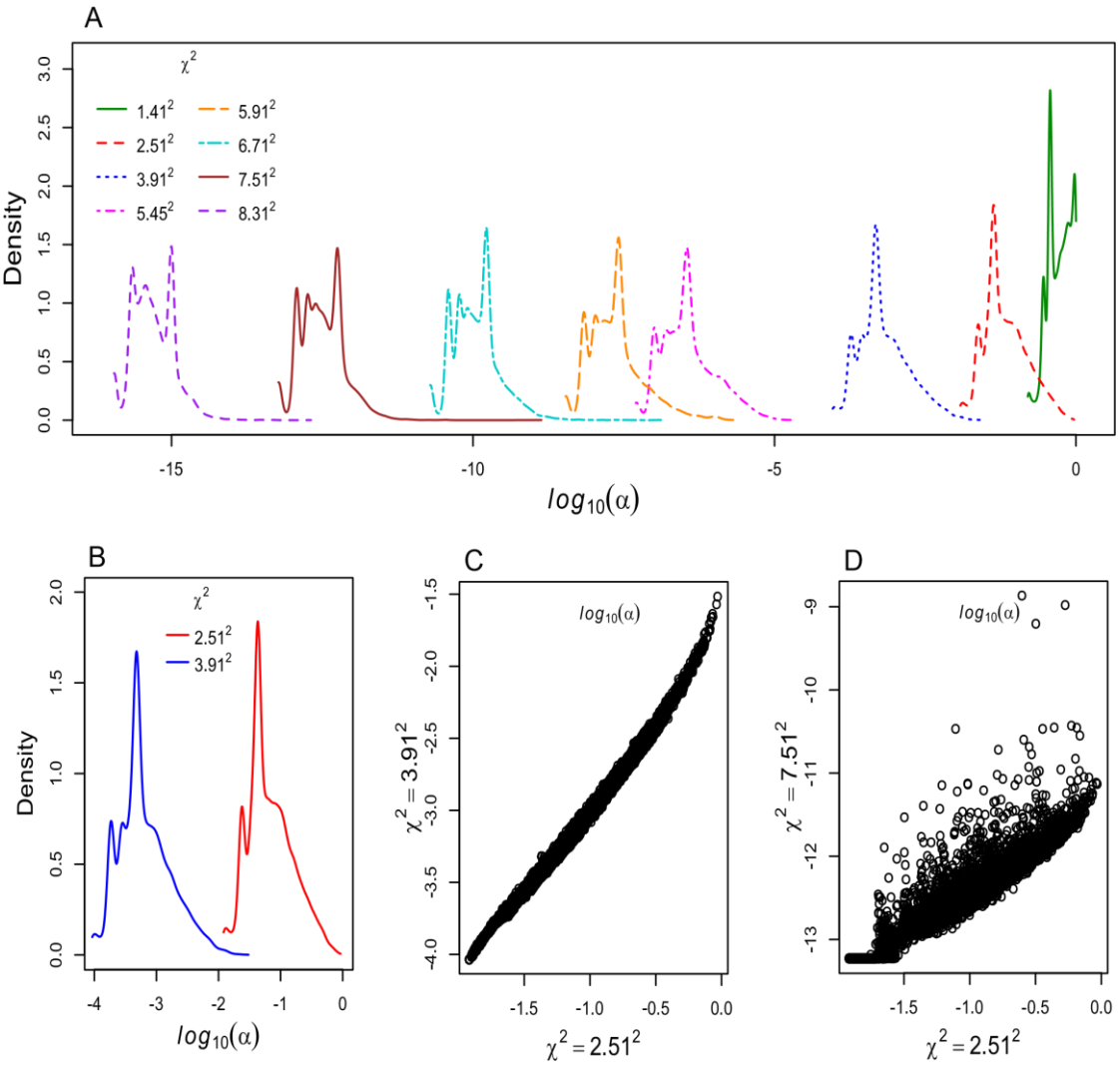



**Figure 12. The overall heterogeneity of $p$-values for multiple cutoffs.** (A) The distributions of $\log_{10}(p)$ for multiple cutoffs, each different colored subpanel of which present the distribution of every different $\chi^2$-cutoff. (B) The distributions of $\log_{10}(p)$-values for two cutoff $\chi^2$-values, $2.51^2$ and $3.91^2$, respectively. The horizontal and vertical axes of (A) and (B) present $\log_{10}(p)$ and corresponding density values, respectively. The scatter plots of $\log_{10}(p)$ for $\chi^2$-cutoffs, (C) $2.51^2$ and $3.91^2$. (D) $2.51^2$ and $7.51^2$.

Figure 12A is presenting the results from the evaluation of all the gene-based SNP subsets ($N$ subsets) for multiple cutoff $\chi^2$-values, including the GWAS cutoff. The eight colored subpanels in this figure depict the distributions of $log_{10}(p)$-values for eight different cutoffs. From here, it was observable that there was a shift of $p$-values for every change in $\chi^2$. The shape of the distributions and their modality at the left tails for each subpanel are changing along with the changes in cutoffs. For example, the shape of the distribution for $\chi^2 = 5.45^2$ (magenta) is very dissimilar from that of $\chi^2 = 8.31^2$ (purple). The evenly spaced distributions were observed along the horizontal log scale, indicating that the distributions are shifting with respect to $p$, keeping the width of the spaces almost the same. The Figure 12A demonstrates how the distribution function shapes vary with multiple cutoff $\chi^2$-values and the changes in the shape is not simple. Therefore, it is advisable to apply our method for individual gene sets by considering the corresponding LD structures and for target $\chi^2$-values individually.

