

Entropy-Based Information Gain Approaches to Detect and to Characterize Gene-Gene and Gene-Environment Interactions/Correlations of Complex Diseases

R Fan¹, M Zhong², S Wang^{1,3}, Y Zhang¹, A Andrew⁴, M Karagas⁴,
H Chen⁵, CI Amos⁶, M Xiong⁷, J Moore⁸

Department of Statistics¹, Texas A&M University, College Station, TX 77843

Global Pharmaceutical Research and Development², Abbott Laboratories
100 Abbott Park Rd, R436 AP9A-1, Abbott Park, IL 60064

School of Information Science and Engineering³
Yunnan University, Kunming 650091, P. R. China

Department of Community and family Medicine⁴, Department of Genetics⁸
Dartmouth Medical School, Lebanon, NH 03756

Surveillance Research Program⁵, National Cancer Institute
6116 Executive Blvd. #5016, Rockville, Maryland 20892

Department of Epidemiology⁶, MD Anderson Cancer Center
University of Texas, Houston, TX 77030

Human Genetics Center⁷, University of Texas, P. O. Box 20334, Houston, Texas 77225

Short title: Gene-gene and Gene-Environment Interactions of Complex Diseases
Correspondence to: Dr. Ruzong Fan, Tel. 979-845-3152 or 3141
(main office), Fax 979-845-3144, E-mail: rfan@stat.tamu.edu

April 3, 2011

Summary

For complex diseases, the relationship between genotypes/environment-factors and phenotype is usually complex and nonlinear. The entropy-based approach of information theory is likely to be very useful to identify and to interpret the gene-gene and gene-environment interactions/correlations of complex diseases. In this paper, we develop entropy-based approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. Such as variance partitioning and ANOVA of linear statistical models, we develop entropy decomposition approaches for model selection, i.e, to select important genetic and environmental factors which interactively/nonlinearly influence the development of complex diseases. For 2-way interactions and correlations, an information gain (IG) approach is proposed using mutual information. The information gain in the presence of disease is defined as a one-dimensional variable through mutual information and entropy function of two genetic markers or one marker and an environmental factor. Based on the one-dimensional information gain, a test statistic T_{IG} is constructed and is showed to be χ^2_1 -distributed with 1 degree of freedom. For 3-way and higher order interactions, an interaction information gain (IIG) based approach is proposed; and for 3-way and higher order correlations, a total correlation information gain (TCIG) based approach is proposed. Such as the 2-way case, the IIG and TCIG are defined as one-dimensional variables. The related test statistics T_{IIG} and T_{TCIG} are constructed to test 3-way or higher order interactions and correlations, respectively. One advantage of the proposed method is that it collapses high-dimensional genetic data into a single dimension, and this makes it possible to build test statistics for sparse data to detect and to characterize gene-gene and gene-environment interactions/correlations. Compared with the naive χ^2 test statistics which usually have high degrees of freedom, the proposed information gain tests can have higher or similar power; in addition, the naive χ^2 tests are not always implementable due to sparse nature of genetic data. The proposed information gain tests are reasonably robust and conservative. We apply the methods to analyze bladder cancer data to investigate the complex interactions between DNA repair gene SNPs, smoking status, and bladder cancer susceptibility. We use the bladder cancer data to show a forward selection procedure for the final model selection.

Key Words: gene-gene and gene-environment interactions/correlations, entropy, mutual information, interaction information, total correlation information.

1 Introduction

Complex diseases are consequences of mutual interactions among genetic variants and environmental factors. Although there has been much effort to dissect complex traits, the identification and characterization of susceptibility genes of common complex human diseases remains a great challenge for human geneticists. It is challenging both conceptually and technically. Conceptually, it is not always clear how to define the interactions. There are two arguments about gene-gene and gene-environment interactions: (1) statistical interaction, (2) biological interaction.

In traditional statistical models, i.e., linear models and generalized linear models such as logistic regressions, the genetic and environmental effects are decomposed into main linear effects and interaction effects.¹ The statistical interactions of genetic variants and environmental effects are deviation from the main linear effects. In the absence of the main linear effects, the statistical interactions does not make sense. Moreover, the traditional statistical models may not work for high dimension sparse data. For instance, logistic regression models which include interaction terms may fail to converge as reported in Andrew et al. (2006)² for bladder cancer data possibly due to small number of individuals in some cells. One advantage of using traditional statistical models to analyze genetic data is that the related theory is very mature and the user-friendly softwares are available. For instance, variance partitioning and ANOVA are standard procedure in SAS for data analysis and model selection.

Biological interaction, on the other hand, happens at the cellular level in an individual and it is the results of physical interactions of biomolecules such as DNA, RNA and proteins.³⁻⁵ The biological gene-gene and gene-environment interaction is the interdependence of genetic and environmental factors that may cause complex diseases. The relationship between genotypes/environment-factors and disease phenotypes is usually complex and is nonlinear for complex diseases. Thus, biological interaction makes sense and it is valid in describing the complicated relation between genetic/environmental factors and disease phenotypes. In the absence of main effects, the biological gene-gene and gene-environment interactions may exist and can be significant and important.⁶ However, the related theory to detect and to characterize the biological gene-gene and gene-environment interactions is not well-developed. There is a need to develop powerful methods and user-friendly softwares to identify and to interpret the complex genetic architecture and nonlinear biological interactions of complex traits.

In recent years, there has been great enthusiasm to detect and to characterize gene-gene and gene-environment interactions of complex diseases using genome data.^{5,7-10} By using multiple genetic markers and environmental factors in analysis, it is usually a high-dimensional problem. For instance, assume we have two single nucleotide polymorphism (SNP) markers. Each of the two SNP has 3 genotypes, and then there are 9 genotype combinations if we consider the two SNPs simultaneously. If we add one environmental factor which has 2 categories, e.g., smoking vs non-smoking, there are $2 \times 9 = 18$ genotype-environment combinations if we consider the two SNPs and the environmental factor simultaneously. Hence, one needs to handle the high-dimensional data.

In the multifactor dimensionality reduction (MDR) procedure, high dimensional genetic data are collapsed into a single dimensional variable and this makes it possible to analyze high-dimensional sparse genetic data.¹¹⁻¹⁷ One may want to notice that MDR is a non-parametric procedure and it makes no assumption of linear relationship between the phenotypes and the genetic/environmental factors. Since there was no alternative and powerful procedure, Andrew et al. (2006)² ran logistic regression models to test three way interaction to replicate the findings of MDR. Unfortunately, the logistic regression models failed to converge due to the sparse nature of bladder cancer data. Thus, it is not only interesting but also necessary and important to develop novel statistical methods to detect and to characterize the complex biological gene-gene and gene-environment interactions of complex traits.

Technically, traditional statistical approaches may not be useful because of the complexity and non-linearity between complex traits and genetic/environment-factors. The traditional statistical models can not properly fit the nonlinear relationship between genotypes/environment-factors and disease phenotypes in the absence of main effects, and it may not be necessarily able and useful to model biological interactions. For the bladder cancer data of Andrew et al. (2006)², the main effects of genetic polymorphisms were not observed and it is unclear if logistic regressions may fit the data well. The failure of convergence may be due to invalidness of the logistic regression model itself.

It is well-known that information theory based on entropy function is widely used to study nonlinear problems and complex system. The entropy function is a nonlinear transformation of interested variables. The entropy is commonly used in information theory to measure the uncertainty of random variables. The entropy-based approach is likely to be very useful to study the nonlinear relationship between

genotypes/environment-factors and phenotypes and to interpret the gene-gene and gene-environment interactions of complex diseases.¹⁸⁻²⁰ In this article, we develop entropy-based approaches to detect and to characterize gene-gene and gene-environment interactions of complex diseases. We start with the definition of entropy for genetic markers and environmental factors. Then, we introduce 2-way mutual information and information gain (IG) to describe gene-gene and gene-environment interactions, and we construct test statistics to detect the 2-way interactions.

One idea of this article is to reduce high dimensional data to be a one-dimensional variable, and then to construct a χ^2 -distribution statistic to test gene-gene interaction of complex diseases. We considered two di-allelic markers A and B in a case-control design. Correspondingly, there are 9 genotype combinations for the two markers. By using the information gain function, we reduce the 9-dimensional data to be a one-dimensional variable to construct the information gain based test T_{IG} . The method can be applied to test 2-way gene-environment interaction. To test interaction between a di-allelic marker and an environmental factor, we may reformulate the problem. For instance, if one di-allelic marker and a two-category environmental factor (e.g., smoking and non-smoking) are concerned, then 6-dimensional data instead of 9-dimensional data need to be reduced to a one-dimensional variable via information gain to construct the test statistics.

To generalize the 2-way methods to handle 3-way and multiple K -way cases, we need to distinguish two different concepts in information theory: interaction information and total correlation information (TCI). In 2-way case, the two concepts are the same. However, they are different in 3-way and multiple K -way cases. Roughly, the interaction information among multiple factors is the amount of information that is common to all the factors. The total correlation information, however, describes the total amount of dependence among all the factors. The 3-way total correlation information can be decomposed into a summation of all 2-way mutual information and the 3-way interaction information. Similarly, the K -way total correlation information can be decomposed into a summation of all lower and same order k -way interaction information, $k \leq K$.²³

We generalize the 2-way methods to detect and to characterize 3-way and multiple K -way gene-gene/gene-environment interactions and correlations. For 3-way and multiple K -way interactions, the 3-way and K -way interaction information is proposed to extend the 2-way mutual information; and

for 3-way and multiple K-way correlations, total correlation information is used. Correspondingly, the interaction information gain (IIG) and total correlation information gain (TCIG) can be defined as one-dimensional variables for case-control data. Thus, high-dimensional genetic data are collapsed as one-dimension variables via the information gains. Using the one-dimensional variables, test statistics are constructed which are χ_1^2 -distributed. Compared with the naive χ^2 test statistics which usually have high degrees of freedom, the proposed information gain tests are easy to implement and the naive χ^2 tests are not always implementable due to sparse nature of high dimension genetic data.

Simulation study is performed to evaluate the robustness of the test statistics by type I error rate calculations. In addition, power analysis is carried out to show the usefulness of the proposed methods. The method is applied to bladder cancer data to explore gene-gene and gene-environment interactions and correlations of SNPs and smoking status with the disease.² We use the bladder cancer data to show a forward selection procedure for the final model selection, and the procedure can be applied to the study of other complex traits.

2 Methods

Entropy, defined by Shannon in 1948²¹, is a measure of the uncertainty of a random variable/system. The entropy $H(X)$ of a discrete random variable X is defined by²²

$$H(X) = -E[\log P(X)] = -\sum_{x \in \mathcal{X}} p(x) \log p(x), \quad (1)$$

where $p(x) = P(X = x)$, $x \in \mathcal{X}$, is the probability mass function of the random variable X , and \mathcal{X} is a finite or enumerable infinite set such as $\{1, 2, \dots, n\}$ or $\{1, 2, \dots\}$. The log is to the base 2. By definition, $0 \log 0 = 0$. The higher the entropy $H(X)$, the less reliable we may predict the outcome about the variable X . The concept of the Shannon entropy was used to select interesting combinations of polymorphisms for evaluating and for visualizing the information gain to detect gene-gene and gene-environment interactions²³⁻²⁸.

2.1 Genotype and Environment Based Entropy

For case-control data, let us denote the disease status by D , i.e., $D = 0$ means normal and $D = 1$ means affected. For the purpose of explanation, consider two di-allelic markers A and B such as SNP data and

an environment exposure E . Let us denote the two alleles of marker A by A and a ; and denote the two alleles of marker B by B and b , respectively. At the marker A , there are three genotypes AA, Aa, aa ; similarly, there are three genotypes BB, Bb, bb at the marker B . For the environmental factor E , let us code it as $E = 0, 1, 2$ (e.g., non-smoking, < 35 pack-years, ≥ 35 pack-years). For the convenience of presentation, let us code the genotypes at marker A by G_A by counting the number of allele A and similarly code the genotypes at marker B by G_B , i.e.,

$$G_A = \begin{cases} 2 & AA \\ 1 & Aa \\ 0 & aa \end{cases} \quad \text{and} \quad G_B = \begin{cases} 2 & BB \\ 1 & Bb \\ 0 & bb \end{cases}. \quad (2)$$

Notice that markers A and B and the environmental factor E are treated as attribute in literature.^{24,26} In practice, environmental factor can be replaced by a genetic marker, and then the explanation needs to change, accordingly. Using the entropy definition (1), we may define the entropy $H(A)$ of marker A in the general population and the conditional entropy $H(A|D)$ in the disease population as

$$\begin{aligned} H(A) &= - \sum_{i=0}^2 P(G_A = i) \log P(G_A = i), \\ H(A|D) &= - \sum_{i=0}^2 P(G_A = i|D = 1) \log P(G_A = i|D = 1). \end{aligned} \quad (3)$$

Under the Hardy-Weinberg equilibrium, we may calculate that $P(G_A = 0) = P_a^2, P(G_A = 1) = 2P_AP_a, P(G_A = 2) = P_A^2$, where P_A and P_a are allele frequencies of marker A . In this paper, however, we don't need to assume the Hardy-Weinberg equilibrium. Similarly, we may define the entropy $H(B)$ of marker B in the general population and the conditional entropy $H(B|D)$ in the disease population. For the environmental factor E , its entropy $H(E)$ in the general population and its conditional entropy $H(E|D)$ in the disease population can be defined, accordingly.

Combining both markers A and B , we may define the entropy $H(A, B)$ in the general population and the conditional entropy $H(A, B|D)$ in the disease population as

$$\begin{aligned} H(A, B) &= - \sum_{i=0}^2 \sum_{j=0}^2 P(G_A = i, G_B = j) \log P(G_A = i, G_B = j), \\ H(A, B|D) &= - \sum_{i=0}^2 \sum_{j=0}^2 P(G_A = i, G_B = j|D = 1) \log P(G_A = i, G_B = j|D = 1). \end{aligned} \quad (4)$$

Combining marker A and environmental factor E , we may define the entropy $H(A, E)$ in the general population and the conditional entropy $H(A, E|D)$ in the disease population as

$$H(A, E) = - \sum_{i=0}^2 \sum_{e=0}^2 P(G_A = i, E = e) \log P(G_A = i, E = e),$$

$$H(A, E|D) = - \sum_{i=0}^2 \sum_{e=0}^2 P(G_A = i, E = e|D = 1) \log P(G_A = i, E = e|D = 1). \quad (5)$$

The entropy $H(B, E)$ in the general population and the conditional entropy $H(B, E|D)$ in the disease population can be defined, accordingly. Combining both markers A and B and environmental factor E , we may define the entropy $H(A, B, E)$ in the general population and the conditional entropy $H(A, B, E|D)$ in the disease population as

$$\begin{aligned} H(A, B, E) &= - \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 P(G_A = i, G_B = j, E = e) \log P(G_A = i, G_B = j, E = e), \\ H(A, B, E|D) &= - \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 P(G_A = i, G_B = j, E = e|D = 1) \log P(G_A = i, G_B = j, E = e|D = 1). \end{aligned} \quad (6)$$

2.2 2-Way Mutual Information and Information Gain

In the general population, the mutual information of markers A and B is defined as^{21,22}

$$\begin{aligned} I(A, B) &= H(A) + H(B) - H(A, B) \\ &= \sum_{i=0}^2 \sum_{j=0}^2 P(G_A = i, G_B = j) \log \frac{P(G_A = i, G_B = j)}{P(G_A = i)P(G_B = j)}. \end{aligned} \quad (7)$$

$I(A, B)$ measures the dependency or correlation between A and B . For any two markers A and B , $I(A, B) \geq 0$ and $I(A, B) = 0$ if and only if G_A and G_B are independent [i.e., $P(G_A = i, G_B = j) = P(G_A = i)P(G_B = j)$, p28 of Cover and Thomas (2006)²²].

