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

Tapati Basak¹ · Kazuhisa Nagashima¹ · Satoshi Kajimoto¹ ·
Takahisa Kawaguchi² · Yasuharu Tabara² · Fumihiko Matsuda² · Ryo Yamada¹

Received: 28 August 2019 / Revised: 22 January 2020 / Accepted: 18 February 2020 /
Published online: 5 March 2020
© The Author(s) 2020

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 χ^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

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12561-020-09271-6>) contains supplementary material, which is available to authorized users.

✉ Ryo Yamada
ryamada@genome.med.kyoto-u.ac.jp

¹ Unit of Statistical Genetics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

² Human Disease Genomics, Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

1 Introduction

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 [12]. For example, in case of high-dimensional data obtained from the field of genetics, medicine, molecular biology, bioinformatics, agricultural science etc. [13]. 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 [4, 18]. 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 [4]. 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.

Therefore, the correction of p -value is obvious for multiple testing to control the type I error [1]. 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 [4]. 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 [3]. 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 forms of this test is the popular chi-square (χ^2) test [3]. Selection of the largest test statistic (MAX-test) from different genetic models is a powerful approach as it provides safeguard against model uncertainty [7].

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 [13]. Moreover, correction of multiple testing p -values is also required for the SNPs in linkage disequilibrium (LD) with each other [19]. 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 [6]. The permutation test can be applied [11, 20]. But, an enormous number of permutations are required to accurately estimate small p -values, constituting a computational burden [14].

We combined the geometry of multi-way contingency table [21] and the χ^2 -test [9, 22] 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 [2] 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. Therefore, the main aim of this paper is to propose a multiple testing correction method for the contingency table tests as well as to study its features using a real data set.

2 Methods

2.1 Spherization

This method combines the idea of algebraic geometry of a contingency table [21] and χ^2 -statistic [9, 22] to develop the MAX-test from the spherical geometry.

2.1.1 Geometry of Multi-way Contingency Tables

Let Y be a m -dimensional multi-way contingency table with non-negative elements Y_i , where $i \equiv (i_1, i_2, \dots, i_m)$, $1 \leq i_j \leq I_j$ and I_j is the number of categories of j -th dimension. To consider the geometry of contingency tables, we define the vectorized form of Y as y by defining the index. The index of i -th element is

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

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.

2.1.2 Geometry of the χ^2 -Test

2.1.2.1 Dimension Reduction of Contingency Tables For the χ^2 -test, the tables sharing same marginal counts can be located in the same $df \equiv \prod_j (I_j - 1)$ dimensional

space, which is smaller than M -dimensional space as the tables are restricted by the 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

$$d \equiv o - e, \quad (2)$$

where o and e are the observed and expected frequency vectors for arbitrary tables in M -dimensional space, respectively. Next, a rotation matrix R is defined depending on I_j :

$$R_{M \times M} \equiv \otimes \prod_{j=1}^m R(I_j). \quad (3)$$

Equation (3) defines the Kronecker product of m rotation matrices of m -dimensional multi-way contingency table. Each of m matrices is a rotation matrix in an n -dimensional space, of which n is the number of categories in I_j for the j -th dimension. The structure for one variable is

$$R(n) \equiv \begin{pmatrix} r_1^T \\ r_2^T \\ \vdots \\ r_n^T \end{pmatrix} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,n} \end{pmatrix}, \quad (4)$$

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

Rotation of the Eq. (2) by (3) 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. 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:

$$v \equiv \otimes \prod_{j=1}^m u(I_j), \quad (5)$$

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 2×3 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 for the vectors of tables are reduced without loss of information.