To visualize the variations in $p$-values more obviously, only the two cutoff $\chi^2$-values of $2.51^2$ and $3.91^2$ were considered. The two distributions of $log_{10}(p)$-values of SNP subsets are presented in Figure 12B. The shapes of the distributions are almost identical, but a modality change was observed in the left tail. The two evenly spaced distributions in the horizontal axis with an almost total shift in $p$-values indicate the fold changes in $p$ with a change in the cutoff. For a typical investigation of heterogeneity in $p$-values, we considered the evaluation of three cutoff $\chi^2$-values: $2.51^2$, $3.91^2$ and $7.51^2$, respectively. The scatter plot between $log_{10}(p)$-values for $\chi^2 = 2.51^2$ with the other two cutoffs is presented in Figures 12C and 12D, respectively. From both figures, it was observed that the $p$-values are positively correlated having quite different strength and patterns for a change in the cutoffs. The values in Figure 12C are linearly related to a quite strong, positive correlation ($r^2 = 0.98$), whereas the strength of the relation is quite weak ($r^2 = 0.07$) in Figure 12D.

These results show that the $p$-values are heterogeneous for the gene-based SNP subsets over the genome for multiple cutoff $\chi^2$-values. Also, they are not uniform for a fixed cutoff. For example, the shape of each colored subpanel of Figure 12A is far from the shape of uniform distribution, indicating that some genes tend to produce type I error more

frequently than others based on their individual characteristics, such as LD pattern and number of SNPs per gene.

The $p$-value correction by our method for gene-subsets is reasonably simple and straightforward. But, the biological interpretation is not so simple, because the number of truly associated genetic variants in one particular gene is not always one, and its association structure between multiple genetic variants and phenotypes is believed complicated (Lamparter et al. 2016). In this study, the features of the proposed method were extracted only for the gene-based SNP subsets of a real GWAS data. This is the limitation of our study to apply these gene specific extracted features for the other genetic studies as the gene-coding regions comprise a small proportion of the human genome.

## 8.3 Discussion

This is a geometric approach of multiple testing for contingency tables such as MAX-test of multiple genetic models (additive, recessive, dominant). It seems one of the realistic approaches to know the variations of the type I error among subsets, each of which is considered of multiple tests but their dependency is heterogeneous. Our results suggested that the type I error may vary based on the individual structure of genes. Therefore, our proposed method offers the estimates of the probability density of MAX-test $p$-values. Because, there can be the functional relation among variants in each gene, it is advisable not to correct simple correction based on our method estimate without considering other potential factors. However, still our method will give meaningful information for GWAS interpretation.

We illustrated an example from the genetics using a real GWAS data for the SNP subsets per gene. But, because of the big data era, the simple positive signal detection by multiple testing seems to exist in various fields. This method can be applicable to other arbitrary fields, as far as the positive signal detection is similar. For example, for large categorical data as well as non-categorical data set, where multiple $\mathrm{df} = 1$ tests are repeated according to the procedure of the MAX-based test.

Our spherization transformation is based on the algebraic geometry of correspondence analysis (CA) (Greenacre 1984; Cuadras CM and Cuadras D 2006) with the squared Euclidean distance as a measure of $\chi^2$ in df-dimensional space. This $\chi^2$-distance is a standardized form of the Mahalanobis distance (Mahalanobis 1936). A detailed illustration of the relationship among the CA, Mahalanobis distance and Euclidean distance.is provided as online resource in the Appendix 1.

The sum-of-chi-squares (S-O-C-S) from multiple $df = 1$ tests can also be used as the gene-wise test statistic for calculating multiple testing $p$-value (Appendix 2) (Li et al. 2011; Lamparter et al. 2016) instead of the maximum-of-chi-squares (M-O-C-S). The geometric features of multiple testing can also be evaluated by this S-O-C-S approach. But, this method is not suitable for our Sph-Btv approach. Because, the Botev method (Botev 2017) considers the computational problems that are based on the $d$-dimensional multivariate normal law under linear restrictions. Also, this method is applicable when the number of linear inequalities is less than the dimension of the space. But, the S-O-C-S approach generates the non-linear contour hyperplanes for the tests that are not linear (Appendix 3). The A3 Figure A is presenting the contour hyperplanes of the tables for a $2 \times 3$ table test from the S-O-C-S approach, where all possible tables sharing the same marginal counts with the observed table are evaluated using three genetic models (additive, dominant and recessive). From this figure, it was observed that the each contour hyperplane is consisted of a linear summation of $\chi^2$-values, which means that the contour hyperplane contains many straight segments. Since, there are many segments, overall they look like "curved". Also, A3 Figure B is showing the non-linearity among the relations of the two co-ordinate values of the contour hyperplanes and the S-O-C-S values. So, this non-linear and too complicated contour hyperplanes are not feasible for Botev's approach.

There are multiple potential directions to extend our approach. One example is to handle the combinations of variables. In this paper, we demonstrated the multiple tests that were consisted of one particular test for multiple items (SNPs). However, the combinatorial effect of multiple SNPs is also the active research target, and our method can take the combinations along with single SNPs, as far as the combinatorial effect is expressed as linear that corresponds to the hyperplane in our geometric approach.

Another example is to apply our method to estimate type II errors rather than type I errors that we demonstrated in this paper. To estimate type II errors in the context of MAX-test, only one difference is to be introduced to our method. The difference is the parallel shift of normal distribution without moving the convex polytope of MAX-test, to the location in the space where an alternative hypothesis indicates.

## 8.4 Concluding remarks

The results shows that the multiple testing $p$-values are not uniform over the genome due to the heterogeneous structures of genes. This provides a strong evidence of the variations of type I errors among the genes. Because, the false detection rates are extremely dependent on the threshold value of the tests. Hence, consideration of uniform significance threshold for the genome-wide evaluation will provide over and under estimation of type I error rates depending on the gene structures. Therefore, the calculation of gene-wise significance cutoff by considering the individual structures of the genes will help to reduce the false positive signals from the studies.

# Chapter 9

# Conclusions and future works

## 9.1 Dissertation conclusions

Our study proposes a geometric testing procedure of contingency table in the context of MAX-test, that enables to estimate the multiple testing corrected $p$-value having multiple extendibility. This is a new addition for multiple testing problems in GWAS by considering of gene-based heterogeneity in terms of LD structures over the genome. The proposed Sph-Btv method is a geometrical demonstration of $MAX_{st}$ defined from the linear algebraic transformation. The new method is able to select the appropriate numbers of independent markers to be tested by defining the inner product matrix $(\Delta)$, which considers the individual LD structures of genes.