Such as Venn diagram in set theory, Figure 1 a) shows an I-diagram of $H(A)$, $H(B)$ and $I(A, B)$ for two markers A and B in information theory.^{29,31,33} In Figure 1 a), the black region corresponds to the magnitude of the $I(A, B)$, the magnitude of $H(A)$ corresponds to the left rectangle and the magnitude of $H(B)$ corresponds to the right one, and each of the two rectangles includes the black region as the common part. In the disease population, the mutual information of markers A and B is defined as

$$\begin{aligned} I(A, B|D) &= H(A|D) + H(B|D) - H(A, B|D) \\ &= \sum_{i=0}^2 \sum_{j=0}^2 P(G_A = i, G_B = j|D = 1) \log \frac{P(G_A = i, G_B = j|D = 1)}{P(G_A = i|D = 1)P(G_B = j|D = 1)}. \end{aligned} \quad (8)$$

$I(A, B|D)$ measures the interaction between markers A and B given the disease. For any two markers A and B , $I(A, B|D) \geq 0$ and $I(A, B|D) = 0$ if and only if G_A and G_B are conditional independent given the disease (i.e., $P(G_A = i, G_B = j|D = 1) = P(G_A = i|D = 1)P(G_B = j|D = 1)$).

Following a few previous studies, the information gain of markers A and B in the presence of disease D can be defined as the difference^{24–27,29}

$$IG(AB | D) = I(A, B|D) - I(A, B). \quad (9)$$

If the disease and the two markers are independent (i.e., $P(G_A = i, G_B = j|D = 1) = P(G_A = i, G_B = j)$), $I(A, B|D) = I(A, B)$ and so their difference or information gain $IG(AB | D)$ is equal to 0. Hence, we may test the gene-gene interaction between the disease and the two markers A and B by testing if the difference is zero. Based on this rationale, we may build test statistics for practical application.

For marker A and environmental factor E , the mutual information $I(A, E)$ in the general population and the conditional mutual information $I(A, E|D)$ in the disease population can be defined along the lines above. Similarly, we may define the mutual information $I(B, E)$ in the general population and the conditional mutual information $I(B, E|D)$ in the disease population for marker B and environmental factor E . The information gain of marker A and environmental factor E in the presence of disease D can be defined as the difference $IG(AE | D) = I(A, E|D) - I(A, E)$. If the marker A and the environmental factor E are independent of the disease status D , then there is no information gain, i.e., $IG(AE | D) = 0$. Similarly, if the marker B and the environmental factor E are independent of the disease status D , then there is no information gain, i.e., $IG(BE | D) = I(B, E|D) - I(B, E) = 0$.

2.3 3-Way Interaction Information and Total Correlation Information

The information gains $IG(AB | D)$, $IG(AE | D)$ and $IG(BE | D)$ represent 2-way interaction gain of two attributes given the disease. If we consider the three attributes A , B and E simultaneously, we may define 3-way interaction information gain and total correlation information as follows.^{29–33} In the general population, the 3-way interaction information of markers A and B and environmental factor E is defined as (Cover and Thomas 2006²², p49).

$$I(A, B, E) = -H(A) - H(B) - H(E) + H(A, B) + H(A, E) + H(B, E) - H(A, B, E).$$

The 3-way interaction information $I(A, B, E)$ contains interactions that can not be explained by the 2-way mutual information $I(A, B)$, $I(A, E)$, and $I(B, E)$. It represents the gain or loss of information by adding one attribute to a pair of attributes. Hence, the 3-way interaction information among attributes

A , B and E can be understood as the amount of information that is common to all the attributes, but not present in any subset. The interaction information may be negative or positive.

Figure 1 b) shows an I-diagram of $H(A)$, $H(B)$, $H(E)$, and $I(A, B, E)$ for two markers A and B and an environment factor E .^{29,31,33} In Figure 1 b), the black region corresponds to the magnitude of the $I(A, B, E)$, and we add a rectangle to represent the magnitude of $H(E)$ compared with Figure 1 a). If one attribute is independent of the other two which are dependent, the interaction information $I(A, B, E)$ will be 0. For instance, if the genetic markers A and B are independent of the environment factor E but A and B are dependent or in linkage disequilibrium, the interaction information $I(A, B, E)$ will be zero. To show this, notice that $P(G_A = i, G_B = j, E = e) = P(G_A = i, G_B = j)P(E = e)$ implies $P(G_A = i, E = e) = P(G_A = i)P(E = e)$ and $P(G_B = j, E = e) = P(G_B = j)P(E = e)$, and then it can be seen $I(A, B, E) = 0$. Certainly, if all the three attributes are independent, the interaction information $I(A, B, E)$ is equal to 0. Hence, $I(A, B, E)$ is an interaction among all three attributes.

The total correlation information is defined as the difference of joint entropy $H(A, B, E)$ and the three individual entropies $H(A)$, $H(B)$ and $H(E)$, i.e.,

$$\begin{aligned} TCI(A, B, E) &= H(A) + H(B) + H(E) - H(A, B, E) \\ &= I(A, B) + I(A, E) + I(B, E) + I(A, B, E). \end{aligned} \quad (10)$$

The total correlation information describes the total amount of dependence among the three attributes A , B and E . It is always positive, or zero if and only if all the three attributes are independent, i.e., $P(G_A = i, G_B = j, E = e) = P(G_A = i)P(G_B = j)P(E = e)$. It will not be zero even if only a pair of attributes are dependent. For instance, if the genetic markers A and B are independent of the environment factor E but A and B are dependent or in linkage disequilibrium, the total correlation information $TCI(A, B, E)$ will be non-zero. Figure 1 c) shows an I-diagram of $H(A)$, $H(B)$, $H(E)$, and $TCI(A, B, E)$ for two markers A and B and an environment factor E .^{29,31,33}

The second equality of relation (10) shows that the total correlation information $TCI(A, B, E)$ is equal to the summation of all 2-way mutual information $I(A, B)$, $I(A, E)$, $I(B, E)$, and 3-way interaction information $I(A, B, E)$. Thus, the 2-way mutual information and the 3-way interaction information can be seen as a decomposition of a 3-way dependency into a sum of 2-way and 3-way interactions.²³ The existence of 3-way correlations indicate the existence of some 2-way or 3-way interactions. On the other

hand, the existence of 2-way or 3-way interactions can lead to 3-way correlations.

In the disease population, the 3-way interaction information $I(A, B, E|D)$ and total correlation information $TCI(A, B, E|D)$ of markers A and B and environmental factor E are defined as

$$\begin{aligned} I(A, B, E|D) &= -H(A|D) - H(B|D) - H(E|D) \\ &\quad + H(A, B|D) + H(A, E|D) + H(B, E|D) - H(A, B, E|D), \\ TCI(A, B, E|D) &= H(A|D) + H(B|D) + H(E|D) - H(A, B, E|D). \end{aligned}$$

The interaction information gain $IIG(ABE | D)$ and total correlation information gain $TCIG(ABE | D)$ of markers A and B and the environmental factor E in the presence of disease D can be defined as the differences^{24–27,29}

$$\begin{aligned} IIG(ABE | D) &= I(A, B, E|D) - I(A, B, E), \\ TCIG(ABE | D) &= TCI(A, B, E|D) - TCI(A, B, E). \end{aligned} \quad (11)$$

If the disease is independent of the two markers and the environmental factor E , $I(A, B, E|D) = I(A, B, E)$ and so their difference or information gain $IIG(ABE | D)$ is equal to 0; similarly, we have $TCI(A, B, E|D) = TCI(A, B, E)$ and so their difference $TCIG(ABE | D)$ is equal to 0. Hence, we may test the gene-gene and gene-environment interactions/correlations between the disease and two markers A and B and the environmental factor E by testing if the differences are zero. Based on this rationale, we may build test statistics for practical application.

2.4 K-Way Interaction Information and Total Correlation Information

Suppose that we are interested in interactions/correlations between the disease and an arbitrary number K of attributes $\mathcal{A} = (A_1, \dots, A_K)$, which can be genetic markers or environmental factors. For simplicity, we assume that each A_i can take three values 0, 1, 2. For a vector of realization $\vec{a} = (a_1, \dots, a_K)$ of $\mathcal{A} = (A_1, \dots, A_K)$, denote the joint probabilities as $P_{\vec{a}} = P(A_1 = a_1, \dots, A_K = a_K) = P_{a_1 \dots a_K}$ in the general population and $Q_{\vec{a}} = P(A_1 = a_1, \dots, A_K = a_K | D = 1) = Q_{a_1 \dots a_K}$ in the disease population. Based on the joint probabilities, we may define the entropies $H(\mathcal{A}) = H(A_1, \dots, A_K) = -\sum_{\vec{a}} P_{\vec{a}} \log P_{\vec{a}}$ and $H(\mathcal{A} | D) = H(A_1, \dots, A_K | D) = -\sum_{\vec{a}} Q_{\vec{a}} \log Q_{\vec{a}}$.

For a subset $\mathcal{S} = (A_{i_1}, A_{i_2}, \dots, A_{i_n}) \subseteq \mathcal{A} = (A_1, A_2, \dots, A_K)$, we may define the related entropies $H(\mathcal{S})$ and $H(\mathcal{S} | D)$, accordingly. Here \subseteq means that \mathcal{S} is a subset of \mathcal{A} and it can be equal to $\mathcal{A} = (A_1, A_2, \dots, A_K)$. For a realization \vec{s} of $\mathcal{S} = (A_{i_1}, A_{i_2}, \dots, A_{i_n})$, the marginal probabilities are denoted as $P_{\vec{s}}$ and $Q_{\vec{s}}$. For individual attributes A_1, \dots, A_K , the marginal probabilities and entropies are denoted as $P_{a_1, \dots, \dots}, \dots, P_{\dots, \dots, a_K}, H(A_1), \dots, H(A_K), Q_{a_1, \dots, \dots}, \dots, Q_{\dots, \dots, a_K}, H(A_1 | D), \dots, H(A_K | D)$. For a subset \mathcal{S} , let us denote $|\mathcal{S}| = |(A_{i_1}, A_{i_2}, \dots, A_{i_n})| = n$, i.e., the number of attributes of \mathcal{S} . The K-way interaction information can be defined as^{29,32,33}

$$\begin{aligned} I(\mathcal{A}) &= I(A_1, A_2, \dots, A_K) = - \sum_{\mathcal{S} \subseteq \mathcal{A}} (-1)^{|\mathcal{A}|-|\mathcal{S}|} H(\mathcal{S}), \\ I(\mathcal{A} | D) &= I(A_1, A_2, \dots, A_K | D) = - \sum_{\mathcal{S} \subseteq \mathcal{A}} (-1)^{|\mathcal{A}|-|\mathcal{S}|} H(\mathcal{S} | D). \end{aligned}$$

The K-way interaction information gain is defined as $IIG(\mathcal{A} | D) = I(\mathcal{A} | D) - I(\mathcal{A})$.

In the general population, the K-way total correlation information is defined as the difference of the summation of individual entropies $H(A_1), \dots, H(A_K)$ and the joint entropy $H(\mathcal{A})$,^{23,30,31} i.e.,

$$\begin{aligned} TCI(\mathcal{A}) &= H(A_1) + \dots + H(A_K) - H(A_1, \dots, A_K) \\ &= \sum_{\mathcal{S} \subseteq \mathcal{A}, |\mathcal{S}| \geq 2} I(\mathcal{S}). \end{aligned} \tag{12}$$

The total correlation information $TCI(\mathcal{A})$ is the total amount of dependence among all the attributes $\mathcal{A} = (A_1, \dots, A_K)$. As the second equality in relation (12) shows, the total correlation information $TCI(\mathcal{A})$ is equal to the summation of all interaction information $I(\mathcal{S})$ including 2-way mutual information, $\mathcal{S} \subseteq \mathcal{A}$. Thus, the interaction information can be seen as a decomposition of a K -way dependency into a sum of k -way interactions, $k \leq K$.²³ The existence of K -way correlations indicate the existence of some k -way interactions, $k \leq K$. On the other hand, the existence of low order k -way interactions can lead to high order K -way correlations.

In the disease population, the K-way total correlation information is defined as the difference of the summation of individual entropies $H(A_1 | D), \dots, H(A_K | D)$ and the joint entropy $H(\mathcal{A} | D)$, i.e.,

$$TCI(\mathcal{A} | D) = H(A_1 | D) + \dots + H(A_K | D) - H(A_1, \dots, A_K | D).$$

The K-way total correlation information gain is defined as $TCIG(\mathcal{A} | D) = TCI(\mathcal{A} | D) - TCI(\mathcal{A})$. If the disease is independent of the attributes, the interaction information gain $IIG(\mathcal{A} | D)$ is equal to 0

and similarly, the total correlation information gain $TCIG(\mathcal{A} | D)$ is equal to 0. The test statistics can be built accordingly to test the interaction between the disease and the attributes $\mathcal{A} = (A_1, A_2, \dots, A_K)$.

2.5 Test Statistics Based on the 2-Way Mutual Information Gain

Based on the discussion in the Subsection 2.2 about mutual information and information gain, we may construct test statistics to detect gene-gene and gene-environment interaction. In the following presentation, we only discuss the construction of test statistic to detect gene-gene interaction between markers A and B . The procedure can be accordingly applied to construct test statistic to detect gene-environment interaction between marker A (or B) and environment-factor E .

Consider a case-control design with M controls from an unaffected population and N cases from an affected population. Assume each individual in the sample is typed at both markers A and B . Let us denote by X_{ij} the count of controls whose genotypes are $(G_A = i, G_B = j)$; and similarly denote by Y_{ij} the count of cases whose genotypes are $(G_A = i, G_B = j)$, $i, j = 0, 1, 2$. The test statistics can be built based on the column vectors $X = (X_{00}, X_{01}, X_{02}, X_{10}, X_{11}, X_{12}, X_{20}, X_{21})^\tau$ and $Y = (Y_{00}, Y_{01}, Y_{02}, Y_{10}, Y_{11}, Y_{12}, Y_{20}, Y_{21})^\tau$. Hereafter, the superscript τ denote the transpose of a vector/matrix. In addition, X_{22} is not included in X , and Y_{22} is not included in Y to remove the redundancy. Before defining our test statistics, let us introduce some notations.

In the general population, denote the joint genotype probabilities for markers A and B by $P_{ij} = P(G_A = i, G_B = j)$; and for the disease population, denote the joint conditional genotype probabilities for markers A and B by $Q_{ij} = P(G_A = i, G_B = j | D = 1)$. It is easy to see that P_{ij} sums to 1 and Q_{ij} does so, respectively; and so the redundancy of parameters exists. Let us denote $P = (P_{00}, P_{01}, P_{02}, P_{10}, P_{11}, P_{12}, P_{20}, P_{21})^\tau$, and $Q = (Q_{00}, Q_{01}, Q_{02}, Q_{10}, Q_{11}, Q_{12}, Q_{20}, Q_{21})^\tau$. The column counting vector $\begin{pmatrix} X \\ X_{22} \end{pmatrix}$ follows a multinomial distribution $\left(M, \begin{pmatrix} P \\ P_{22} \end{pmatrix}\right)$; and the column counting vector $\begin{pmatrix} Y \\ Y_{22} \end{pmatrix}$ follows a multinomial distribution $\left(N, \begin{pmatrix} Q \\ Q_{22} \end{pmatrix}\right)$. The mean vector of X is MP , and the mean vector of Y is NQ , respectively. The variance-covariance matrix of X is $M[\text{diag}(P) - PP^\tau]$, and the variance-covariance matrix of Y is $N[\text{diag}(Q) - QQ^\tau]$, respectively. In the following, let us denote $\Sigma = \text{diag}(P) - PP^\tau$ and $\Sigma_D = \text{diag}(Q) - QQ^\tau$.