2.1.2.2 Spherization of the χ^2 -Statistic Using vectorized form of Y , the χ^2 -statistic in df -dimensional space is:

$$\chi_{df}^2 = d^T E^{-1} d, \quad (6)$$

where E is a diagonal matrix having e in the main diagonals. For dimension reduction, the Eq. (6) can be written as

$$\chi_{df}^2 = d^T R^T R E^{-1} R^T R d, \quad (7)$$

$$\chi_{df}^2 = d^T R^T V^T V R E^{-1} R^T V^T V R d, \quad (8)$$

where VRd is a df -dimensional vector, and $VRE^{-1}R^TV^T$ is a $(df \times df)$ -dimensional symmetric matrix. Using LU decomposition, Eq. (8) can be transformed into

$$\chi_{df}^2 = d^T R^T V^T X^T X V R d, \quad (9)$$

$$\chi_{df}^2 = g^T g, \quad (10)$$

where X is an LU decomposed matrix of $VRE^{-1}R^TV^T$ and $g = XVRd$. Hence, Eq. (10) 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.

2.1.2.3 Geometry of the Proportion Trend Test The geometry of the χ^2 -statistic can be applied to deploy the proportion trend test in df -dimensional space. Let us consider a 2×3 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 χ_w is the function of a weighted sum of the differential table:

$$\chi_w = f(w^T d). \quad (11)$$

To deploy Eq. (11) in df-dimensional space, we use the deformation of the definition of g in Eq. (10):

$$g \equiv \mathcal{P}d, \quad (12)$$

$$g = XVRd, \quad (13)$$

$$VRd = X^{-1}g, \quad (14)$$

$$V^T VRd = V^T X^{-1}g, \quad (15)$$

$$Rd = V^T X^{-1}g, \quad (16)$$

$$d = R^T V^T X^{-1}g, \quad (17)$$

$$d = \mathcal{P}^\dagger g, \quad (18)$$

where $\mathcal{P}^\dagger = R^T V^T X^{-1}$ is a sparse version of the pseudo-inverse of $\mathcal{P}$. The deformation from Eq. (15) to (16) only holds for cases, in which the dimension reduction to df-dimensional space of spherization is applicable. Then, from Eq. (11):

$$\chi_w = f(w^T \mathcal{P}^\dagger g). \quad (19)$$

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

$$\tau^T = \frac{w^T \mathcal{P}^\dagger}{|w^T \mathcal{P}^\dagger|}. \quad (20)$$

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

$$\tau^T g = |g| \cos \theta \quad (21)$$

Equation (21) defines the inner product, which is the relation between observations and weight, where θ is the angle between observation and weight. If the observation is in the direction of $\theta = 0$ so that $(\tau^T g)^2 = |g|^2 = \chi_{df}^2$, then the χ^2 -statistic for a vectorized weight is

$$\chi_w^2 = |\tau^T g|^2, \quad (22)$$

where χ_w^2 follows a χ_1^2 -distribution, and the p -value is the proportion of tables having χ^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 MAX-test. Pairs of truncating hyperplanes can be drawn in df -dimensional space depending on a set of test vectors and χ_w . The hyperplanes are defined as vertical to the test vector, and the difference from the origin to the hyperplane is χ_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.

3 Geometric Features of Convex Polytopes in Spherized Space

The geometric configuration of the convex polytope defines the multiple testing p -values. Figure 1a shows the case of $df = 2$ with $df = 1$ test for a single test. Tables with the same χ^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 the tables on the ellipsoid are not equidistant from the origin (Fig. 1a). 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 the tables are placed on the transformed df -dimensional n -sphere. The χ^2 -value is calculated as the squared Euclidean distance from the expected table to the hyperplane with distance χ , and the p -value is the proportion of tables outside of the convex polytope (Fig. 1b). The definition of the hyperplane for a single test is

Fig. 1 The spherical geometry of tests in df -dimensional space. **a** The contour hyperplanes of the distribution of tables with the same χ^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**, **c** 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 χ^2 -value is the squared Euclidean distance from the center to the hyperplane having distance χ . **c** Multiple testing: Each pair of parallel hyperplanes of the same color presents a contour hyperplane of the tables having the same χ^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 (Color figure online)

$$-\chi_w \leq \tau^T Z \leq \chi_w, \quad (23)$$

where Z is the axis of df -dimensional space.