The evaluation indicates that the heterogeneity of the $p$-values affects the likelihood of false detection of a gene. The results confirm that the $p$-values are not uniform for the SNP subsets over the genome for multiple cutoff $\chi^2$-values.

In this research, we estimated the $p$-values of GWAS multiple testing using a geometric testing procedure that considers the genome-wide heterogeneity of individual SNP subsets. The results indicate that the $p$-values or the false detection probabilities per gene vary based on the change in $\chi^2$-cutoff when we perform GWAS. This variation is an important indicator of significant variation in type I error rates per gene. The appropriate calculation of cutoff $p$-values per gene considering gene-based heterogeneity could reduce the false positive detection rate significantly.

## 9.2 Future works

Many different adaptations can be introduced from various aspects that can be considered for the future due to lack of time. This study considers only one typical genetic scenario that is the additive model tests of the individual SNPs were used. But, the idea can be generalized to the MAX-test for an arbitrary number of tests using the geometry of the spherization method. The TND algorithm [Botev, 2017] is suitable for the case when $n_{test} \leq df$ and consequently fails to estimate the $p$-values if the covariance matrix of the distribution becomes singular as per the definition of the algorithm. The power of the proposed method would be useful to correctly assess whether or not the test correctly rejects the $H_0$. Hence, the convex polytope of the test's contour hyperplanes also expected to have significant effect on the estimated $p$-values under the $H_a$. Future directions of this research could include to:

1. generalize the approach for the arbitrary number of tests having the arbitrary contour lines, which can be done by generating directional unit vectors for estimation rather than using the Monte Carlo samples.

2. modify the algorithm to solve the issue of non-singularity of the variance-covariance matrix.

3. handle the combinations of variables that combinations of multiple SNPs.

4. estimate the power of the methods.

# References


1. Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955; **11**: 375-386.

2. Agresti A. *Categorical data analysis,* 2nd edn. John Wiley and Sons: New York, 2002.

3. Abel Z. Barycentric Coordinates. *An unpublished mathematical writings* 2007 (http://zacharyabel.com/papers/Barycentric_A07.pdf).

4. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.

5. Bender R, Lange S (2001) Adjusting for multiple testing—when and how? Journal of Clinical Epidemiology 54: 343-349. https://doi: 10.1016/S0895-4356(00)00314-0

6. Bickel PJ, Klassen CAJ, Ritov Y, Wellner JA. *Efficient and adaptive estimation in semiparametric models*, The Johns Hopkins University Press: Baltimore, 1993.

7. Botev ZI. The normal law under linear restrictions: simulation and estimation via minimax tilting. *J. R. Statist. Soc. B* 2017; **79**: 125-148.

8. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology* 2012; **8**: e1002822.

9. Buzdugan L, Kalisch M, Navarro A, Schunk D, Fehr E, Bühlmann P. Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics* 2016; **32**: 1990-2000.

10. Bühlmann P. Statistical significance in high-dimensional linear models. *Bernoulli* 2013; **19**: 1212–1242.

11. Bishop Y, Fienberg S, Holland P. *Discrete multivariate analysis*, 1st edn. Springer: New York, 2007.

12. Cargill M, Schrodi SJ, Chang M, Garcia VE, Brandon R, Callis KP *et al.* A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. *Am J Hum Genet* 2007; **80**: 273-290.

13. Chen SY, Feng Z, Yi X (2017) A general introduction to adjustment for multiple comparisons. Journal of Thoracic Disease, 9:1725-1729. https:// doi: 10.21037/jtd.2017.05.34

14. Cochran WG. Some methods for strengthening the common $\chi^2$ tests. *Biometrics* 1954; **10**: 417-451.

15. Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* 2002; **70**: 124-141.

16. Carlini E, Rapallo F. The geometry of statistical models for two-way contingency tables with fixed odds ratios. *Rend. Istit. Mat. Univ. Trieste* 2005; **XXXVII**: 71-84.

17. Delaneau O, Zagury JF, Marchini J. Improved whole chromosome phasing for disease and population genetic studies. *Nat Methods* 2013; **10**: 5-6.

18. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* 2016; **48**: 1284–1287.

19. Efron B. The length heuristic for simultaneous hypothesis tests. *Biometrika* 1997; **84**, 143–157.

20. Fienberg SE. The geometry of an $r \times c$ contingency table. *Annals of Mathematical Statistics* 1968; **39**: 1186-1190.

21. Fienberg SE, Gilbert JP. The geometry of a two by two contingency table. *Journal of the American Statistical Association* 1970; **65**: 694-701.

22. Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 2002; **53**: 146-152.

23. Gao X, Starmer J, Martin ER (2008) A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms. Genetic Epidemiology, 32:361–369. https://doi: 10.1002/gepi.20310

24. González JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V. Maximizing Association Statistics Over Genetic Models, *Genetic Epidemiology* 2008; **32**: 246-254.

25. Genz A. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* 1992; **1**: 141-149.

26. Gastwirth, JL. On robust procedures. *J. Am. Statist. Assoc.* 1966; **61**: 929–948.

27. Hirosawa K, Kawaguchi T, Matsuda F, Yamada R. Estimation of $p$-value of MAX Test with Double Triangle Diagram for $2 \times 3$ SNP Case-Control Tables. *Genetic Epidemiology* 2010; **34**: 543-551.

28. Higasa K, Miyake N, Yoshimura J, Okamura K, Niihori T, Saitsu H *et al.* Human genetic variation database, a reference database of genetic variations in the Japanese population. *Journal of Human Genetics* 2016; **61**: 547-553.

29. Hill WJ, Robertson A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 1968; **38**: 226–231.

30. Henk M, Richter-Gebert J, Ziegler GM. Basic properties of convex polytopes. In: Goodman JE, O'Rourke J, Tóth CD (eds). *Preliminary version of the Handbook of Discrete and Computational Geometry*, 3rd edn. CRC Press: Boca Raton, FL, 2017.