The sample mean $\bar{X} = X/M$ serves as the estimate of P ; and sample mean $\bar{Y} = Y/N$ serves as the estimate of Q . Assume that the sample sizes M and N are large enough that the large sample

theory applies. By the multivariate central limit theorem of large sample theory, $\sqrt{M}[\bar{X} - P]$ tends to a multivariate normal distribution with a zero mean vector and variance-covariance matrix Σ ; and $\sqrt{N}[\bar{Y} - Q]$ tends also to a multivariate normal distribution with a zero mean vector and variance-covariance matrix Σ_D (Lehmann 1983³⁴, Theorem 5.1.8, p343).

Now, let us define

$$f_{ij} = P_{ij} \log \frac{P_{ij}}{P_{i.} P_{.j}}, \quad g_{ij} = Q_{ij} \log \frac{Q_{ij}}{Q_{i.} Q_{.j}},$$

where $P_{i.} = \sum_{j=0}^2 P_{ij}$, $P_{.j} = \sum_{i=0}^2 P_{ij}$, $Q_{i.} = \sum_{j=0}^2 Q_{ij}$ and $Q_{.j} = \sum_{i=0}^2 Q_{ij}$. Let $f = I(A, B) = \sum_{i=0}^2 \sum_{j=0}^2 f_{ij}$ and $g = I(A, B|D) = \sum_{i=0}^2 \sum_{j=0}^2 g_{ij}$. Then, the information gain can be expressed as

$$IG(AB | D) = g - f = \sum_{i=0}^2 \sum_{j=0}^2 g_{ij} - \sum_{i=0}^2 \sum_{j=0}^2 f_{ij}.$$

For functions f and g , denote their partial derivatives as $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ which are column vectors. The elements of $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ are given in the Appendix A as

$$\begin{aligned} \frac{\partial f}{\partial P_{ij}} &= \log \frac{P_{ij}}{P_{i.} P_{.j}} - \log \frac{P_{22}}{P_{2.} P_{.2}}, \\ \frac{\partial g}{\partial Q_{ij}} &= \log \frac{Q_{ij}}{Q_{i.} Q_{.j}} - \log \frac{Q_{22}}{Q_{2.} Q_{.2}}. \end{aligned} \quad (13)$$

Denote $\Lambda = [\frac{\partial f}{\partial P}]^T \Sigma \frac{\partial f}{\partial P} / M + [\frac{\partial g}{\partial Q}]^T \Sigma_D \frac{\partial g}{\partial Q} / N$.

We denote the estimate of P_{ij} as $\hat{P}_{ij} = X_{ij}/M$, and the estimate of Q_{ij} as $\hat{Q}_{ij} = Y_{ij}/N$. Similarly, we denote the estimates of other parameters as $\hat{P}_{i.} = \sum_{j=0}^2 \hat{P}_{ij}$, etc. Then, the estimates \hat{f} , $\hat{\Lambda}$, and \hat{g} of f , Λ , and g can be calculated by \hat{P}_{ij} and \hat{Q}_{ij} , accordingly.

By large sample theory, $\sqrt{M}(\hat{f} - f)$ tends to a normal distribution with zero mean and variance $[\frac{\partial f}{\partial P}]^T \Sigma \frac{\partial f}{\partial P}$; similarly, $\sqrt{N}(\hat{g} - g)$ tends to a normal distribution with zero mean and variance $[\frac{\partial g}{\partial Q}]^T \Sigma_D \frac{\partial g}{\partial Q}$ (Lehmann 1983³⁴, Theorem 2.5.3, p112). Notice that $f = I(A, B) = I(A, B|D) = g$ under the null hypothesis of independence of the disease and the two markers A and B ; and so $\hat{g} - \hat{f} = (\hat{g} - g) - (\hat{f} - f)$ tends to a normal distribution with zero mean and variance Λ . With these discussions in mind, the statistical tests to test the dependence of markers A and B and the disease can be constructed by

$$\begin{aligned} T_{IG} &= (\hat{g} - \hat{f})^2 / \hat{\Lambda}, \\ T &= (\hat{P} - \hat{Q})^T \left(\frac{\hat{\Sigma}}{M} + \frac{\hat{\Sigma}_D}{N} \right)^{-1} (\hat{P} - \hat{Q}). \end{aligned} \quad (14)$$

The test T_{IG} is based on the information gain $IG(AB | D)$, the test T is a naive χ^2 -distributed statistic which is based on the 2 by 8 contingency table to compare the counts of case and controls for genotype combinations of markers A and B . Under the null hypothesis that the two markers are independent of the disease, the test T_{IG} is centrally χ_1^2 -distributed with 1 degree of freedom and the test T is centrally χ_8^2 -distributed with 8 degrees of freedom. Under the alternative hypothesis that the disease and the two markers are not independent, the test T_{IG} is non-centrally χ_1^2 -distributed with a non-centrality parameter $\lambda_{IG} = (g - f)^2/\Lambda$ and the test T is non-centrally χ_8^2 -distributed with a non-centrality parameter $\lambda_T = (P - Q)^\tau \left(\frac{\Sigma}{M} + \frac{\Sigma_D}{N} \right)^{-1} (P - Q)$.

The statistics T_{IG} and T are overall test statistics to test the dependence of markers A and B and the disease. In case that the markers are associated with the disease (i.e., the markers are not independent of the disease), we need to know which genotypes are associated with the disease. For genotype $(G_A = i, G_B = j)$, we may test if it is associated with the disease by either of the two tests as follows

$$\begin{aligned} T_{E,ij} &= \frac{(\hat{g}_{ij} - \hat{f}_{ij})^2}{\hat{\Lambda}_{ij}}, \\ T_{ij} &= \frac{(\hat{P}_{ij} - \hat{Q}_{ij})^2}{\widehat{\text{Var}}(\hat{P}_{ij} - \hat{Q}_{ij})}, \end{aligned} \quad (15)$$

where $\hat{\Lambda}_{ij}$ is the estimate of Λ_{ij} , and $\widehat{\text{Var}}(\hat{P}_{ij} - \hat{Q}_{ij})$ is the estimate of $\text{Var}(\hat{P}_{ij} - \hat{Q}_{ij})$ which are given by

$$\begin{aligned} \Lambda_{ij} &= \frac{\left[\frac{\partial f_{ij}}{\partial P} \right]^\tau \Sigma \frac{\partial f_{ij}}{\partial P}}{M} + \frac{\left[\frac{\partial g_{ij}}{\partial Q} \right]^\tau \Sigma_D \frac{\partial g_{ij}}{\partial Q}}{N}, \\ \text{Var}(\hat{P}_{ij} - \hat{Q}_{ij}) &= \frac{P_{ij}(1 - P_{ij})}{M} + \frac{Q_{ij}(1 - Q_{ij})}{N}. \end{aligned}$$

The test T_{ij} simply compares the difference $\hat{P}_{ij} - \hat{Q}_{ij}$ of the proportions of cases and controls with genotype $(G_A = i, G_B = j)$, and the test $T_{E,ij}$ is based on the difference $\hat{f}_{ij} - \hat{g}_{ij}$. If genotype $(G_A = i, G_B = j)$ is strongly associated with the disease, the differences $\hat{P}_{ij} - \hat{Q}_{ij}$ and $\hat{f}_{ij} - \hat{g}_{ij}$ tend to be different from 0 and significant result is likely to be found by the test T_{ij} and/or test $T_{E,ij}$.

Under the null hypothesis that the disease and the genotype $(G_A = i, G_B = j)$ are independent, the test $T_{E,ij}$ and T_{ij} are centrally χ_1^2 -distributed with 1 degree of freedom. Under the alternative hypothesis that the disease and the genotype $(G_A = i, G_B = j)$ are not independent, the test $T_{E,ij}$ and T_{ij} are non-centrally χ_1^2 -distributed with non-centrality parameters $\lambda_{E,ij} = (g_{ij} - f_{ij})^2/\Lambda_{ij}$ and $\lambda_{ij} = (P_{ij} - Q_{ij})^2/\text{Var}(\hat{P}_{ij} - \hat{Q}_{ij})$, respectively.

2.6 Test Statistics Based on the 3-Way Interaction Information Gain and Total Correlation Information Gain

Again, consider a case-control design with M controls from an unaffected population and N cases from an affected population. Let us denote by X_{ije} the count of controls whose genotypes are ($G_A = i, G_B = j, E = e$); and similarly denote by Y_{ije} the count of cases whose genotypes are ($G_A = i, G_B = j, E = e$), $i, j, e = 0, 1, 2$. The test statistics can be built based on two column vectors X and Y , where X includes all X_{ije} except X_{222} , and Y includes all Y_{ije} except Y_{222} to remove the redundancy.

In the general population, denote the joint genotype probabilities for markers A and B and environmental factor E by $P_{ije} = P(G_A = i, G_B = j, E = e)$. In the disease population, denote the joint conditional genotype probabilities by $Q_{ije} = P(G_A = i, G_B = j, E = e | D = 1)$. Let us denote column vector P which includes all P_{ije} except P_{222} , and we denote column vector Q which includes all Q_{ije} except Q_{222} . The column counting vector $\begin{pmatrix} X \\ X_{222} \end{pmatrix}$ follows a multinomial distribution $\left(M, \begin{pmatrix} P \\ P_{222} \end{pmatrix}\right)$; and the column counting vector $\begin{pmatrix} Y \\ Y_{222} \end{pmatrix}$ follows a multinomial $\left(N, \begin{pmatrix} Q \\ Q_{222} \end{pmatrix}\right)$ distribution. The mean vector of $\bar{X} = X/M$ is P , and the mean vector of $\bar{Y} = Y/N$ is Q , respectively. The variance-covariance matrix of X is $M\Sigma$, and the variance-covariance matrix of Y is $N\Sigma_D$, where $\Sigma = \text{diag}(P) - PP^T$ and $\Sigma_D = \text{diag}(Q) - QQ^T$. Assume that the sample sizes M and N are large enough that the large sample theory applies. By the multivariate central limit theorem of large sample theory, $\sqrt{M}[\bar{X} - P]$ tends to a multivariate normal distribution with a zero mean vector and variance-covariance matrix Σ ; and $\sqrt{N}[\bar{Y} - Q]$ tends also to a multivariate normal distribution with a zero mean vector and variance-covariance matrix Σ_D (Lehmann 1983³⁴, Theorem 5.1.8, p343).

Denote $P_{i..} = \sum_{j=0}^2 \sum_{e=0}^2 P_{ije}$, $P_{.j.} = \sum_{i=0}^2 \sum_{e=0}^2 P_{ije}$, $P_{..e} = \sum_{i=0}^2 \sum_{j=0}^2 P_{ije}$; similarly, we may define $Q_{i..}$, $Q_{.j.}$ and $Q_{..e}$. Now, let us define

$$f_{ije} = P_{ije} \log \frac{P_{ije}}{P_{i..} P_{.j.} P_{..e}}, \quad g_{ije} = Q_{ije} \log \frac{Q_{ije}}{Q_{i..} Q_{.j.} Q_{..e}},$$

Let $f = TCI(A, B, E) = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 f_{ije}$ and $g = TCI(A, B, E | D) = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 g_{ije}$. Then, the total correlation information gain can be expressed as

$$TCIG(ABE | D) = g - f = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 g_{ije} - \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 f_{ije}.$$

For functions f and g , denote their partial derivatives as $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ which are column vectors. The

elements of $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ are given in the Supplementary Materials Appendix A as

$$\begin{aligned}\frac{\partial f}{\partial P_{ije}} &= \log \frac{P_{ije}}{P_{i..}P_{.j.}P_{..e}} - \log \frac{P_{222}}{P_{2..}P_{.2.}P_{..2}}, \\ \frac{\partial g}{\partial Q_{ije}} &= \log \frac{Q_{ije}}{Q_{i..}Q_{.j.}Q_{..e}} - \log \frac{Q_{222}}{Q_{2..}Q_{.2.}Q_{..2}}.\end{aligned}\quad (16)$$

Denote $\Lambda = [\frac{\partial f}{\partial P}]^{\tau} \Sigma \frac{\partial f}{\partial P} / M + [\frac{\partial g}{\partial Q}]^{\tau} \Sigma_D \frac{\partial g}{\partial Q} / N$.

To build a 3-way interaction information gain based test statistic, let us denote $P_{ij.} = \sum_{e=0}^2 P_{ije}$, $P_{.je} = \sum_{i=0}^2 P_{ije}$, $P_{i.e} = \sum_{j=0}^2 P_{ije}$; similarly, we may define $Q_{ij.}$, $Q_{.je}$ and $Q_{i.e}$. Now, let us define

$$h_{ije} = P_{ije} \log \frac{P_{ije}P_{i..}P_{.j.}P_{..e}}{P_{ij.}P_{.je}P_{i.e}}, \quad \ell_{ije} = Q_{ije} \log \frac{Q_{ije}Q_{i..}Q_{.j.}Q_{..e}}{Q_{ij.}Q_{.je}Q_{i.e}},$$

Let $h = I(A, B, E) = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 h_{ije}$ and $\ell = I(A, B, E|D) = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 \ell_{ije}$. Then, the 3-way interaction information gain of markers A and B and environmental factor E can be expressed as

$$IIG(ABE | D) = \ell - h = \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 \ell_{ije} - \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 h_{ije}.$$

For functions h and ℓ , denote their partial derivatives as $\frac{\partial h}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$ which are column vectors. The elements of $\frac{\partial h}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$ are given in the Supplementary Materials Appendix B as

$$\begin{aligned}\frac{\partial h}{\partial P_{ije}} &= \log \frac{P_{ije}P_{i..}P_{.j.}P_{..e}}{P_{ij.}P_{i.e}P_{.je}} - \log \frac{P_{222}P_{2..}P_{.2.}P_{..2}}{P_{22.}P_{2.2}P_{.22}}, \\ \frac{\partial \ell}{\partial Q_{ije}} &= \log \frac{Q_{ije}Q_{i..}Q_{.j.}Q_{..e}}{Q_{ij.}Q_{i.e}Q_{.je}} - \log \frac{Q_{222}Q_{2..}Q_{.2.}Q_{..2}}{Q_{22.}Q_{2.2}Q_{.22}}.\end{aligned}\quad (17)$$

Denote $\Gamma = [\frac{\partial h}{\partial P}]^{\tau} \Sigma \frac{\partial h}{\partial P} / M + [\frac{\partial \ell}{\partial Q}]^{\tau} \Sigma_D \frac{\partial \ell}{\partial Q} / N$.

We denote the estimate of P_{ije} as $\hat{P}_{ije} = X_{ije}/M$, and the estimate of Q_{ije} as $\hat{Q}_{ije} = Y_{ije}/N$. Similarly, we denote the estimates of other parameters as $\hat{P}_{i..} = \sum_{j=0}^2 \sum_{e=0}^2 \hat{P}_{ije}$, etc. Then, the estimates \hat{f} , \hat{g} , \hat{h} , $\hat{\ell}$, $\hat{\Lambda}$, and $\hat{\Gamma}$ of f , g , h , ℓ , Λ , and Γ can be calculated by \hat{P}_{ije} and \hat{Q}_{ije} , accordingly. The statistical tests to test the correlations and interactions of markers A and B and environmental factor E with the disease can be constructed by

$$\begin{aligned}T_{TCIG} &= (\hat{g} - \hat{f})^2 / \hat{\Lambda}, \\ T_{IIG} &= (\hat{h} - \hat{\ell})^2 / \hat{\Gamma}.\end{aligned}\quad (18)$$

The test T_{TCIG} is based on the total correlation information gain $TCIG(ABE | D)$, and it can be used to test the 3-way correlations. The test T_{IIG} , on the other hand, is based on 3-way interaction information gain $IIG(ABE | D)$, and it can be used to test the 3-way interactions. Under the null hypothesis that the two markers A and B and the environmental factor E are independent of the disease, the test statistics T_{TCIG} and T_{IIG} are centrally χ_1^2 -distributed. Under the alternative hypothesis that the two markers and the environmental factor are not independent of the disease, the test statistics T_{TCIG} and T_{IIG} are non-centrally χ_1^2 -distributed with non-centrality parameters $\lambda_{TCIG} = (g - f)^2/\Lambda$ and $\lambda_{IIG} = (h - \ell)^2/\Gamma$, respectively.