In multiple testing, multiple pairs of parallel hyperplanes form a convex polytope (Fig. 1c, light-yellow-colored area), in which each pair corresponds to the tables with same χ^2 -values from the individual $df = 1$ test. The definition of the convex polytope in Eq. (23) is generalized for n_{test} as:

$$-\chi_{\max} \mathbf{1}_{n_{\text{test}}} \leq AZ \leq \chi_{\max} \mathbf{1}_{n_{\text{test}}}, \quad (24)$$

where $\chi_{\max} \equiv \max_k \chi_{w_k}$; $k = 1, 2, \dots, n_{\text{test}}$, A is a $(n_{\text{test}} \times df)$ matrix with rows that are test vectors, $\chi_{\max}$ is the largest value of the test statistics among n_{test} , and $\mathbf{1}_{n_{\text{test}}}$ is a $(n_{\text{test}} \times 1)$ 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 χ^2 -values. Hence, the probability of observing tables with representative χ^2 -values greater than or equal to the observed table's χ^2 -value is the integral of the probability density of a multivariate normal distribution over the space out of the convex polytope. Therefore, the p -value for n_{test} is

$$p = 1 - \int N(Z|0, I) \delta(-\chi_{\max} \mathbf{1}_{n_{\text{test}}} \leq AZ \leq \chi_{\max} \mathbf{1}_{n_{\text{test}}}) dZ, \quad (25)$$

where δ 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 testing H_0 . Because of the relationship between n -dimensional normal distributions and χ^2 -distributions, we define χ as the distance from the center to the point corresponding to the observed table, and also obtain the spherized space (Fig. 1b, circle) with a multivariate normal distribution demarcated by a convex polytope, producing TND (Fig. 1c).

4 The Spherization and Botev's (Sph–Btv) Approach

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 [2] as:

$$1 - \Pr(l \leq AZ \leq u) = 1 - \int_l^u \phi(Z; 0, I_d) dZ, \quad (26)$$

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 probability density function 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. (26) as [2]:

$$1 - \Pr(l - A\mu \leq A\mathbb{Z} \leq u - A\mu) \quad (27)$$

$$= 1 - \Pr(l \leq A\mathbb{Z} \leq u) \quad (28)$$

$$= 1 - \int_l^u \phi(\mathbb{Z}; 0, A\Sigma_Z A^T) d\mathbb{Z}, \quad (29)$$

where $\mathbb{Z} \sim N(0, A\Sigma_Z A^T)$, $\Delta_{n_{\text{test}} \times n_{\text{test}}} = A\Sigma_Z A^T = AA^T$ is an inner product matrix. Here, the p -values were estimated by the Botev's [2] approach after the transformation of the space with spherization. The definition of the convex polytope for n_{test} can be defined from the Eq. (28):

$$l \leq A\mathbb{Z} \leq u, \quad (30)$$

where $A_{n_{\text{test}} \times df}$ is the df -dimensional matrix, the rows of which are the test vectors defined by the test models ($W_{n_{\text{test}} \times k}$) 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 $(n_{\text{test}} \times 1)$ 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 χ , where χ 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 χ , 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. (30):

$$p = 1 - \Pr(l_\chi \leq A\mathbb{Z} \leq u_\chi) = 1 - \int_{l_\chi}^{u_\chi} \phi(\mathbb{Z}; 0, \Delta) d\mathbb{Z}, \quad (31)$$

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

The Δ in Eq. (31) is a positive definite variance–covariance matrix. The upper or lower triangle of Δ consists of $q (= n_{\text{test}}(n_{\text{test}} - 1)/2)$ 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 probability density function of the MAX-test.

5 Application to the Real Data

5.1 Genotype Data

The genotype data from the Nagahama Study [10, 23] of 1813 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.

Multiple cutoff χ^2 -values were used by selecting various cutoff χ -values from (0.01, 8.5) by increasing by 0.1 along with the GWAS cutoff of 5.45 ($\chi^2 = 5.45^2$, and $p = 5 \times 10^{-08}$). For each subset, the calculation was repeated 1,000 times, and 1,000 Monte Carlo samples were used for each repetition. More detailed data description is provided as Online Resource in the Supplementary File, ESM_1.pdf.