31. Hothorn LA, Hothorn T. Order-restricted Scores Test for the Evaluation of Population-based Case–control Studies when the Genetic Model is Unknown. Biometrical Journal 2009; **51**: 659-669.

32. Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Research, 11:2115–2119. https://doi: 10.1101/gr.204001

33. Hsu JC (1996) Multiple comparisons: theory and methods. Chapman & Hall: CRC Press, London.

34. Jain AK, Dubes RC. *Algorithms for Clustering Data*, Upper Saddle River NJ: Prentice Hall, 1988.

35. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, O'Brien SJ (2010) Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics, 11:724. https://doi: 10.1186/1471-2164-11-724

36. Kanai M, Toshihiro T, Okada Y. Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *Journal of Human Genetics* 2016; **61**: 861-866.

37. Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I. Fewer permutations, more accurate *p*-values. *Bioinformatics* 2009; **25**: i161-i168.

38. Kroese DP, Taimre T, Botev ZI. *Handbook of Monte Carlo Methods*, 2nd edn. John Wiley and Sons: New York, 2011.

39. Kuo C L, Feingold E. What's the best statistic for a simple test of genetic association in a case-control study?. *Genetic Epidemiology* 2010; **34**: 246–253.

40. Legendre P, Legendre L. *Numerical ecology*, 2nd edn. Elsevier Science BV: Amsterdam, 1998.

41. Lauritzen SL. *Graphical models*, 1st edn. Oxford University Press: USA, 1996.

42. Li Q, Zheng G, Li Z, Yu K. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. *Ann Hum Genet* 2008; **72**: 397-406.

43. Li Q, Zheng G, Liang X, Yu K. Robust tests for single-marker analysis in case-control genetic association studies. *Ann Hum Genet* 2009; **73**: 245-252.

44. Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 2005a; **21**: 781–787.

45. Lin DY. On rapid stimulation of *P* values in association studies. *Am J Hum Genet* 2005b; **77**: 513–515.

46. L'Ecuyer P, Blanchet JH, Tuffin B, Glynn PW. Asymptotic robustness of estimators in rare-event simulation. *ACM Trans. Model. Comput. Simul.* 2010; **20**: article 6.

47. Loley C, König IR, Hothron L, Ziegler A. A unifying framework for robust association testing, estimation, and genetic model selection using the generalized linear model. *European Journal of Human Genetics* 2013; **21**: 1442-1448.

48. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLOS Computational Biology* 2016; **12**: e1004714.

49. Li MX, Gui HS, Kwan JSH, Sham PC. GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *The American Journal of Human Genetics* 2011; **88**: 283-293.

50. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM *et al*. A Versatile Gene-Based Test for Genome-wide Association Studies. *The American Journal of Human Genetics* 2010; **87**: 139-145.

51. McDonald JH (2014) Handbook of biolological statistics. Sparky House Publishing, U.S.A.

52. Mantel N. Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *J Am Stat Assoc* 1963; **58**: 690–700.

53. Mandozzi J, Bühlmann P. Hierarchical testing in the high-dimensional setting with correlated variables. *J. Am. Stat. Assoc.* 2015; **111**: 331–343.

54. Meinshausen,N, Meier L, Bühlmann P. *p*-Values for high-dimensional regression. *JASA* 2009; **104**: 1671–1681.

55. Milne RL, Ribas G, González NA, Fagerholm R, Salas A, González E *et al.* ERCC4 associated with breast cancer risk: a two-stage case-control study using high-throughput genotyping. *Cancer Res* 2006; **66**: 9420-9427.

56. Metaxoglou K, Pettenuzzo D, Smith A. Exponential Tilting with Variance Risk Premia: Equity Premium and Oil Prices. *An unpublished preliminary draft* 2015.

57. Nyholt DR (2004) A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. American Journal of Human Genetics, 74:765–769. https:// doi: 10.1086/383251

58. Pooley KA, Healey CS, Smith PL, Pharoah PD, Thompson D, Tee L *et al.* Association of the progesterone receptor gene with breast cancer risk: a single-nucleotide polymorphism tagging approach. *Cancer Epidemiol Biomarkers Prev* 2006; **15**: 675-682.

59. Puig P. A note on testing segregation between two groups of animals using entropy. *Biometrical Journal* 1999; **40**: 155–163.

60. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002; **70**: 425–434.

61. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American Journal of Human Genetics, 69:138–147. https://doi: 10.1086/321276

62. Slavković AB, Fienberg SE. Algebraic geometry of $2 \times 2$ contingency tables. In: Gibilisco P, Riccomagno E, Rogantin MP, Wynn HP (eds). Algebraic and Geometric Methods in Statistics, 1st edn. Cambridge University Press: UK, 2010.

63. Slager SL, Schaid DJ. Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Hum Hered* 2001; **52**: 149–153.

64. So HC, Sham PC. Robust Association Tests Under Different Genetic Models, Allowing for Binary or Quantitative Traits and Covariates. *Behav Genet* 2011; **41**: 768-775.

65. Strasser H, Weber C. On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics* 1999; **8**, 220–250.

66. Simes RJ. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrica* 1986; **73**: 751-754.

67. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* 2017; **101**: 5-22.

68. Van BG, Fisher LD, Heagerty PJ, Lumley T. *Biostatistics a methodology for the health sciences*, 2nd edn. Chapter 7. Hoboken, NJ: Wiley-Interscience, 2004.

69. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* 2014; **42**: D1001-D1006.

70. Warren J, Schaefer S, Hirani AN, Desbrun M. Barycentric Coordinates for Convex Sets. *Advances in Computational Mathematics* 2005; **27**: 319-338.

71. Wang K, Sheffield VC. A constrained-likelihood approach to marker-trait association studies. *Am J Hum Genet* 2005; **77**: 768–780.

72. Yoshimura K, Nakayama T, Sekine A, Matsuda F, Kosugi S, Yamada R *et al.* B-Type Natriuretic Peptide as an Independent Correlate of Nocturnal Voiding in Japanese Women. *Neurourology and Urodynamics* 2012; **31**: 1266-1271.