2.7 Test Statistics Based on the K-Way Interaction Information Gain and Total Correlation Information Gain

Given a case-control sample with M controls and N cases, we are going to construct test statistics to test K-way interaction of K attributes A_1, \dots, A_K . Hereafter, $|\vec{s}|$ is the number of elements in a vector \vec{s} . Such as 2-way and 3-way cases, let us denote

$$\begin{aligned} TCIG(\mathcal{A} | D) &= g - f = \sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 [g_{a_1 \cdots a_K} - f_{a_1 \cdots a_K}], \\ g_{a_1 \cdots a_K} &= Q_{a_1 \cdots a_K} \log \frac{Q_{a_1 \cdots a_K}}{\prod_{\substack{\vec{s} \subset (a_1, \dots, a_K) \\ |\vec{s}|=1}} Q_{\vec{s}}}, \\ f_{a_1 \cdots a_K} &= P_{a_1 \cdots a_K} \log \frac{P_{a_1 \cdots a_K}}{\prod_{\substack{\vec{s} \subset (a_1, \dots, a_K) \\ |\vec{s}|=1}} P_{\vec{s}}}, \end{aligned}$$

where $\prod_{\substack{\vec{s} \subset (a_1, \dots, a_K) \\ |\vec{s}|=1}} Q_{\vec{s}} = Q_{a_1, \dots, \cdot} \cdots Q_{\cdot, \dots, a_K}$ is the product of all individual marginal probabilities of A_1, \dots, A_K in the disease population such as those in 2-way and 3-way cases, and $\prod_{\substack{\vec{s} \subset (a_1, \dots, a_K) \\ |\vec{s}|=1}} P_{\vec{s}} = P_{a_1, \dots, \cdot} \cdots P_{\cdot, \dots, a_K}$ is the product of all individual marginal probabilities of A_1, \dots, A_K in the general population. For functions f and g , denote their partial derivatives as $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ which are column vectors. The elements of $\frac{\partial f}{\partial P}$ and $\frac{\partial g}{\partial Q}$ are given as

$$\begin{aligned} \frac{\partial f}{\partial P_{a_1 \cdots a_K}} &= \log \frac{P_{a_1 \cdots a_K}}{P_{a_1, \dots, \cdot} \cdots P_{\cdot, \dots, a_K}} - \log \frac{P_{2 \cdots 2}}{P_{2, \dots, \cdot} \cdots P_{\cdot, \dots, 2}}, \\ \frac{\partial g}{\partial Q_{a_1 \cdots a_K}} &= \log \frac{Q_{a_1 \cdots a_K}}{Q_{a_1, \dots, \cdot} \cdots Q_{\cdot, \dots, a_K}} - \log \frac{Q_{2 \cdots 2}}{Q_{2, \dots, \cdot} \cdots Q_{\cdot, \dots, 2}}, \end{aligned} \quad (19)$$

which can be proved along the vein of relation (16) in Supplementary Materials Appendix C. Denote $\Lambda = [\frac{\partial f}{\partial P}]^\tau \Sigma \frac{\partial f}{\partial P} / M + [\frac{\partial g}{\partial Q}]^\tau \Sigma_D \frac{\partial g}{\partial Q} / N$, where P is a column vector which includes all $P_{a_1 \cdots a_K}$ except

$P_{2\dots 2}$ and $\Sigma = \text{diag}(P) - PP^\tau$, and Q is a column vector which includes all $Q_{a_1\dots a_K}$ except $Q_{2\dots 2}$ and $\Sigma_D = \text{diag}(Q) - QQ^\tau$.

Similarly, let us denote

$$\begin{aligned} IIG(\mathcal{A} | D) &= \ell - h = \sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 [\ell_{a_1\dots a_K} - h_{a_1\dots a_K}], \\ \ell_{a_1\dots a_K} &= Q_{a_1\dots a_K} \log \frac{Q_{a_1\dots a_K} \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} Q_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} Q_{\bar{s}}}, \\ h_{a_1\dots a_K} &= P_{a_1\dots a_K} \log \frac{P_{a_1\dots a_K} \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}}, \end{aligned}$$

where $|(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)} = 0$ means that the subset $(a_1, \dots, a_K) \setminus \bar{s}$ contains an even number of elements, and $|(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)} = 1$ means that the subset $(a_1, \dots, a_K) \setminus \bar{s}$ contains an odd number of elements. Moreover, the product $\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} Q_{\bar{s}}$ does not contain Q_{a_1, \dots, a_K} since $\bar{s} \subset (a_1, \dots, a_K)$ means that \bar{s} is a real subset of (a_1, \dots, a_K) [i.e., $\bar{s} \neq (a_1, \dots, a_K)$], and so for the other products. For functions ℓ and h , denote their partial derivatives as $\frac{\partial \ell}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$ which are column vectors. The elements of $\frac{\partial h}{\partial P}$ and $\frac{\partial \ell}{\partial Q}$ are given as

$$\begin{aligned} \frac{\partial h}{\partial P_{a_1\dots a_K}} &= \log \frac{P_{a_1\dots a_K} \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}} - \log \frac{P_{2\dots 2} \prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}}, \\ \frac{\partial \ell}{\partial Q_{a_1\dots a_K}} &= \log \frac{Q_{a_1\dots a_K} \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} Q_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} Q_{\bar{s}}} - \log \frac{Q_{2\dots 2} \prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=0}} Q_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=1}} Q_{\bar{s}}}, \end{aligned} \quad (20)$$

which can be proved along the vein of relation (17) in Supplementary Materials Appendix C. Denote $\Gamma = [\frac{\partial h}{\partial P}]^\tau \Sigma \frac{\partial h}{\partial P} / M + [\frac{\partial \ell}{\partial Q}]^\tau \Sigma_D \frac{\partial \ell}{\partial Q} / N$. To test the K-way interactions and correlations between the disease and the attributes A_1, \dots, A_K , the χ_1^2 -distributed test statistics can be constructed as $T_{TCIG} = (\hat{g} - \hat{f})^2 / \hat{\Lambda}$ and $T_{IIG} = (\hat{h} - \hat{\ell})^2 / \hat{\Gamma}$, respectively. Here $\hat{f}, \hat{g}, \hat{h}, \hat{\ell}, \hat{\Lambda}$ and $\hat{\Gamma}$ are estimates of f, g, h, ℓ, Λ and Γ .

2.8 Association Test Statistics based on 1-way Entropy Loss

Suppose that one is interested in testing association between an attribute and the disease in a case-control study. The entropy of the attribute can be used as the basis to construct test statistics. The attribute here can be a single marker or an environment factor; in addition, if two or more markers are in strong linkage disequilibrium and their haplotype data are available, one may treat the haplotype data

as an attribute. For the convenience of presentation, we use marker A as the attribute in the following. It is well-known that the entropy maximizes when a system reaches its equilibrium state. In one locus case, an equilibrium state refers to the Hardy-Weinberg equilibrium. In the general population, the entropy $H(A)$ may reach the maximum since an assumption of Hardy-Weinberg equilibrium is likely to be true. In the disease population, the assumption of Hardy-Weinberg equilibrium may not be true and the conditional entropy $H(A|D)$ may decrease. The entropy loss of marker A in the presence of disease D can be defined as follows:

$$EL(A | D) = H(A) - H(A|D). \quad (21)$$

If the disease and marker A are independent, $H(A|D) = H(A)$ and so their difference or entropy loss $EL(A | D)$ is equal to 0. Hence, we may test the association between marker A and disease by testing if the difference is zero. Based on this rationale, we may build test statistics for practical application.

In the general population, denote the genotype probabilities for marker A by $P_i = P(G_A = i)$; and in the disease population, denote the conditional genotype probabilities for marker A by $Q_i = P(G_A = i | D = 1)$. Let us denote $P = (P_0, P_1)^\tau$, and $Q = (Q_0, Q_1)^\tau$. The entropy loss can be expressed as

$$EL(A | D) = H(A) - H(A|D) = \sum_{i=0}^2 P_i \log P_i - \sum_{i=0}^2 Q_i \log Q_i.$$

For entropy functions $H(A)$ and $H(A|D)$, denote their partial derivatives as $\frac{\partial H(A)}{\partial P}$ and $\frac{\partial H(A|D)}{\partial Q}$ which are column vectors. It is easy to show that the elements of $\frac{\partial H(A)}{\partial P}$ and $\frac{\partial H(A|D)}{\partial Q}$ are given by

$$\begin{aligned} \frac{\partial H(A)}{\partial P_i} &= \log P_i - \log P_2, \\ \frac{\partial H(A|D)}{\partial Q_i} &= \log Q_i - \log Q_2. \end{aligned} \quad (22)$$

For a case-control design with M controls and N cases, let us denote by X_i the count of controls whose genotypes are $(G_A = i)$; and similarly, denote by Y_i the count of cases whose genotypes are $(G_A = i)$, $i = 0, 1, 2$. We denote the estimate of P_i as $\hat{P}_i = X_i/M$, and the estimate of Q_i as $\hat{Q}_i = Y_i/N$. Then, the estimates $\hat{H}(A)$, $\hat{\Omega}$, and $\hat{H}(A|D)$ of $H(A)$, Ω , and $H(A|D)$ can be calculated by \hat{P}_i and \hat{Q}_i , accordingly. Denote $\Omega = \left[\frac{\partial H(A)}{\partial P} \right]^\tau \Sigma \frac{\partial H(A)}{\partial P} / M + \left[\frac{\partial H(A|D)}{\partial Q} \right]^\tau \Sigma_D \frac{\partial H(A|D)}{\partial Q} / N$, where $\Sigma = \text{diag}(P) - PP^\tau$ and $\Sigma_D = \text{diag}(Q) - QQ^\tau$. The entropy loss based statistics to test the dependence of marker A and the disease can be constructed by

$$T_{EL} = \left(\hat{H}(A) - \hat{H}(A|D) \right)^2 / \hat{\Omega}.$$

Under the null hypothesis, T_{EL} is centrally χ_1^2 -distributed. Under the alternative hypothesis, T_{EL} is non-centrally χ_1^2 -distributed with a non-centrality parameter $\lambda_{EL} = (H(A) - H(A|D))^2 / \Omega$.

3 Results

In this section, we apply the methods to the bladder cancer data of Andrew et al. (2006) to search for interactions of the disease and the genetic/smoking factors. Then, we investigate the robustness of the proposed test statistics by type I error rate calculation using the joint genotype frequencies of the bladder cancer data. We perform power analysis using the analytical non-centrality parameters of 2-way tests under a few interaction models taken from literature.^{12,35}

3.1 Application to Bladder Cancer Data

We apply the proposed methods to the bladder cancer data of Andrew et al. (2006),² and the results are presented in Table 1. The dataset consists of 355 cases and 559 controls, and the genotype data of 7 SNPs are available, i.e., BER pathway: APE1 148, XRCC1 399, and XRCC1 194; DSB pathway: XRCC3 241; and NER pathway: XPC PAT, XPD 751, and XPD 312. In addition to the bladder cancer status, the following information of each individual is also available: gender, age, and smoking variable Pack_years (e.g., non-smoking, < 35 pack-years, \geq 35 pack-years).

In the MDR analysis of Andrew et al. (2006),² the combination of XPD 751 and XPD 312 was the best two-factor model, which was confirmed by the interaction dendrogram and logistic regression analysis; the three-factor model added Pack-years of smoking to XPD 751 and XPD 312 for the most accurate model, which however was not confirmed by the interaction dendrogram and logistic regression analysis (the logistic regression model failed to converge).

For 2-way interaction, we confirm the result of Andrew et al. (2006).² The combination of XPD 751 and XPD 312 is the only significant one by our 2-way mutual information gain test statistic T_{IG} ($T_{IG} = 51.62$, p-value = $6.75e-13$), and none of the rest two-factor combinations provides significant result by T_{IG} . By adding each of the rest 5 SNPs and Pack_years, the 3-way total correlation information gain test statistic T_{TCIG} provides significant result (p-value $\leq 2.67e-9$); however, the 3-way interaction information gain test statistic T_{IIG} provides significant result for none of the three-factor combinations of XPD 751 and XPD 312 and one of the rest 5 SNPs and Pack_years (p-value ≥ 0.23).

The only significant result of 3-way interaction information gain test statistic T_{IIG} at 5% significance level is from the combination of XRCC1 399, XRCC1 194, XRCC3 241 ($T_{IIG} = 4.25$, p-value = 0.04). However, the result is hardly significant under an adjustment of multiple tests such as Bonferroni procedure. Therefore, there is no 3-way interaction combination based on our analysis. The very significant results of 3-way total correlation information gain test statistic T_{TCIG} of Table 1 are most likely due to the 2-way combination of XPD 751 and XPD 312.

3.2 Type I Error Rates

Using the joint SNP genotype frequencies of bladder cancer data,² we perform simulation to evaluate the type I error rates of 2-way information gain based test statistic T_{IG} and 3-way test statistics T_{IIG} and T_{TCIG} , and the results are presented in Table 2. Each entry of the empirical type I error rates in Table 2 is calculated by 100,000 simulations, i.e., we simulate 100,000 random samples of $N = M = 100, 150, 200, 250, 300, 400, 500, 600, 700$ cases and controls, respectively. In each sample, M controls and N cases are generated under multinomial distributions (M, P) and (N, Q) , respectively. Here $P = Q$ is the joint genotype frequencies estimated by the bladder cancer data. For instance, the joint genotype frequencies of the combination of Xpd 751 and Xpd 312 is $P = Q = (156, 42, 15, 60, 193, 19, 5, 33, 36)^T / (156+42+15+60+193+19+5+33+36) = (156, 42, 15, 60, 193, 19, 5, 33, 36)^T / 559$, which is used to generate simulation data to calculate the empirical type I error rates of the 2-way test statistic T_{IG} .

We assume $P = Q$ in our simulation to calculate the type I error rates, i.e., the disease status D is independent of genetic/environmental factors. For each sample, an empirical test value is calculated. The empirical type I error rates at nominal levels $\alpha = 0.01$ and $\alpha = 0.05$ are reported in Table 2, which are the proportions of the test values of the 100,000 samples that exceed 99-th and 95-th percentiles of the χ_1^2 -distribution, respectively. Because the disease status D is independent of genetic/environmental factors, the empirical type I error rates reported in Table 2 can be thought as false positives.

For each combination of genotype frequencies, 9 empirical type I error rates are calculated for sample sizes $N = M = 100, 150, 200, 250, 300, 400, 500, 600, 700$, respectively. We calculate the type I error rates of 2-way test statistic T_{IG} based on the joint SNP genotype frequencies of the combination of Xpd 751 and Xpd 312 of bladder cancer data, which provide the only very significant result of T_{IG} (Table 1).

For the 3-way test statistic T_{TCIG} , we add one of the other five SNPs or Pack_years in addition to Xpd 751 and Xpd 312 to calculate the joint SNP genotype frequencies and to calculate the empirical type I error rates. This leads to six combinations of three factors/attributes, i.e., Xpd 751 and Xpd 312 plus one SNP or Pack_years. One may want to notice these six combinations provide very significant results of total correlation between the bladder cancer and the three factors/attributes (Table 1). The results of Table 2 show that the empirical type I error rates of 2-way test statistic T_{IG} and 3-way test statistic T_{TCIG} are around the nominal level $\alpha = 0.01$ or $\alpha = 0.05$ when the sample sizes $M = N \geq 300$. Therefore, the test statistics T_{IG} and T_{TCIG} are reasonably conservative and robust. The very significant results of T_{IG} and T_{TCIG} in Table 1 are most likely from the strong interaction between the bladder cancer and two SNPs Xpd 751 and Xpd 312.