5.2 Results

The estimated p -values for the two genes with their individual LD structures are presented in Fig. 2. For example, the gene *FES* in Fig. 2a has a relatively small number of SNPs with quite strong pairwise LD structures, compared to the gene *NCSI* in Fig. 2b. 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. The p -values are highly deviated from the $df = 1$ test for the gene *NCSI*, whereas the gene *FES*, is closer to the values of $df = 1$ (Fig. 2c). 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 $p = 5 \times 10^{-08}$ ($\chi^2 = 5.45^2$) for all the χ^2 -values along the horizontal axis of Fig. 2c, 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 (Fig. 2). 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 Fig. 2a will be higher than that of the gene in Fig. 2b.

Figure 3a is presenting the results from the evaluation of all the gene-based SNP subsets (N subsets) for multiple cutoff χ^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 χ^2 . The shape of the distributions and their modality at the left tails 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. Figure 3a demonstrates how the distribution function shapes vary with multiple cutoff χ^2 -values and the changes in the shape is not simple. Therefore, it is advisable

Fig. 2 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 χ^2 -values (Color figure online)

to apply our method for individual gene sets by considering the corresponding LD structures and for target χ^2 -values individually.

To visualize the variations in p -values more obviously, only the two cutoff χ^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 Fig. 3b. 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 χ^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 Fig. 3c and d, 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 Fig. 3c 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 Fig. 3d.

Fig. 3 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 χ^2 -cutoff. **b** The distributions of $\log_{10}(p)$ -values for two cutoff χ^2 -values, 2.51² and 3.91², 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 χ^2 -cutoffs, **c** 2.51² and 3.91². **d** 2.51² and 7.51² (Color figure online)

These results show that the p -values are heterogeneous for the gene-based SNP subsets over the genome for multiple cutoff χ^2 -values. Also, they are not uniform for a fixed cutoff. For example, the shape of each colored subpanel of Fig. 3a 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 [15]. 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.

6 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. However, 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 $df = 1$ 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) [5, 8] with the squared Euclidean distance as a measure of χ^2 in df -dimensional space. This χ^2 -distance is a standardized form of the Mahalanobis distance [17]. A detailed illustration of the relationship among the CA, Mahalanobis distance and Euclidean distance is provided as Online Resource in the Supplementary File, ESM_2.pdf.

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 (Supplementary File, ESM_3.pdf) [15, 16] 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. However, this method is not suitable for our Sph-Btv approach because the Botev method [2] 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 (Supplementary File, ESM_4.eps). In ESM_4.eps, Fig. (a) is presenting the contour hyperplanes of the tables for a 2×3 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 χ^2 -values, which means that the contour hyperplane contains many straight segments. Since, there are many segments, overall they look like “curved”. Also, in ESM_4.eps, Fig. (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.

In summary, we demonstrated 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.

Acknowledgements The authors thank all the members of Statistical Genetics and the members of Human Disease Genomics, the Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan. This work was supported by the Core Research for Evolutional Science and Technology (CREST) (Grant No. 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 Nos. 25293141, 26670313, 26293198, 17H04182, 17H04126) from the Ministry of Education, Culture, Sports, Science and Technology in Japan (MEXT); the Practical Research Project for Rare/Intractable Diseases (Grant Nos. ek0109070, ek0109196, ek0109348); the Comprehensive Research on Aging and Health Science Research Grants for Dementia R&D (Grant No. dk0207006); the Program for an Integrated Database of Clinical and Genomic Information (Grant No. kk0205008); the Practical Research Project for Life-Style related Diseases including Cardiovascular Diseases and Diabetes Mellitus (Grant No. ek0210066, ek0210096) from Japan Agency for Medical Research and Development (AMED), and the Takeda Medical Research Foundation.