73. Yamada R, Okada Y. An optimal dose-effect mode trend test for SNP genotype tables. *Genetic Epidemiology* 2009; **33**: 114-127.

74. Zhang Y. On The Use of P-Values in Genome Wide Disease Association Mapping. *Journal of Biometrics & Biostatistics* 2016; **7**:1000297.

75. Zheng G, Freidlin B, Gastwirth JL. Comparison of robust tests for genetic association using case-control studies. *IMS Lecture Notes-Monograph Series, 2nd Lehmann Symposium-Optimality* 2006; **49**: 253-265.

76. Zang Y, Fung WK. Robust Mantel-Haenszel Test Under Genetic Model Uncertainty Allowing for Covariates in Case-Control Association Studies. *Genetic Epidemiology* 2011; **35**: 695-705.

77. Zang Y, Fung WK, Zheng G. Simple algorithms to calculate asymptotic null distributions of robust tests in case-control genetic association studies in R. *J Stat Softw* 2010; **33**:1–24.

78. Zhang CH, Zhang S. Confidence intervals for low dimensional parameters in high dimensional linear models. J. *R. Stat. Soc. Ser. B: Stat. Methodol.* 2014; **76**: 217–242.

79. Zhang J, Deng B, Liu Z, Patané G, Bouaziz S, Hormann K *et al*. Local Barycentric Coordinates. *ACM Transactions on Graphics* 2014; **33**: 188.1-188.12.

80. Zheng G, Joo J, Yang Y. Pearson's test, trend test, and MAX are all trend tests with different types of scores. Ann Hum Genet. 2009; 73: 133-140.

# Appendix 1

## **A1** Relation among the correspondence analysis (CA), Mahalanobis distance and Euclidean distance.

CA is a multivariate technique to visualize categorical data (Greenacre 1984; Cuadras CM and Cuadras D 2006). This method is used to analyze simple 2-way and multi-way contingency tables containing some measure of correspondence between the rows and columns, which explores the structure of categorical variables included in the table using the chi-square ($\chi^2$) distance between the profiles of rows (and columns) (Hill and Lewicki 2006; Cuadras CM and Cuadras D 2006). According to Cuadras CM, Cuadras D (2006) and Greenacre (1984) the CA can be described as below:

Let $N = (n_{ij})$ be an $I \times J$ contingency table and $P = n^{-1}N$ the correspondence matrix, where $n = \sum_{ij} n_{ij}$. Let $K = min\{I \times J\}$ and $r = P1$, $D_r = diag(r)$, $c = P'1$, $D_c = diag(c)$, the vectors and diagonal matrices with the marginal frequencies of $P$, where 1 is the vector of ones. Using the singular value decomposition,

$$D_r^{1/2}(P - rc')D_c^{1/2} = UD_\lambda V' \tag{S1}$$

where $D_\lambda = diag(\lambda_1, \dots, \lambda_{K-1})$ is a diagonal matrix of singular values in descending order, $U$ is an orthogonal matrix and the columns of $V$ are orthogonal. To represent the $I$ rows of $N$ the principal coordinates of the rows of $A$ are used, where,

$$A = D_r^{1/2}UD_\lambda \tag{S2}$$

Then the squared Euclidean distance between rows $i$, $i'$of $A$ equals the $\chi^2$distance,

$$\delta_{ii'}^2 = \sum_{j=1}^{J} \left( \frac{p_{ij}}{r_i c_j} - \frac{p_{i'j}}{r_{i'} c_j} \right)^2 c_{j.} \tag{S3}$$

Similarly, to represent the $J$ columns of $N$, the principal coordinates in the rows of $B$ or the standard coordinates $B_0$ are used, where,

$$B = D_c^{1/2} V D_\lambda, B_0 = D_c^{1/2} V \quad \text{(S4)}$$

The transitive relations,

$$A = D_r^{-1} P B {D_\lambda}^{-1}, B = D_c^{-1} P' A {D_\lambda}^{-1} \quad \text{(S5)}$$

allow to perform a joint representation of rows and columns, called the symmetric representation (Cuadras CM and Cuadras D 2006).

Geometrically, CA displays the rows and columns of a contingency table as points in dual low-dimensional vector spaces. The row and column points reside in a stretched barycentric coordinate space having the dimension of $(K-1)$ (Greenacre 1984). For example, the rows and columns of a $(2 \times 3)$ table (2-way table) can be realized in a 1-dimensional space. For a $(5 \times 18 \times 3)$ table (3-way table), where the frequencies of 3 column categories are within 18 row categories of 5 variables, the row profiles are three vectors with elements summing to 1 for the corresponding table of the relative frequencies. These three row vectors can be realized in a 3-dimensional space as an equilateral triangle with vertices $(1 \quad 0 \quad 0)^T$, $(0 \quad 1 \quad 0)^T$ and $(0 \quad 0 \quad 1)^T$. This triangle represents all three vectors whose elements are non-negative and add up to one with the vertices representing the three most polarized profiles. All the row profiles can be represented exactly in a 2-dimensional triangle by taking the triangle out of the space. This is the triangular (or barycentric) co-ordinate system, which is often used to represent the data of a contingency table consisting of sets of three values with fixed marginal counts.

This algebraic geometry of contingency tables can be generalized for multi-way tables, where multiple row (and column) vectors can be realized in the multidimensional vector space. When the dimensions (columns) of this vector space are correlated to each other, the

Euclidean distance between a point and the center of the points (distribution) is not appropriate. As the Euclidean distance is a distance between two points only, it does not consider how the rest of the points in the dataset vary.

In this situation, the Mahalanobis distance (Mahalanobis 1936) is applicable that measures the distance between a point and a distribution, which is effectively a multivariate equivalent of the Euclidean distance (Greenacre 1984). It transforms the columns into uncorrelated variables, scale the columns to make their variance equal to one, and finally, calculates the Euclidean distance. So, the Mahalanobis distance is a standardized form of Euclidean distance.

**References**

Cuadras CM, Cuadras D. A parametric approach to correspondence analysis. *Linear Algebra and its Applications* 2006; **417**:64–74.

Greenacre MJ. *Theory and Applications of Correspondence Analysis*. Academic Press: London, 1984.