In our simulation to calculate the entries of Table 2, the null hypothesis of T_{IG} is that the disease status D is independent of genetic markers $A = \text{Xpd 751}$ and $B = \text{Xpd 312}$, i.e., $Q_{ij} = P(G_A = i, G_B = j | D = 1) = P(G_A = i, G_B = j) = P_{ij}$, but the genotype of SNP A is not independent of the genotype of SNP B . The null hypothesis of T_{TCIG} is that $Q_{ije} = P(G_A = i, G_B = j, E = e | D = 1) = P(G_A = i, G_B = j, E = e) = P_{ije}$, i.e., the disease status D is independent of genetic/environmental factors, but the pair-wise dependence and three-way dependence of genetic/environmental factors are allowed. Actually, the genotypes of the SNP pair of Xpd 751 and Xpd 312 are strongly dependent of each other (Pearson $\chi^2 = 256.83$, p-value = 0.0000). In addition, Xpd 751 and Xpd 312 are in strong linkage disequilibrium.² Therefore, the simulated data are generated under the null hypothesis of either T_{IG} or T_{TCIG} since the two SNPs, Xpd 751 and Xpd 312, are correlated to each other. The empirical type I error rates of tests T_{IG} and T_{TCIG} reported in Table 2 are around the nominal levels, and the two tests are reasonably robust.

To calculate the empirical type I error rates of the interaction information gain based test statistic T_{IIG} , we screen the three SNP combinations which are significantly correlated to each other. Consider three SNPs A, B , and C . By significantly correlated to each other for the three SNPs, we mean all the four null hypotheses,

$$H_1 : P(G_A, G_B, G_C) = P(G_A, G_B)P(G_C),$$

$$H_2 : P(G_A, G_B, G_C) = P(G_A)P(G_B, G_C),$$

$$H_3 : P(G_A, G_B, G_C) = P(G_B)P(G_A, G_C),$$

$$H_4 : P(G_A, G_B, G_C) = P(G_A)P(G_B)P(G_C),$$

all unlikely to be true. We use the Pearson χ^2 test to screen the three SNP combinations. In Table D.1 of the Supplementary Materials Appendix Appendix D , we present these three attribute combinations of SNPs. Utilizing the joint genotype frequencies of the three SNP combinations in Table D.1 of the Supplementary Materials, we perform simulation to calculate the empirical type I error rates for the 3-way test statistic T_{IIG} . Since each of the three SNP combinations is selected based on the existence of significant correlations of the three SNPs, the simulated data are likely to be generated under the null hypothesis of T_{IIG} , i.e., $I(A, B, C|D) = I(A, B, C)$. The empirical type I error rates of the 3-way test statistic T_{IIG} reported in Table 2 are generally slight lower than the nominal levels. Hence, the test T_{IIG} is conservative and robust. Basically, the test T_{IIG} is more conservative than the test T_{TCIG} since the test T_{IIG} is constructed to detect the 3-way or higher order interactions and the test T_{TCIG} is constructed to detect the correlations. The existence of 2-way or 3-way interactions implies 3-way correlations, but 3-way correlations are not necessarily due to 3-way interactions.

In Table E.1 of the Supplementary Materials Appendix E , we present the type I error rates of 1-way entropy loss test statistic T_{EL} . The test statistic T_{EL} is reasonably robust and conservative.

3.3 Power Comparison

After evaluating the robustness of the test statistic T_{IG} by type I error rate calculation in subsection 3.2, we perform power calculation for the information gain based test T_{IG} and the naive test T . We mainly concern with the performance of the test statistic T_{IG} when the interaction is nonlinear, and the main effect may be absent. To achieve the goal, six models of two-locus penetrance functions and allele frequencies are taken from Moore et al. (2002)³⁵, Figures 5-10, and the penetrance functions and allele frequencies are presented in Table 3. Similarly, four models are taken from Ritchie et al. (2003)¹², Figure 2, and the related parameters are presented in Table 4.

To make a comparison, we calculate the theoretical power curves of both test statistics T_{IG} and T based on their non-centrality parameters $\lambda_{T_{IG}}$ and λ_T , and the results are plotted in Figures 2 and 3, respectively. Generally, the power of information gain based test T_{IG} has similar power as naive test T

or higher power than T . For models 2 and 4-6 in Table 3 and model 1-2 in Table 4, the power curves of T_{IG} are higher than that of T ; for other models, the power is similar. Hence, the information gain based test T_{IG} is advantageous over the naive test T in terms of power comparison.

4 Discussion

In this paper, we propose information gain based test statistics to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. For 2-way interaction, an information gain based approach is proposed using mutual information. The information gain in the presence of disease is defined as a one-dimensional variable through mutual information and entropy function of genetic markers, i.e., $IG(AB | D) = I(A, B|D) - I(A, B)$. Based on the one-dimensional information gain, a test statistic T_{IG} is constructed and is showed to be χ_1^2 -distributed with 1 degree of freedom. As equation (14) shows, the information gain based test T_{IG} does not involve matrix inverse calculation which facilitates the implementation in practical applications because it is based on the normalization of a one-dimensional random variable $\hat{g} - \hat{f} = \hat{I}(A, B|D) - \hat{I}(A, B)$. On the other hand, the calculation of the naive test T involves matrix inverse calculation and it is almost impossible to use it for sparse data as in our simulation calculation of empirical type I error rates. One way is to calculate the generalized inverse of matrix to implement the naive test T , and then its degrees of freedom changes from dataset to dataset. By power comparison, we clearly showed that the naive test T does not have an advantage over the information gain test T_{IG} .

In Wu et al. (2009),²⁸ a mutual information based approach was proposed to construct a statistic to test 2-way gene-environment interaction by using a multi-dimensional vector. Under the null hypothesis of independence of the genetic marker and the environmental factor, the test statistic was showed to be a χ_2^2 -distributed variable with 2 degrees of freedom.²⁸ Some of the theoretical justification in our discussion such as mutual information is similar to that of Wu et al. (2009).²⁸ However, our way to construct the test statistic T_{IG} is different. In addition, our test statistic T_{IG} is χ_1^2 -distributed with 1 degree of freedom no matter under the null hypothesis of independence of disease status and genetic/environmental factors or under the alternative hypothesis. Under the null hypothesis, T_{IG} is centrally χ_1^2 -distributed; under the alternative hypothesis, the T_{IG} is non-centrally χ_1^2 -distributed.

The methods are generalized to test 3-way or high order K-way interactions and correlations of genetic markers and environmental factors. Two approaches are proposed: (1) an interaction information gain (IIG) based approach and a total correlation information gain (TCIG) based approach. Such as the 2-way case, the interaction information gain and total correlation information gain are defined as one-dimensional variables. The related χ_1^2 -distributed test statistics T_{IIG} and T_{TCIG} are constructed to test 3-way or higher order interactions and total correlations, respectively. The test statistic T_{IIG} is based on interaction information gain and it can test 3-way or higher order interactions; the test statistic T_{TCIG} , however, is based on total correlation information gain and it can test 3-way or higher order correlations. One may want to notice that correlation can be treated as the interaction in 2-way case, but they are not the same for 3-way or high order K-way cases.

The proposed method is applied to bladder cancer data of Andrew et al. (2006).² We are able to confirm the significant result of 2-way interaction/correlation combination of XPD 751 and XPD 312 in Andrew et al. (2006).² However, we find that there is no significant result of 3-way interaction combinations for the bladder cancer data after adjusting for multiple tests. In the meantime, significant 3-way correlations are found for the bladder cancer data which are basically from the 2-way interaction/correlation combination of XPD 751 and XPD 312.

In practice, one may use forward procedure to detect the interactions using test statistics T_{IG} and T_{IIG} . That is, one may test 2-way interactions by T_{IG} first. In the presence of 2-way interactions, one may search for evidence of 3-way and higher order interactions by T_{IIG} . If there are multiple genetic markers, the proposed method can be used to construct genet network to interpret the relation among the markers and environmental factors with the disease. Our analysis of the bladder cancer data provides an example of the forward procedure to detect the interactions. Similarly, one may use test statistics T_{IG} and T_{TCIG} to detect the correlations, but the high order correlations may be actually from low order interactions/correlations.

One advantage of the proposed method is that it collapses high-dimensional genetic and environment data into a single dimension, and this makes it possible to build test statistic for high-dimensional sparse data to detect and to characterize gene-gene and gene-environment interactions and correlations. For instance, there are 27 genotype combinations if we consider 3 di-allelic markers. By using 3-way

interaction information gain and total correlation information gain of the three markers, we may reduce the 27-dimensional data to be one-dimensional variables to construct the three-way information gain based test statistics. The principle applies to high order K-way interactions and correlations.

To our knowledge, there is no much research about gene-gene and gene-environment interactions using entropy-based approaches, although investigators are paying more and more attention to the research.^{18,19,27,28,31} It is a new and an interesting area which deserves more attention and investigation. In this article, we make no assumption about population history. It is unclear which kind of impact would appear in the presence of population structure, genotyping error, phenocopy, and genetic heterogeneity. It would be interesting and important to systematically investigate the issues in the future study. So far, we focus on qualitative trait of complex trait, i.e., either with disease or no disease. It would be interesting to extend the method for quantitative traits. Besides, new methods and models need to be developed to analyze other data type such as sibling and nuclear family.^{36,37} These can be exciting areas for future investigation.

Appendix A Proof of Relation (13)

A.1 The Subscripts ij Do Not Contain 2

Notice

$$\begin{aligned}\frac{\partial f_{00}}{\partial P_{00}} &= \frac{\partial}{\partial P_{00}} \left[P_{00} \log \frac{P_{00}}{P_{0.}P_{.0}} \right] = \log \frac{P_{00}}{P_{0.}P_{.0}} + \left[1 - \frac{P_{00}}{P_{0.}} - \frac{P_{00}}{P_{.0}} \right] \log e, \\ \frac{\partial f_{01}}{\partial P_{00}} &= \frac{\partial}{\partial P_{00}} \left[P_{01} \log \frac{P_{01}}{P_{0.}P_{.1}} \right] = -\frac{P_{01}}{P_{0.}} \log e, \\ \frac{\partial f_{10}}{\partial P_{00}} &= \frac{\partial}{\partial P_{00}} \left[P_{10} \log \frac{P_{10}}{P_{1.}P_{.0}} \right] = -\frac{P_{10}}{P_{.0}} \log e, \\ \frac{\partial f_{11}}{\partial P_{00}} &= \frac{\partial}{\partial P_{00}} \left[P_{11} \log \frac{P_{00}}{P_{1.}P_{.1}} \right] = 0.\end{aligned}$$

In addition, we have

$$\begin{aligned}\frac{\partial f_{02}}{\partial P_{00}} &= \frac{\partial}{\partial P_{00}} \left[P_{02} \log \frac{P_{02}}{P_{0.}P_{.2}} \right] = \left[-\frac{P_{02}}{P_{0.}} + \frac{P_{02}}{P_{.2}} \right] \log e, \\ \frac{\partial f_{12}}{\partial P_{00}} &= \frac{\partial}{\partial P_{00}} \left[P_{12} \log \frac{P_{12}}{P_{1.}P_{.2}} \right] = \frac{P_{12}}{P_{.2}} \log e, \\ \frac{\partial f_{20}}{\partial P_{00}} &= \frac{\partial}{\partial P_{00}} \left[P_{20} \log \frac{P_{20}}{P_{2.}P_{.0}} \right] = \left[\frac{P_{20}}{P_{2.}} - \frac{P_{20}}{P_{.0}} \right] \log e,\end{aligned}$$

$$\frac{\partial f_{21}}{\partial P_{00}} = \frac{\partial}{\partial P_{00}} \left[P_{21} \log \frac{P_{21}}{P_{2 \cdot} P_{\cdot 1}} \right] = \frac{P_{21}}{P_{2 \cdot}} \log e.$$

Moreover, we have

$$\frac{\partial f_{22}}{\partial P_{00}} = \frac{\partial}{\partial P_{00}} \left[P_{22} \log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}} \right] = -\log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}} + \left[-1 + \frac{P_{22}}{P_{2 \cdot}} + \frac{P_{22}}{P_{\cdot 2}} \right] \log e.$$

Therefore, we have

$$\frac{\partial f}{\partial P_{00}} = \sum_{i=0}^2 \sum_{j=0}^2 \frac{\partial f_{ij}}{\partial P_{00}} = \log \frac{P_{00}}{P_{0 \cdot} P_{\cdot 0}} - \log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}}.$$

Similarly, we have for $i, j = 0, 1$

$$\frac{\partial f}{\partial P_{ij}} = \log \frac{P_{ij}}{P_{i \cdot} P_{\cdot j}} - \log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}}.$$

A.2 The Subscripts ij Contain One 2

Notice

$$\begin{aligned} \frac{\partial f_{00}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{00} \log \frac{P_{00}}{P_{0 \cdot} P_{\cdot 0}} \right] = -\frac{P_{00}}{P_{\cdot 0}} \log e, \\ \frac{\partial f_{01}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{01} \log \frac{P_{01}}{P_{0 \cdot} P_{\cdot 1}} \right] = 0, \\ \frac{\partial f_{02}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{02} \log \frac{P_{02}}{P_{0 \cdot} P_{\cdot 2}} \right] = \frac{P_{02}}{P_{\cdot 2}} \log e, \\ \frac{\partial f_{10}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{10} \log \frac{P_{10}}{P_{1 \cdot} P_{\cdot 0}} \right] = -\frac{P_{10}}{P_{\cdot 0}} \log e, \\ \frac{\partial f_{11}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{11} \log \frac{P_{11}}{P_{1 \cdot} P_{\cdot 1}} \right] = 0, \\ \frac{\partial f_{12}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{12} \log \frac{P_{12}}{P_{1 \cdot} P_{\cdot 2}} \right] = \frac{P_{12}}{P_{\cdot 2}} \log e, \\ \frac{\partial f_{20}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{20} \log \frac{P_{20}}{P_{2 \cdot} P_{\cdot 0}} \right] = \log \frac{P_{20}}{P_{2 \cdot} P_{\cdot 0}} + \left[1 - \frac{P_{20}}{P_{\cdot 0}} \right] \log e, \\ \frac{\partial f_{21}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{21} \log \frac{P_{21}}{P_{2 \cdot} P_{\cdot 1}} \right] = 0, \\ \frac{\partial f_{22}}{\partial P_{20}} &= \frac{\partial}{\partial P_{20}} \left[P_{22} \log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}} \right] = -\log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}} + \left[-1 + \frac{P_{22}}{P_{\cdot 2}} \right] \log e. \end{aligned}$$

Therefore, we have

$$\frac{\partial f}{\partial P_{20}} = \log \frac{P_{20}}{P_{2 \cdot} P_{\cdot 0}} - \log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}}.$$

Similarly, we have for $i, j = 0, 1$

$$\begin{aligned} \frac{\partial f}{\partial P_{i2}} &= \log \frac{P_{i2}}{P_{i \cdot} P_{\cdot 2}} - \log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}}, \\ \frac{\partial f}{\partial P_{2j}} &= \log \frac{P_{2j}}{P_{2 \cdot} P_{\cdot j}} - \log \frac{P_{22}}{P_{2 \cdot} P_{\cdot 2}}. \end{aligned}$$

Acknowledgment. The research was supported by a Research and Travel Support from the Intergovernmental Personnel Act (IPA), National Cancer Institute, NIH for Fan R., the National Cancer Institute grant R01-CA133996 for Amos C., and NIH grant LM009012 for Moore J. H.

References

1. Fisher, R.A. (1918) The correlations between relatives on the supposition of Mendelian inheritance, Trans. R. Soc. Edinburgh 52:399-433.
2. Andrew, A. S., Nelson, h. H., Kelsey, K. T., Moore, J. H., Meng, A. C., Casella, D. P., Tosteson, T. D., Schned, A. R., and Karagas, M. R. (2006) Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. Carcinogenesis 27:1030-1037.
3. Bateson, B. (2002) William Bateson: A biologist ahead of his time. J. Genet. 81:49-58.
4. Bateson, W. (1909) Mendel's Principles of Heredity, Cambridge University Press, Cambridge.
5. Moore, J.H., and Williams, S. M. (2009) Epistasis and its implications for personal genetics. The American Journal of Human Genetics 85:309-320.
6. Frankel, W. N., and Schork, N. J. (1996) Who's afraid of epistasis. Nature Genetics 14:371-373.