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 approach are available free online via a GitHub repository: https://github.com/statgenetJimu/Spherization/blob/master/pvalueCalculation_Botev.R. The algorithm and the codes are also within the Supplementary Files (ESM_5.pdf, ESM_6.pdf).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bender R, Lange S (2001) Adjusting for multiple testing—when and how? *J Clin Epidemiol* 54:343–349. [https://doi.org/10.1016/S0895-4356\(00\)00314-0](https://doi.org/10.1016/S0895-4356(00)00314-0)
- Botev ZI (2017) The normal law under linear restrictions: simulation and estimation via minimax tilting. *J R Stat Soc B* 79:125–148. <https://doi.org/10.1111/rssb.12162>
- Bush WS, Moore JH (2012) Chapter 11: genome-wide association studies. *PLOS Comput Biol* 8(12):e1002822. <https://doi.org/10.1371/journal.pcbi.1002822>
- Chen SY, Feng Z, Yi X (2017) A general introduction to adjustment for multiple comparisons. *J Thorac Dis* 9:1725–1729. <https://doi.org/10.21037/jtd.2017.05.34>
- Cuadras CM, Cuadras D (2006) A parametric approach to correspondence analysis. *Linear Algebra Appl* 417:64–74. <https://doi.org/10.1016/j.laa.2005.10.029>
- Gao X, Starmer J, Martin ER (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol* 32:361–369. <https://doi.org/10.1002/gepi.20310>
- González JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V (2008) Maximizing association statistics over genetic models. *Genet Epidemiol* 32:246–254. <https://doi.org/10.1002/gepi.20299>
- Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic Press, London
- Hirosawa K, Kawaguchi T, Matsuda F, Yamada R (2010) Estimation of p-value of MAX test with double triangle diagram for 2×3 SNP case-control tables. *Genet Epidemiol* 34:543–551. <https://doi.org/10.1093/nar/gkt1229>
- Higasa K, Miyake N, Yoshimura J et al (2016) Human genetic variation database, a reference database of genetic variations in the Japanese population. *J Human Genet* 61:547–553. <https://doi.org/10.1038/jhg.2016.12>
- Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* 11:2115–2119. <https://doi.org/10.1101/gr.204001>
- Hsu JC (1996) Multiple comparisons: theory and methods. Chapman & Hall, CRC Press, London
- 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 Genom* 11:724. <https://doi.org/10.1186/1471-2164-11-724>
- Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I (2009) Fewer permutations, more accurate p-values. *Bioinformatics* 25:i161–i168. <https://doi.org/10.1093/bioinformatics/btp211>
- Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S (2016) Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLOS Comput Biol* 12(1):e1004714. <https://doi.org/10.1371/journal.pcbi.1004714>
- Li MX, Gui HS, Kwan JSH et al (2011) GATES: a rapid and powerful gene-based association test using extended simes procedure. *Am J Hum Genet* 88:283–293. <https://doi.org/10.1016/j.ajhg.2011.01.019>
- Mahalanobis PC (1936) On the generalized distance in statistics. *Proc Natl Inst Sci India* 2:49–55
- McDonald JH (2014) Handbook of biological statistics. Sparky House Publishing, Baltimore
- Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 74:765–769. <https://doi.org/10.1086/383251>
- 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. *Am J Hum Genet* 69:138–147. <https://doi.org/10.1086/321276>
- Slavković AB, Fienberg SE (2010) Algebraic and geometric methods in statistics. In: Gibilisco P, Riccomagno E, Rogantin MP, Wynn HP (eds) Algebraic geometry of 2×2 contingency tables. Cambridge University Press, Cambridge, pp 63–81
- Yamada R, Okada Y (2009) An optimal dose-effect mode trend test for SNP genotype tables. *Genet Epidemiol* 33:114–127. <https://doi.org/10.1002/gepi.20362>
- Yoshimura K, Nakayama T, Sekine A et al (2012) B-type natriuretic peptide as an independent correlate of nocturnal voiding in Japanese women. *Neurourol Urodyn* 31:1266–1271. <https://doi.org/10.1002/nau.22250>