Hill T, Lewicki P. *Statistics: methods and applications-A comprehensive reference for science, industry and data mining*. StatSoft, Inc.: United States, 2006.

Mahalanobis PC. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* 1936; **2**:49–55.

# Appendix 2

## **A2** Correction of $p$-values by the sum-of-chi-squares (S-O-C-S) and the maximum-of-chi-squares (M-O-C-S) approaches.

The S-O-C-S and the M-O-C-S can be used as the gene-wise statistics to calculate the corrected $p$-value (Li et al. 2011; Lamparter et al. 2016). They measure the strongest and the average association signals in the gene-based association studies, respectively (Lamparter et al. 2016). Some fast and accurate tools like Pascal (Pathway scoring algorithm) (Lamparter et al. 2016) having high power has been developed for gene and pathway-level analysis of GWAS association results without accessing the original raw genotype data. The gene-wise $p$-value is calculated efficiently from the gene score by aggregating the SNP-wise $p$-values from the GWAS meta-analysis.

The two options S-O-C-S and M-O-C-S based on the most significant SNP and the average association signal across the region, respectively, can be used for aggregating SNP-wise summary statistics into the gene score. The $p$-values are corrected for the gene sizes and LD structures for the given information from a reference population. In the absence of the original genotype data, the correlation matrix is estimated from ethnicity-matched, publicly available genotype data assuming that the null distributions of the S-O-C-S and M-O-C-S statistics depend solely on the pairwise correlation matrix of the contributing genotypes.
In the Pascal (Lamparter et al. 2016), the correlation structure between the $Z$-scores of SNP association values are estimated from a reference population. The $Z$-scores of $n$ SNPs follow a multivariate normal distribution under the null hypothesis, i.e., $Z \sim N_n(0, \Sigma)$, where $\Sigma$ is the pairwise SNP-by-SNP correlation matrix. The base statistics, S-O-C-S ($T_{sum}$) and M-O-C-S ($T_{max}$) are defined as,

$$T_{sum} = \sum_{i=1}^{n} Z_i^2 \qquad \text{(S6)}$$

and

$$T_{max} = \max(Z_i^2) \tag{S7}$$

, respectively. The $T_{sum}$ is distributed according to the weighted sum of $\chi_1^2$-distributed random variables (Lamparter et al. 2016),

$$T_{sum} \sim \sum_{i=1}^{n} \lambda_i \chi_1^2 \tag{S8}$$

where $\lambda_i$ is the $i$-th eigenvalue of Σ. The null distribution of $T_{max}$ can be obtained as,

$$\begin{aligned} Pr(T_{max} \geq t) &= Pr\big(\max(|Z_i| \geq t)\big) \\ &= 1 - Pr(|Z_i| < t, i = 1,2,\cdots,n) \end{aligned} \tag{S9}$$

The equation (S9) is a rectangular integration over a multivariate normal. The gene-wise $p$-values were approximated by multiplying the minimum $p$-value of SNPs in the region with the effective number of tests ($n_{test}$) within the gene.

On the other hand, a rapid gene-based association test that uses extended Simes procedure (GATES) (Li at al. 2011) to assess the gene-wise statistical association significance was proposed by assuming the availability of $p$-values from a test of association between the disease and each of the SNPs within a gene and pairwise correlation coefficients ($r$) for all the SNPs. This is a modification of the Simes test (Simes 1986) that can rapidly combine the $p$-values of SNPs within a gene, without relying on raw, individual phenotype and genotype data, to produce valid gene-based $p$-values. Let $p_{(i)}, i = 1,2,\cdots,n$ are the ascending ordered $p$-values of $n$-SNPs within a gene. In GATES, the overall $p$-value is obtained after combining these $n$ SNP-based $p$-values for the gene as,

$$p_G = \min\left(\frac{n_e p_{(j)}}{n_{e(j)}}\right) \tag{S10}$$

where $n_e$ and $n_{e(j)}$ are the effective number of independent $p$-values among the $n$ SNPs and the top $j$ SNPs, respectively.

The value of $n_e$ is estimated from,

$$n - \sum_{i=1}^{n} [I(\lambda_i > 1)(\lambda_i - 1)]\lambda_i > 0 \tag{S11}$$

where $I(x)$ is an indicator function and $\lambda_i$ is the $i$-th eigenvalue of the pairwise SNP-by-SNP correlation matrix Σ such that,

$$\begin{array}{ccll} \lambda_i = 1, n_e = n & i = 1,2,\cdots,n & \text{all SNPs are independent} & \\ \lambda_1 = n, \lambda_i = 0, n_e = 1 & i = 2,3,\cdots,n & \text{all SNPs are in complete LD} & \text{(S12)} \\ \lambda_i = 0 & i = 1,2,\cdots,n & \text{for negative eigenvalues} & \end{array}$$

For the other intermediate situations, the equation (S10) also provides an appropriate effective number of SNP $p$-values and a uniform approximation of the distribution of $p_G$ (Li et al. 2011).

The gene-based test can be further extended to incorporate differential SNP weights as follows,

$$p_G = \min\left(\frac{n_e p_{(j)}}{\sum_{k=1}^{j} w_{(k)}}\right) \tag{S13}$$

This calculation considers the sorted SNPs according, $w_{(i)} = c\left(n_{e(i)} - n_{e(i-1)}\right)r_{(i)}$, $n_{e(0)} = 0$ and $c = \frac{n_e}{\sum_{i=1}^{n}\left(n_{e(i)} - n_{e(i-1)}\right)r_{(i)}}$. The weights, $w_{(i)}$ are calculated from prior weights $r_{(i)}, i = 1,2,\cdots,n$, which are being set according to the relative functional importance of the SNPs, where $\sum_{i=1}^{n} w_{(i)} = n_e$ and $w_{(i)} > 0$.

The use of weights is expected to increase statistical power if SNPs with higher weights are more likely to be associated with the disease than SNPs with lower weights and equal weights can be used if information is not available.