7. Mahdi, H., Fisher, B. A., Källberg, H., Plant, D., Malmström, V., Rönnelid, J., Charles, P., Ding, B., Alfredsson, L., Padyukov, L., Symmons, D. P. M., Venables, P. J., Klareskog, L. and Lundberg, K. (2009) Specific interaction between genotype, smoking and autoimmunity to citrullinated α -enolase in the etiology of rheumatoid arthritis. *Nature Genetics* 41:1319-1324.
8. van der Woude, D., Alemayehu, W. D., Verduijn, W., de Vries, R. R. P., Houwing-Duistermaat, J. J., Huizinga, T. W. J., and Toes, R. E. M. (2010) Gene-environment interaction influences the reactivity of autoantibodies to citrullinated antigens in rheumatoid arthritis. *Nature Genetics* 42:814-816.
9. Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., Yu, Y. (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics* 87:325-340.
10. Zhang, Y., and Liu, J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* 39:1167-1173.
11. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., Moore, J.H. (2001) Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics* 69:138-147.
12. Ritchie, M.D., Hahn, L.W., Moore, J.H. (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, phenocopy, and genetic heterogeneity. *Genetic Epidemiology* 24:150-157.
13. Ritchie, M.D., White, B.C., Parker, J.S., Hahn, L.W., Moore, J.H. (2003) Optimization of neural network architecture using genetic programming improves the detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinform.* 4, 28.

14. Ritchie, M.D., Coffey, C.S., Moore, J.H. (2004) Genetic programming neural networks as a bioinformatics tool in human genetics. *Lect. Notes Comput. Sci.* 3102:438-448.
15. Hahn, L.W., Ritchie, M.D., and Moore, J. H. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19:376-382.
16. Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore J.H. (2007) A balanced accuracy metric for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology* 31:306-315.
17. Lou, X.Y., Chen, G.B., Yan, L., Ma, J.Z., Zhu, J., Elston, R.C., and Li, M.D. (2007) A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics* 80:1125-1137.
18. Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W., and Li, Y. (2008) Exploration of gene-gene interaction effects using entropy-based methods. *European Journal of Human Genetics* 16:229-235.
19. Kang, G., Yue, W., Zhang, J., Cui, Y., Zuo, Y., and Zhang, D. (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases. *Journal of Theoretical Biology* 250:362-374.
20. Nothnagel, M., Furst, R., and Rohde K. (2002) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Human Heredity* 54:186-198.
21. Shannon, C.E. (1948) A mathematical theory of communications. *The Bell System Technical Journal* XXVII, 379-423 and 623-656.
22. Cover, T.M., and Thomas, J.A. (2006) *Elements of Information Theory*, 2nd edition. Wiley-Interscience.
23. Jakulin, A. (2005) *Machine Learning Based on Attribute Interactions*. PhD thesis.

24. Jakulin, A., Bratko, I. (2003) Analyzing attribute interactions. *Lect. Notes Artif. Intell.* 2838:229-240.
25. Jakulin, A., Bratko I. (2004) Testing the significance of attribute interactions. *Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 409-416*, edited by Greiner, R., Schuurmans, D.
26. Jakulin, A., Bratko, I., Smrke, D., Demsar, J., Zupan, B. (2003) Attribute interactions in medical data analysis. *Lect. Notes Artif. Intell.* 2780:229-238.
27. Moore, J.H., Gilbert, J.C., Tsai, C.T., Chiang, F.T., Holden, T., Barney, N., and White B.C. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology* 241:252-261.
28. Wu, X., Jin, L., and Xiong, M.M. (2009) Mutual information for testing gene-environmental interaction. *PLoS One*, e4578.
29. McGill, W. J. (1954) Multivariate information transmission. *Psychometrika* 19:97-116.
30. Watanabe, S. (1960) Information theoretical analysis of multivariate correlation. *IBM. J. Res. Dev.* 4:66-82.
31. Chanda, P., Zhang, A., Brazeau, D., Sucheston, L., Freudenheim, J.L., Ambrosone, C., and Ramanathan, M. (2007) Information-theoretic metrics for visualizing gene environment interactions. *The American Journal of Human Genetics* 81:939-863.
32. Han, T. S. (1980) Multiple mutual informations and multiple interactions in frequency data. *Information and Control* 46:26-45.

33. Yeung, R. W. (1991) A new outlook on Shannons information measures. *IEEE Transactions on Information Theory* 37:466-474.
34. Lehmann, E.L. (1983) *Theory of Point Estimation*. John Wiley & Sons.
35. Moore, J. H., Hahn, L. W., Ritchie, M. D., Thornton, T. A., and White, B. C. (2002) Applications of genetic algorithms to the discovery of complex models for simulation studies in human genetics. In Langdon, W. B., Cantu-Paz, E., Mathias, K., Roy, R., Davis, D., Poli, R., Balakrishnan, K., Honavar, V., Rudolph, G., Wegener, J., Bull, L., Potter, M. A., Schultz, A. C., Miller, J. F., Burke, E. and Jonoska, N. (eds), *Proceedings of the Genetic and Evolutionary Computation Conference*. Morgan Kaufmann, San Francisco, 1150-1155.
36. Martin, E.R., Ritchie, M.D., Hahn, L., Kang, S., and Moore, J.H. (2006) A novel method to identify gene-gene effects in nuclear families: The MDR-PDT. *Genetic Epidemiology* 30:111-123.
37. Lou, X.Y., Chen, G.B., Yan, L., Ma, J.Z., Mangold1, J.E., Zhu, J., Elston, R.C., and Li, M.D. (2008) A combinatorial approach to detecting gene-gene and gene-environment interactions in family studies. *The American Journal of Human Genetics* 83:457-467.

Table 1: Results of the 2-way test statistic T_{IG} and the 3-way test statistics T_{IIG} and T_{TCIG} of the bladder cancer data of Andrew et al (2006)². * - the most significant result of T_{IG} ; # - the second most significant result of T_{IG} ; † - the only significant result of 3-way interaction information gain test statistic T_{IIG} at 5% significance level.

No. of Factors	SNPs	Test	P-value
2-way interaction	Xpd.751, Xpd.312*	$T_{IG} = 51.62$	6.75e-13
	XRCC1.194, XPC PAT#	$T_{IG} = 2.47$	0.12
3-way interaction and 3-way correlation	Xpd.751, Xpd.312, APE1 148	$T_{TCIG} = 44.54$	2.49e-11
		$T_{IIG} = 0.19$	0.66
	Xpd.751, Xpd.312, XPC PAT	$T_{TCIG} = 43.87$	3.51e-11
		$T_{IIG} = 0.12$	0.73
	Xpd.751, Xpd.312, Pack.years	$T_{TCIG} = 40.93$	1.58e-10
		$T_{IIG} = 0.11$	0.74
	Xpd.751, Xpd.312, XRCC3.241	$T_{TCIG} = 40.20$	2.29e-10
		$T_{IIG} = 0.12$	0.73
	Xpd.751, Xpd.312, XRCC1.399	$T_{TCIG} = 39.90$	2.68e-10
		$T_{IIG} = 1.46$	0.23
	Xpd.751, Xpd.312, XRCC1.194	$T_{TCIG} = 35.41$	2.67e-9
		$T_{IIG} = 0.30$	0.58
	Xpd.751, XRCC1.194, XRCC3.241	$T_{TCIG} = 5.67$	0.02
		$T_{IIG} = 1.80$	0.18
XRCC1.399, XRCC1.194, XRCC3.241	$T_{TCIG} = 3.67$	0.06	
	$T_{IIG} = 4.25$	0.04 [†]	
XPC PAT, XRCC1.194, Pack.years	$T_{TCIG} = 3.97$	0.05	
	$T_{IIG} = 0.21$	0.65	

Table 2: Type I error rates of 2-way test statistic T_{IG} and 3-way test statistics T_{HIG} and T_{TCIG} based on the joint genotype frequencies of bladder cancer data at nominal levels $\alpha = 0.01$ and $\alpha = 0.05$. Each of the entries is based on 100,000 simulations. Pack_years is used as a SNP in simulating its three category genotypes.

α	Test	SNPs used to generate Joint Genotype Frequency	Sample Sizes $M = N$									
			100	150	200	250	300	400	500	600	700	
0.01	T_{IG}	Xpd_751, Xpd_312	0.01499	0.01357	0.01200	0.01184	0.01157	0.01070	0.01094	0.01130	0.01039	
			0.01701	0.01570	0.01434	0.01382	0.01297	0.01262	0.01159	0.01194	0.01122	
	T_{TCIG}	Xpd_751, Xpd_312, XPC PAT Xpd_751, Xpd_312, Pack_years Xpd_751, Xpd_312, XRCC3_241 Xpd_751, Xpd_312, XRCC1_399 Xpd_751, Xpd_312, XRCC1_194 APE1, XRCC1_399, XRCC1_194 Xpd_751, Xpd_312, XRCC1_194 XRCC3_241, APE1, XRCC1_194 XRCC3_241, APE1, XRCC1_399 XRCC3_241, XRCC1_399, XRCC1_194	0.01886	0.01568	0.01321	0.01280	0.01208	0.01154	0.01058	0.01051	0.01060	
			0.01888	0.01572	0.01468	0.01366	0.01235	0.01249	0.01223	0.01112	0.01063	
			0.01636	0.01484	0.01289	0.01244	0.01175	0.01176	0.01099	0.01074	0.01059	
			0.01751	0.01454	0.01340	0.01212	0.01234	0.01153	0.01098	0.01168	0.01058	
			0.01418	0.01259	0.01212	0.01203	0.01031	0.01041	0.01027	0.00971	0.01001	
			0.00948	0.00976	0.00908	0.00857	0.0079	0.00743	0.00752	0.00795	0.00753	
			0.00808	0.00961	0.01057	0.01133	0.01092	0.01053	0.00950	0.00924	0.00843	
			0.00939	0.00733	0.00648	0.00665	0.00677	0.00663	0.00794	0.00848	0.00889	
0.05	T_{IG}	Xpd_751, Xpd_312	0.0119	0.01005	0.00887	0.00848	0.00796	0.00762	0.00852	0.00820	0.00845	
			0.06148	0.05813	0.05510	0.05575	0.05316	0.05288	0.05173	0.05097	0.05091	
	T_{TCIG}	Xpd_751, Xpd_312, XPC PAT Xpd_751, Xpd_312, Pack_years Xpd_751, Xpd_312, XRCC3_241 Xpd_751, Xpd_312, XRCC1_399 Xpd_751, Xpd_312, XRCC1_194 APE1, XRCC1_399, XRCC1_194 Xpd_751, Xpd_312, XRCC1_194 XRCC3_241, APE1, XRCC1_194 XRCC3_241, APE1, XRCC1_399 XRCC3_241, XRCC1_399, XRCC1_194	0.06852	0.06355	0.06171	0.05974	0.05776	0.0548	0.05584	0.05373	0.05356	
			0.06886	0.06400	0.05987	0.05697	0.05601	0.05478	0.05337	0.05043	0.05220	
			0.07143	0.06579	0.06289	0.06016	0.05851	0.05541	0.05494	0.05439	0.05311	
			0.06608	0.06200	0.05898	0.05620	0.05451	0.05326	0.05137	0.05281	0.05057	
			0.06594	0.06303	0.05968	0.05701	0.05519	0.05466	0.05297	0.05234	0.05223	
			0.05857	0.05689	0.05453	0.05403	0.05321	0.05059	0.05109	0.04992	0.05136	
			0.0573	0.05309	0.05025	0.0463	0.04481	0.04352	0.04516	0.04511	0.04674	
			0.04551	0.04972	0.05394	0.05424	0.05379	0.05187	0.04912	0.04636	0.04599	
T_{HIG}	XRCC3_241, APE1, XRCC1_194 XRCC3_241, APE1, XRCC1_399 XRCC3_241, XRCC1_399, XRCC1_194	0.05195	0.04382	0.04132	0.04125	0.04072	0.04327	0.04446	0.04600	0.04696		
		0.04834	0.04498	0.04412	0.04328	0.04369	0.04312	0.04446	0.04379	0.04317		
			0.06258	0.05403	0.04737	0.04526	0.04447	0.04672	0.04736	0.04800		

Table 3: Six models of two-locus penetrance functions and allele frequencies taken from Moore et al. (2002)³⁵, Figures 5-10.

(a) **Model 1**, $P_A = P_B = 0.5$

	BB	Bb	bb
AA	0.083	0.076	0.964
Aa	0.056	0.508	0.085
aa	0.977	0.098	0.062

(b) **Model 2**, $P_A = P_B = 0.5$

	BB	Bb	bb
AA	0.094	0.905	0.097
Aa	0.967	0.097	0.937
aa	0.027	0.990	0.080

(c) **Model 3**, $P_A = P_B = 0.5$

	BB	Bb	bb
AA	0.967	0.314	0.137
Aa	0.313	0.312	0.742
aa	0.129	0.779	0.075

(d) **Model 4**, $P_A = P_B = 0.5$

	BB	Bb	bb
AA	0.967	0.139	0.799
Aa	0.057	0.655	0.627
aa	0.974	0.544	0.019

(e) **Model 5**, $P_A = P_B = 0.5$

	BB	Bb	bb
AA	0.017	0.451	0.711
Aa	0.520	0.571	0.039
aa	0.640	0.053	0.949

(f) **Model 6**, $P_A = P_B = 0.5$

	BB	Bb	bb
AA	0.954	0.256	0.360
Aa	0.010	0.731	0.300
aa	0.801	0.093	0.808

Table 4: Four models of two-locus penetrance functions and allele frequencies taken from Ritchie et al. (2003)¹², Figure 2.