The type I error rate of the GATES were compared with some other existing gene-based tests (Li et al. 2011). A versatile gene-based test for genome-wide association studies (VEGAS) (Liu et al. 2010) is one of them, where the comparison is carried out with the two versions such as VEGAS-sum and VEGAS-max calculated from the S-O-C-S and M-O-C-S, respectively.

**References**

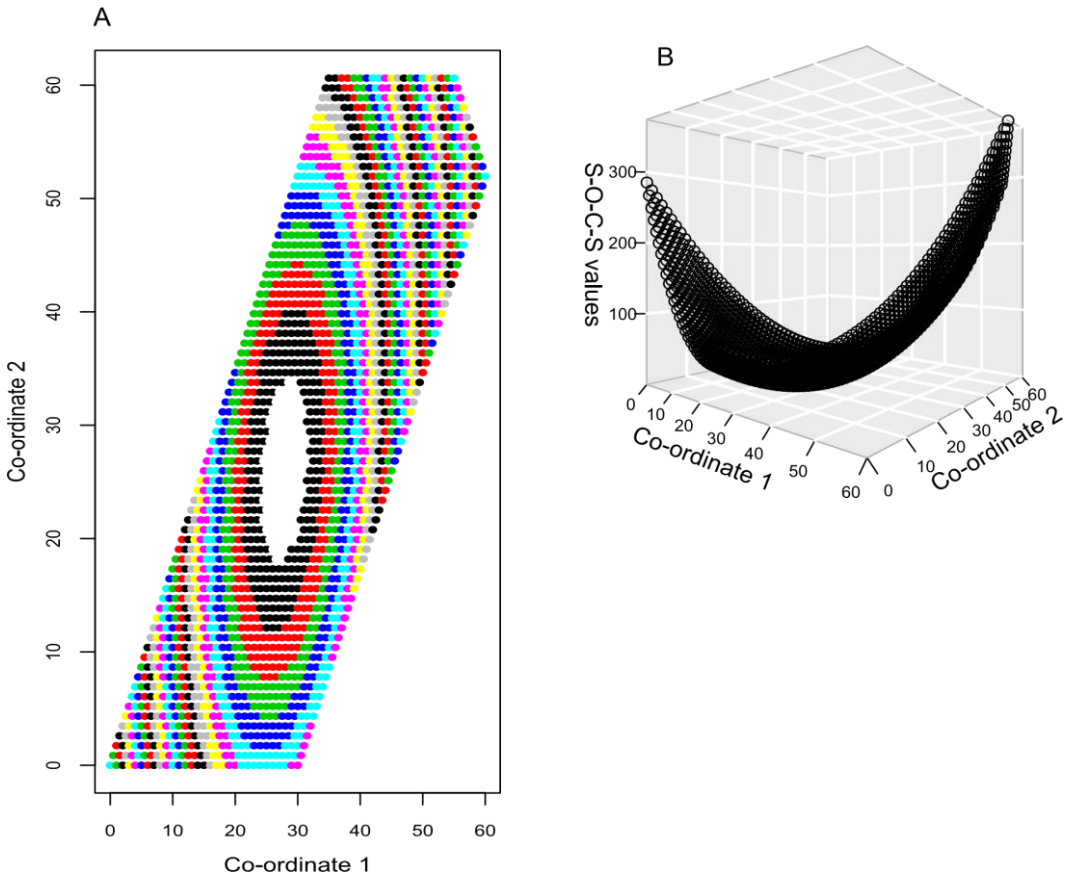
Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLOS Computational Biology* 2016; **12**: e1004714.

Li MX, Gui HS, Kwan JSH, Sham PC. GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *The American Journal of Human Genetics* 2011; **88**: 283-293.

Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM *et al*. A Versatile Gene-Based Test for Genome-wide Association Studies. *The American Journal of Human Genetics* 2010; **87**: 139-145.

Simes RJ. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrica* 1986; **73**:751-754.

# Appendix 3




**A3 Figure. An application of the S-O-C-S approach for a $2 \times 3$ table using three genetic models (additive, dominant and recessive).** (A) The contour hyperplanes of the tables sharing the same marginal counts with an observed table. Each curve having the same color is the hyperplane of the tables with same $\chi^2$-values from the S-O-C-S approach. (B) The relation among the two co-ordinate values of the contour hyperplanes and the integrated $\chi^2$-values.

# Appendix 4

**A4 Algorithm.** The algorithm to calculate $p$-values by the Sph-Btv method.

The Sph-Btv method is useful if the number of tests ($n_{test}$) is less than or equal to $df$. Four arguments, lower and upper truncation limits ($\pm\chi_{max}$), a variance-covariance matrix ($\Delta$) and the number of Monte Carlo samples ($n_{samples}$) are required for calculation. The number of Monte Carlo samples is selected depending on the dimension of the space defined by $n_{test}$.

---

**Require:** $n_{test} \leq df = \prod_i (I_i - 1)$

$R \leftarrow 1$

**for** $i \in I$ **do**

Calculate $R(i)$ by equation 4

$R \leftarrow R \otimes R(i)$

**end for**

Calculate $V$ depending on $I$

Calculate $E^{-1}$ from the given contingency table

Calculate $X$ by decomposing as $X^T X = V\ R\ E^{-1} R^T V^T$

Calculate $X^{-1}$

Calculate $\mathcal{P}^{\dagger} = R^T V^T X^{-1}$

**for** $i = 1,2, \ldots, n_{test}$ **do**

Calculate $\chi_i$ using the proportion trend test with given contingency table and $i$-th test weight.

**end for**

$\chi_{max} = max_i\ \chi_i$

Compute truncation limits, $u = 1_{n_{test}}\chi_{max}$ and $l = -1_{n_{test}}\chi_{max}$

**for** $i = 1,2, \dots, n_{test}$ **do**

Calculate $\tau_i^T = \frac{w_i^T \mathcal{P}^\dagger}{|w_i^T \mathcal{P}^\dagger|}$

**end for**

Compute $A = \begin{pmatrix} \tau_1^T \\ \vdots \\ \tau_{n_{test}}^T \end{pmatrix}$

Compute $\Delta = A\ A^T$

Run the Sph-Btv with four arguments, $(l, u, \Delta, n_{samples})$.

# Appendix 5

**A5 Code.** R-Codes to run the Sph-Btv approach.

```
library(mvtnorm)
library(TruncatedNormal)

# Calculates marginal counts of a contingency table "A".
calc.marg<-function (A){
    d <- dim(A)
    ret <- list()
    for (i in 1:length(d)) {
        ret[[i]] <- apply(A, i, sum)
    }
    ret
}

# Calculates expected table of a contingency table "A".
make.exp.table <- function(A){
  n <- sum(A)
  marg <- calc.marg(A)
  tmp <- marg[[1]]
  for(i in 2:length(marg)){
    tmp <- t(marg[[i]]/n) %x% tmp
  }
  tmp
}

# Calculates a k × k simplex-rotation matrix for an integer k.
```

```
# Example: For a 2 × 3 table, k = 2, 3. For each k, this function returns two rotation matrices with dimensions, 2 × 2 and 3 × 3, respectively.
make.simplex <- function(k){
  ret <- matrix(0,k,k)
  for(i in 1:(k-1)){
    for(j in 1:k){
      if(j < i){
      }else if(j==i){
        ret[i,j] <-  sqrt((k-i)/(k-i+1))
      }else{
        ret[i,j] <- -sqrt(1/((k-i)*(k-i+1)))
      }
    }
  }
  ret[k,] <- sqrt(1/k)
  ret
}

# Calculates the Kronecker product of multiple simplex-rotation matrices.
# r: a vector of numbers of the levels of the variables of a contingency table (the dimension of the table).
# Example: For a 2 × 3 table, r = c(2, 3), this function returns a  6 × 6 matrix.
make.simplex.multi <- function(r){
  X <- make.simplex(r[1])
  k <- length(r)
  if(k > 1){
    for(i in 2:k){
      X <- make.simplex(r[i]) %x% X
    }
```

```
  }
  X
}
```

# Calculates the index vector to define the positions of zero and non-zero elements produced from the rotation.

# Example: For a $2 \times 3$ table, d = c(2, 3).

```
arrive.index <- function(d){
  n <- prod(d-1)
  x <- numeric(n)
  ind <- 1
    for(i in 1:(d[2]-1)){
      for(j in 1:(d[1]-1)){
      x[ind] <- (i-1)*d[1]+j
      ind <- ind+1
    }
  }
  x
}
```

# Spherization of table.

# Returns the matrices that will transfer test vectors to the $df$-dimensional vectors.

```
calc.rotate <- function(table){
  dim <- dim(table)
  exp.table <- make.exp.table(table)
  e.vec <- c(exp.table)
  R <- make.simplex.multi(dim)
  Einv <- diag(1/e.vec)
  d.vec <- c(table)-e.vec
```

```
  Z <- diag(prod(dim))[arrive.index(dim),]
  XtX <- Z %*% R %*% Einv %*% t(R) %*% t(Z)
  eigenout <- eigen(XtX)
  X <- diag((eigenout[[1]])^0.5) %*% solve(eigenout[[2]])
  P <- X %*% Z %*% R
  Pinv <- t(R) %*% t(Z) %*% solve(X)
  return(list(P=P,Pinv=Pinv))
}
```

# Calculates the $df$-dimensional test vectors.

# tests: a list of tables representing test models.

# rotation: the spherization information from the "calc.rotate" function.

```
make.test.vecs <- function(rotation,tests){
  test.vecs <- tests %*% rotation
  L.test.vecs <- sqrt(apply(test.vecs^2,1,sum))
  test.vecs / L.test.vecs
}
```

# Calculates the $p$-value(s) for the Sph-Btv method.

# stat: the $\chi^2$-value(s) of the table from the proportion trend test.

# table: the contingency table.

# lower.tail: logical; if TRUE, probabilities are $P[X \leq x]$, otherwise, $P[X > x]$.

# one.side: logical; if TRUE, the test vectors indicate one-sided test, otherwise, two-sided.

# n: the number of Monte Carlo samples.

```
pmway.table.null.multi.botev <- function(stat,table,tests,lower.tail,one.side,n){
  d <- dim(table)
```

```
  df <- prod(d-1)
  rotate <- calc.rotate(table)$Pinv
  test.vecs <- make.test.vecs(rotate,tests)
  sig <- test.vecs %*% t(test.vecs)
  pr <- numeric(length(stat))
  for(i in 1:length(stat)){
    u <- rep(stat[i],nrow(sig))
    pr[i] <- 1 - mvNcdf(-u,u,sig,n)$prob
  }
  if(lower.tail){
    if(one.side){
      ret <- 1 - (1 - pr)/2
    }else{
      ret <- pr
    }
  }else{
    if(one.side){
      ret <- (1 - pr)/2
    }else{
      ret <- 1 - pr
    }
  }
  ret
}
```

# Achievements

## Presentations

1. **Basak T**, Nagashima K, Kajimoto S, Yamada R. Feature description of truncated normal distributions of additive tests for multiple testing p-value correction – An application to GWAS SNP data. *Conference of the International Federation of Classification Societies (IFCS)* 2017; Tokai University, Tokyo, Japan, August 8-10.

2. **Basak T**, Nagashima K, Kajimoto S, Yamada R. A geometry-based method for multiple testing problems using maximum statistics and notion of distance for multi-way tables. *Human Genome Meeting (HGM)* 2017; Hotel Barceló saints, Barcelona, Spain, February 5-7.

3. **Basak T**, Yamada R. A linear algebraic method for evaluating the relation between power and the pattern of linkage disequilibrium in multiple testing. *International Congress of Human Genetics (ICHG)* 2016; Kyoto International Conference Center, Japan, April 3-7.

## Manuscript submitted for review

1. **Basak T**, Nagashima K, Kajimoto S, Kawaguchi T, Tabara Y, Matsuda F, Yamada R. A Geometry-Based Multiple Testing Correction for Contingency Tables by Truncated Normal Distribution, *Statistics in Biosciences* 2020; **12**:63-77.

# The End