(a) **Model 1**, $P_A = P_B = 0.25$

	BB	Bb	bb
AA	.08	.07	.05
Aa	.10	0	.10
aa	.03	.10	.04

(b) **Model 2**, $P_A = P_B = 0.25$

	BB	Bb	bb
AA	0	.01	.09
Aa	.04	.01	.08
aa	.07	.09	.03

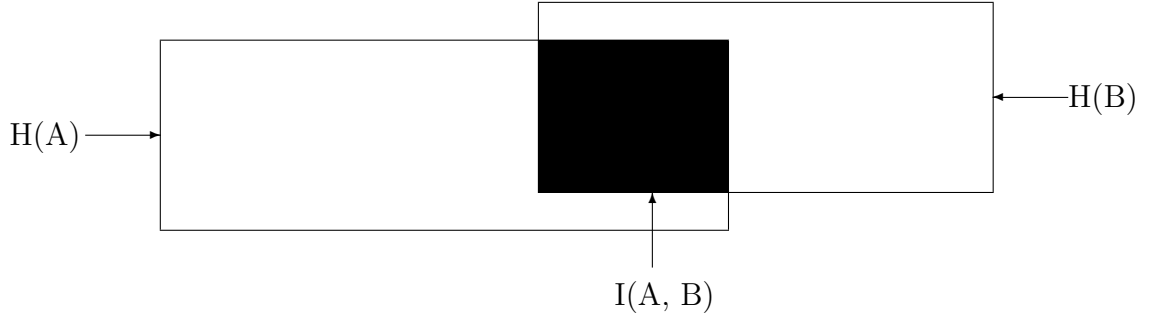
(c) **Model 3**, $P_A = P_B = 0.1$

	BB	Bb	bb
AA	.07	.05	.02
Aa	.05	.09	.01
aa	.02	.01	.03

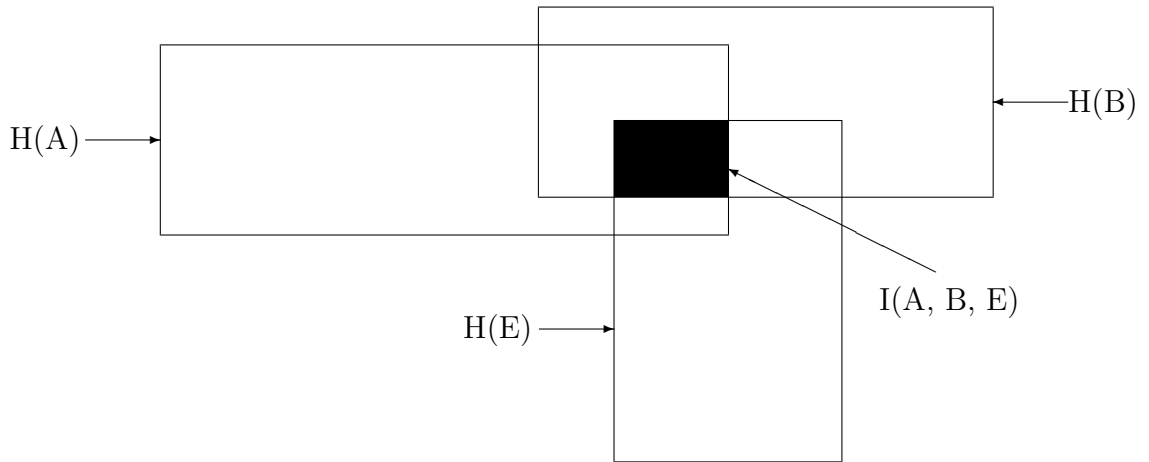
(d) **Model 4**, $P_A = P_B = 0.1$

	BB	Bb	bb
AA	.09	.001	.02
Aa	.08	.07	.005
aa	.003	.007	.02

a) I-diagram of $H(A)$, $H(B)$, and $I(A, B)$



b) I-diagram of $H(A)$, $H(B)$, $H(E)$, and $I(A, B, E)$



c) I-diagram of $H(A)$, $H(B)$, $H(E)$, and $TCI(A, B, E)$

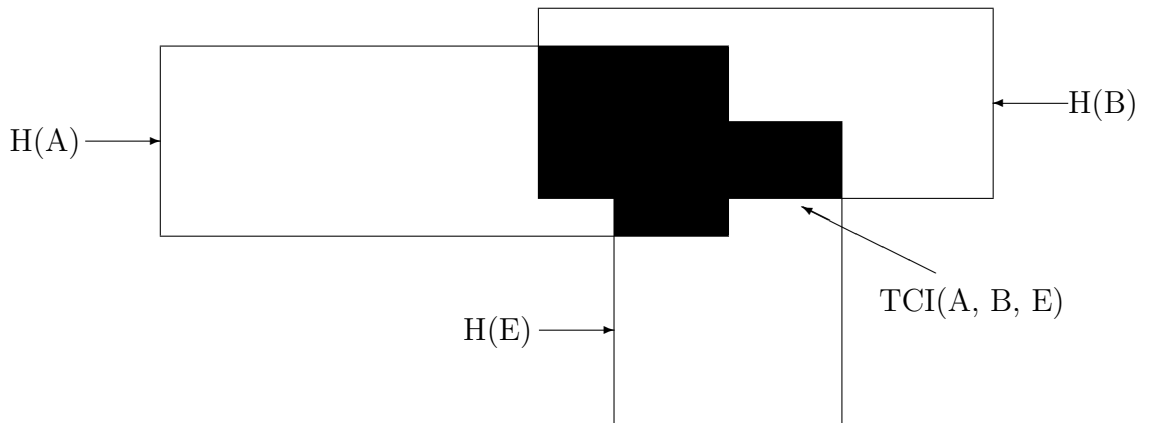


Figure 1: a) The I-diagram of entropies $H(A)$, $H(B)$, and 2-way mutual information $I(A, B)$; b) The I-diagram of entropies $H(A)$, $H(B)$, $H(E)$, and 3-way interaction information $I(A, B, E)$; c) The I-diagram of entropies $H(A)$, $H(B)$, $H(E)$, and 3-way total correlation information $TCI(A, B, E)$.

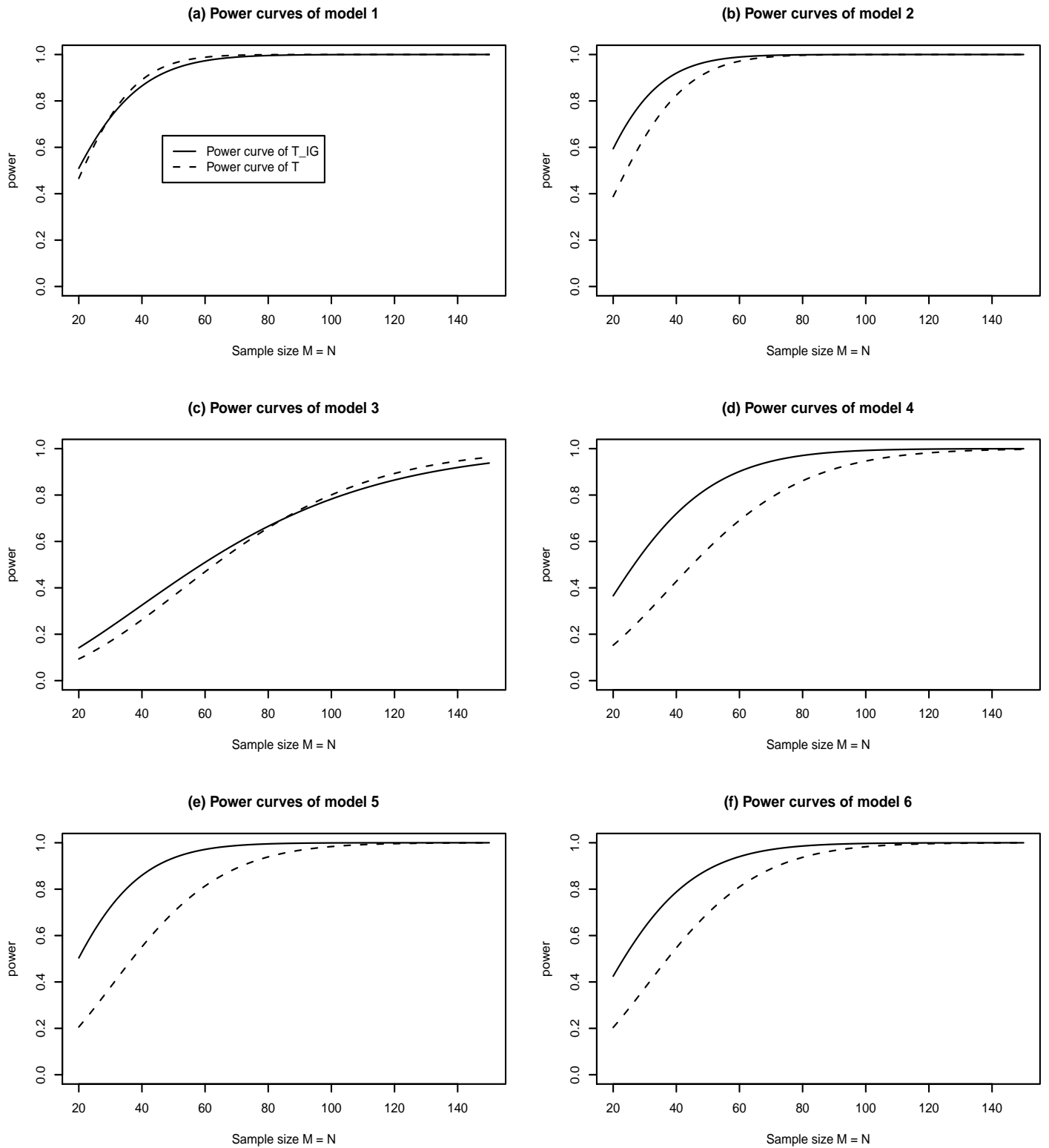


Figure 2: The power curves of test statistics T_{IG} and T at a significance level $\alpha = 0.01$ for the six models of Table 3.

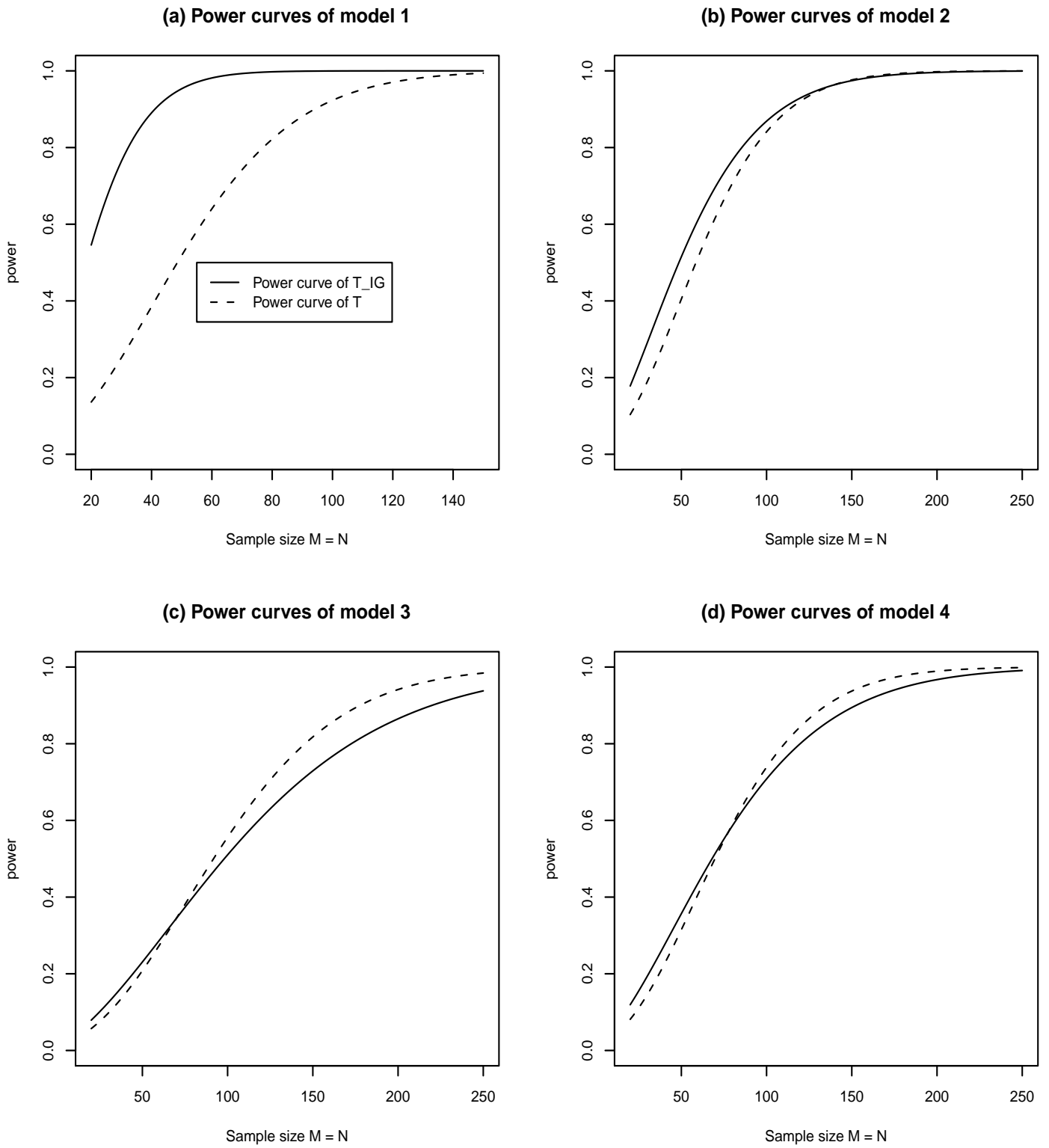


Figure 3: The power curves of test statistics T_{IG} and T at a significance level $\alpha = 0.01$ for the four models of Table 4.

Supplementary Materials

Appendix A Proof of Relations (16)

A.1 The Subscripts ije Do Not Contain 2

Notice

$$\begin{aligned}
 \frac{\partial f_{000}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{000} \log \frac{P_{000}}{P_{0..}P_{.0.}P_{..0}} \right] = \log \frac{P_{000}}{P_{0..}P_{.0.}P_{..0}} + \left[1 - \frac{P_{000}}{P_{0..}} - \frac{P_{000}}{P_{.0.}} - \frac{P_{000}}{P_{..0}} \right] \log e, \\
 \frac{\partial f_{001}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{001} \log \frac{P_{001}}{P_{0..}P_{.0.}P_{..1}} \right] = - \left[\frac{P_{001}}{P_{0..}} + \frac{P_{001}}{P_{.0.}} \right] \log e, \\
 \frac{\partial f_{002}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{002} \log \frac{P_{002}}{P_{0..}P_{.0.}P_{..2}} \right] = \left[-\frac{P_{002}}{P_{0..}} - \frac{P_{002}}{P_{.0.}} + \frac{P_{002}}{P_{..2}} \right] \log e, \\
 \frac{\partial f_{010}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{010} \log \frac{P_{010}}{P_{0..}P_{.1.}P_{..0}} \right] = - \left[\frac{P_{010}}{P_{0..}} + \frac{P_{010}}{P_{.0.}} \right] \log e, \\
 \frac{\partial f_{011}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{011} \log \frac{P_{011}}{P_{0..}P_{.1.}P_{..1}} \right] = -\frac{P_{011}}{P_{0..}} \log e, \\
 \frac{\partial f_{012}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{012} \log \frac{P_{012}}{P_{0..}P_{.1.}P_{..2}} \right] = \left[-\frac{P_{012}}{P_{0..}} + \frac{P_{012}}{P_{..2}} \right] \log e, \\
 \frac{\partial f_{020}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{020} \log \frac{P_{020}}{P_{0..}P_{.2.}P_{..0}} \right] = \left[-\frac{P_{020}}{P_{0..}} + \frac{P_{020}}{P_{.2.}} - \frac{P_{020}}{P_{..0}} \right] \log e, \\
 \frac{\partial f_{021}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{021} \log \frac{P_{021}}{P_{0..}P_{.2.}P_{..1}} \right] = \left[-\frac{P_{021}}{P_{0..}} + \frac{P_{021}}{P_{.2.}} \right] \log e, \\
 \frac{\partial f_{022}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{022} \log \frac{P_{022}}{P_{0..}P_{.2.}P_{..2}} \right] = \left[-\frac{P_{022}}{P_{0..}} + \frac{P_{022}}{P_{.2.}} + \frac{P_{022}}{P_{..2}} \right] \log e.
 \end{aligned} \tag{A.1}$$

In addition, we have

$$\begin{aligned}
 \frac{\partial f_{100}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{100} \log \frac{P_{100}}{P_{1..}P_{.0.}P_{..0}} \right] = - \left[\frac{P_{100}}{P_{.0.}} + \frac{P_{100}}{P_{..0}} \right] \log e, \\
 \frac{\partial f_{101}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{101} \log \frac{P_{101}}{P_{1..}P_{.0.}P_{..1}} \right] = -\frac{P_{101}}{P_{.0.}} \log e, \\
 \frac{\partial f_{102}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{102} \log \frac{P_{102}}{P_{1..}P_{.0.}P_{..2}} \right] = \left[-\frac{P_{102}}{P_{.0.}} + \frac{P_{102}}{P_{..2}} \right] \log e, \\
 \frac{\partial f_{110}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{110} \log \frac{P_{110}}{P_{1..}P_{.1.}P_{..0}} \right] = -\frac{P_{110}}{P_{.0.}} \log e, \\
 \frac{\partial f_{111}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{111} \log \frac{P_{111}}{P_{1..}P_{.1.}P_{..1}} \right] = 0, \\
 \frac{\partial f_{112}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{112} \log \frac{P_{112}}{P_{1..}P_{.1.}P_{..2}} \right] = \frac{P_{112}}{P_{.2.}} \log e, \\
 \frac{\partial f_{120}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{120} \log \frac{P_{120}}{P_{1..}P_{.2.}P_{..0}} \right] = \left[\frac{P_{120}}{P_{.2.}} - \frac{P_{120}}{P_{..0}} \right] \log e,
 \end{aligned} \tag{A.2}$$

$$\begin{aligned}\frac{\partial f_{121}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{121} \log \frac{P_{121}}{P_{1..}P_{.2.}P_{..1}} \right] = \frac{P_{121}}{P_{.2.}} \log e, \\ \frac{\partial f_{122}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{122} \log \frac{P_{122}}{P_{1..}P_{.2.}P_{..2}} \right] = \left[\frac{P_{122}}{P_{.2.}} + \frac{P_{122}}{P_{..2}} \right] \log e,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial f_{200}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{200} \log \frac{P_{200}}{P_{2..}P_{.0.}P_{..0}} \right] = \left[\frac{P_{200}}{P_{2..}} - \frac{P_{200}}{P_{.0.}} - \frac{P_{200}}{P_{..0}} \right] \log e, \\ \frac{\partial f_{201}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{201} \log \frac{P_{201}}{P_{2..}P_{.0.}P_{..1}} \right] = \left[\frac{P_{201}}{P_{2..}} - \frac{P_{201}}{P_{.0.}} \right] \log e, \\ \frac{\partial f_{202}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{202} \log \frac{P_{202}}{P_{2..}P_{.0.}P_{..2}} \right] = \left[\frac{P_{202}}{P_{2..}} - \frac{P_{202}}{P_{.0.}} + \frac{P_{202}}{P_{..2}} \right] \log e, \\ \frac{\partial f_{210}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{210} \log \frac{P_{210}}{P_{2..}P_{.1.}P_{..0}} \right] = \left[\frac{P_{210}}{P_{2..}} - \frac{P_{210}}{P_{.1.}} \right] \log e, \\ \frac{\partial f_{211}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{211} \log \frac{P_{211}}{P_{2..}P_{.1.}P_{..1}} \right] = \frac{P_{211}}{P_{2..}} \log e, \\ \frac{\partial f_{212}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{212} \log \frac{P_{212}}{P_{2..}P_{.1.}P_{..2}} \right] = \left[\frac{P_{212}}{P_{2..}} + \frac{P_{212}}{P_{..2}} \right] \log e, \\ \frac{\partial f_{220}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{220} \log \frac{P_{220}}{P_{2..}P_{.2.}P_{..0}} \right] = \left[\frac{P_{220}}{P_{2..}} + \frac{P_{220}}{P_{.2.}} - \frac{P_{220}}{P_{..0}} \right] \log e, \\ \frac{\partial f_{221}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{221} \log \frac{P_{221}}{P_{2..}P_{.2.}P_{..1}} \right] = \left[\frac{P_{221}}{P_{2..}} + \frac{P_{221}}{P_{.2.}} \right] \log e, \\ \frac{\partial f_{222}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{222} \log \frac{P_{222}}{P_{2..}P_{.2.}P_{..2}} \right] = -\log \frac{P_{222}}{P_{2..}P_{.2.}P_{..2}} + \left[-1 + \frac{P_{222}}{P_{2..}} + \frac{P_{222}}{P_{.2.}} + \frac{P_{222}}{P_{..2}} \right] \log e.\end{aligned} \tag{A.3}$$

By relations (A.1), (A.2) and (A.3), we have

$$\begin{aligned}\frac{\partial f}{\partial P_{000}} &= \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 \frac{\partial f_{ije}}{\partial P_{000}} \\ &= \log \frac{P_{000}}{P_{0..}P_{.0.}P_{..0}} - \log \frac{P_{222}}{P_{2..}P_{.2.}P_{..2}}.\end{aligned}$$

Similarly, we have for $i, j, e = 0, 1$

$$\frac{\partial f}{\partial P_{ije}} = \log \frac{P_{ije}}{P_{i..}P_{.j.}P_{..e}} - \log \frac{P_{222}}{P_{2..}P_{.2.}P_{..2}}.$$

A.2 The Subscripts ije Contain One 2

Notice

$$\frac{\partial f_{000}}{\partial P_{002}} = \frac{\partial}{\partial P_{002}} \left[P_{000} \log \frac{P_{000}}{P_{0..}P_{.0.}P_{..0}} \right] = \left[-\frac{P_{000}}{P_{0..}} - \frac{P_{000}}{P_{.0.}} \right] \log e,$$

$$\begin{aligned}
\frac{\partial f_{001}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{001} \log \frac{P_{001}}{P_{0..}P_{.0}P_{..1}} \right] = \left[-\frac{P_{001}}{P_{0..}} - \frac{P_{001}}{P_{.0}} \right] \log e, \\
\frac{\partial f_{002}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{002} \log \frac{P_{002}}{P_{0..}P_{.0}P_{..2}} \right] = \log \frac{P_{002}}{P_{0..}P_{.0}P_{..2}} + \left[1 - \frac{P_{002}}{P_{0..}} - \frac{P_{002}}{P_{.0}} \right] \log e, \\
\frac{\partial f_{010}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{010} \log \frac{P_{010}}{P_{0..}P_{.1}P_{..0}} \right] = -\frac{P_{010}}{P_{0..}} \log e, \\
\frac{\partial f_{011}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{011} \log \frac{P_{011}}{P_{0..}P_{.1}P_{..1}} \right] = -\frac{P_{011}}{P_{0..}} \log e, \\
\frac{\partial f_{012}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{012} \log \frac{P_{012}}{P_{0..}P_{.1}P_{..2}} \right] = -\frac{P_{012}}{P_{0..}} \log e, \\
\frac{\partial f_{020}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{020} \log \frac{P_{020}}{P_{0..}P_{.2}P_{..0}} \right] = \left[-\frac{P_{020}}{P_{0..}} + \frac{P_{020}}{P_{.2}} \right] \log e, \\
\frac{\partial f_{021}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{021} \log \frac{P_{021}}{P_{0..}P_{.2}P_{..1}} \right] = \left[-\frac{P_{021}}{P_{0..}} + \frac{P_{021}}{P_{.2}} \right] \log e, \\
\frac{\partial f_{022}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{022} \log \frac{P_{022}}{P_{0..}P_{.2}P_{..2}} \right] = \left[-\frac{P_{022}}{P_{0..}} + \frac{P_{022}}{P_{.2}} \right] \log e.
\end{aligned} \tag{A.4}$$

In addition, we have

$$\begin{aligned}
\frac{\partial f_{100}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{100} \log \frac{P_{100}}{P_{1..}P_{.0}P_{..0}} \right] = -\frac{P_{100}}{P_{.0}} \log e, \\
\frac{\partial f_{101}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{101} \log \frac{P_{101}}{P_{1..}P_{.0}P_{..1}} \right] = -\frac{P_{101}}{P_{.0}} \log e, \\
\frac{\partial f_{102}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{102} \log \frac{P_{102}}{P_{1..}P_{.0}P_{..2}} \right] = -\frac{P_{102}}{P_{.0}} \log e, \\
\frac{\partial f_{110}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{110} \log \frac{P_{110}}{P_{1..}P_{.1}P_{..0}} \right] = 0, \\
\frac{\partial f_{111}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{111} \log \frac{P_{111}}{P_{1..}P_{.1}P_{..1}} \right] = 0, \\
\frac{\partial f_{112}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{112} \log \frac{P_{112}}{P_{1..}P_{.1}P_{..2}} \right] = 0, \\
\frac{\partial f_{120}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{120} \log \frac{P_{120}}{P_{1..}P_{.2}P_{..0}} \right] = \frac{P_{120}}{P_{.2}} \log e, \\
\frac{\partial f_{121}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{121} \log \frac{P_{121}}{P_{1..}P_{.2}P_{..1}} \right] = \frac{P_{121}}{P_{.2}} \log e, \\
\frac{\partial f_{122}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{122} \log \frac{P_{122}}{P_{1..}P_{.2}P_{..2}} \right] = \frac{P_{122}}{P_{.2}} \log e,
\end{aligned} \tag{A.5}$$

and

$$\begin{aligned}
\frac{\partial f_{200}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{200} \log \frac{P_{200}}{P_{2..}P_{.0}P_{..0}} \right] = \left[\frac{P_{200}}{P_{2..}} - \frac{P_{200}}{P_{.0}} \right] \log e, \\
\frac{\partial f_{201}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{201} \log \frac{P_{201}}{P_{2..}P_{.0}P_{..1}} \right] = \left[\frac{P_{201}}{P_{2..}} - \frac{P_{201}}{P_{.0}} \right] \log e,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial f_{202}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{202} \log \frac{P_{202}}{P_{2..}P_{.0}P_{..2}} \right] = \left[\frac{P_{202}}{P_{2..}} - \frac{P_{202}}{P_{.0}} \right] \log e, \\
\frac{\partial f_{210}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{210} \log \frac{P_{210}}{P_{2..}P_{.1}P_{..0}} \right] = \frac{P_{210}}{P_{2..}} \log e, \\
\frac{\partial f_{211}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{211} \log \frac{P_{211}}{P_{2..}P_{.1}P_{..1}} \right] = \frac{P_{211}}{P_{2..}} \log e, \\
\frac{\partial f_{212}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{212} \log \frac{P_{212}}{P_{2..}P_{.1}P_{..2}} \right] = \frac{P_{212}}{P_{2..}} \log e, \\
\frac{\partial f_{220}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{220} \log \frac{P_{220}}{P_{2..}P_{.2}P_{..0}} \right] = \left[\frac{P_{220}}{P_{2..}} + \frac{P_{220}}{P_{.2}} \right] \log e, \\
\frac{\partial f_{221}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{221} \log \frac{P_{221}}{P_{2..}P_{.2}P_{..1}} \right] = \left[\frac{P_{221}}{P_{2..}} + \frac{P_{221}}{P_{.2}} \right] \log e, \\
\frac{\partial f_{222}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{222} \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}} \right] = -\log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}} + \left[-1 + \frac{P_{222}}{P_{2..}} + \frac{P_{222}}{P_{.2}} \right] \log e.
\end{aligned} \tag{A.6}$$

By relations (A.4), (A.5) and (A.6), we have

$$\begin{aligned}
\frac{\partial f}{\partial P_{002}} &= \sum_{i=0}^2 \sum_{j=0}^2 \sum_{\epsilon=0}^2 \frac{\partial f_{ije}}{\partial P_{002}} \\
&= \log \frac{P_{002}}{P_{0..}P_{.0}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\frac{\partial f}{\partial P_{012}} &= \log \frac{P_{012}}{P_{0..}P_{.1}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{020}} &= \log \frac{P_{020}}{P_{0..}P_{.2}P_{..0}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{021}} &= \log \frac{P_{021}}{P_{0..}P_{.2}P_{..1}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{102}} &= \log \frac{P_{102}}{P_{1..}P_{.0}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{112}} &= \log \frac{P_{112}}{P_{1..}P_{.1}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{120}} &= \log \frac{P_{120}}{P_{1..}P_{.2}P_{..0}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{121}} &= \log \frac{P_{121}}{P_{1..}P_{.2}P_{..1}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{200}} &= \log \frac{P_{200}}{P_{2..}P_{.0}P_{..0}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{201}} &= \log \frac{P_{201}}{P_{2..}P_{.0}P_{..1}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}},
\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial P_{210}} &= \log \frac{P_{210}}{P_{2..}P_{.1}P_{.0}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{.2}}, \\ \frac{\partial f}{\partial P_{211}} &= \log \frac{P_{211}}{P_{2..}P_{.1}P_{.1}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{.2}}.\end{aligned}$$

A.3 The Subscripts ije Contain Two 2

Notice

$$\begin{aligned}\frac{\partial f_{000}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{000} \log \frac{P_{000}}{P_{0..}P_{.0}P_{.0}} \right] = 0, \\ \frac{\partial f_{001}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{001} \log \frac{P_{001}}{P_{0..}P_{.0}P_{.1}} \right] = 0, \\ \frac{\partial f_{002}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{002} \log \frac{P_{002}}{P_{0..}P_{.0}P_{.2}} \right] = 0, \\ \frac{\partial f_{010}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{010} \log \frac{P_{010}}{P_{0..}P_{.1}P_{.0}} \right] = 0, \\ \frac{\partial f_{011}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{011} \log \frac{P_{011}}{P_{0..}P_{.1}P_{.1}} \right] = 0, \\ \frac{\partial f_{012}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{012} \log \frac{P_{012}}{P_{0..}P_{.1}P_{.2}} \right] = 0, \\ \frac{\partial f_{020}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{020} \log \frac{P_{020}}{P_{0..}P_{.2}P_{.0}} \right] = 0, \\ \frac{\partial f_{021}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{021} \log \frac{P_{021}}{P_{0..}P_{.2}P_{.1}} \right] = 0, \\ \frac{\partial f_{022}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{022} \log \frac{P_{022}}{P_{0..}P_{.2}P_{.2}} \right] = 0.\end{aligned}\tag{A.7}$$

In addition, we have

$$\begin{aligned}\frac{\partial f_{100}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{100} \log \frac{P_{100}}{P_{1..}P_{.0}P_{.0}} \right] = -\frac{P_{100}}{P_{1..}} \log e, \\ \frac{\partial f_{101}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{101} \log \frac{P_{101}}{P_{1..}P_{.0}P_{.1}} \right] = -\frac{P_{101}}{P_{1..}} \log e, \\ \frac{\partial f_{102}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{102} \log \frac{P_{102}}{P_{1..}P_{.0}P_{.2}} \right] = -\frac{P_{102}}{P_{1..}} \log e, \\ \frac{\partial f_{110}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{110} \log \frac{P_{110}}{P_{1..}P_{.1}P_{.0}} \right] = -\frac{P_{110}}{P_{1..}} \log e, \\ \frac{\partial f_{111}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{111} \log \frac{P_{111}}{P_{1..}P_{.1}P_{.1}} \right] = -\frac{P_{111}}{P_{1..}} \log e, \\ \frac{\partial f_{112}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{112} \log \frac{P_{112}}{P_{1..}P_{.1}P_{.2}} \right] = -\frac{P_{112}}{P_{1..}} \log e,\end{aligned}\tag{A.8}$$

$$\begin{aligned}
\frac{\partial f_{120}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{120} \log \frac{P_{120}}{P_{1..}P_{.2}P_{..0}} \right] = -\frac{P_{120}}{P_{1..}} \log e, \\
\frac{\partial f_{121}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{121} \log \frac{P_{121}}{P_{1..}P_{.2}P_{..1}} \right] = -\frac{P_{121}}{P_{1..}} \log e, \\
\frac{\partial f_{122}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{122} \log \frac{P_{122}}{P_{1..}P_{.2}P_{..2}} \right] = \log \frac{P_{122}}{P_{1..}P_{.2}P_{..2}} + \left[1 - \frac{P_{122}}{P_{1..}} \right] \log e,
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial f_{200}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{200} \log \frac{P_{200}}{P_{2..}P_{.0}P_{..0}} \right] = \frac{P_{200}}{P_{2..}} \log e, \\
\frac{\partial f_{201}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{201} \log \frac{P_{201}}{P_{2..}P_{.0}P_{..1}} \right] = \frac{P_{201}}{P_{2..}} \log e, \\
\frac{\partial f_{202}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{202} \log \frac{P_{202}}{P_{2..}P_{.0}P_{..2}} \right] = \frac{P_{202}}{P_{2..}} \log e, \\
\frac{\partial f_{210}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{210} \log \frac{P_{210}}{P_{2..}P_{.1}P_{..0}} \right] = \frac{P_{210}}{P_{2..}} \log e, \\
\frac{\partial f_{211}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{211} \log \frac{P_{211}}{P_{2..}P_{.1}P_{..1}} \right] = \frac{P_{211}}{P_{2..}} \log e, \\
\frac{\partial f_{212}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{212} \log \frac{P_{212}}{P_{2..}P_{.1}P_{..2}} \right] = \frac{P_{212}}{P_{2..}} \log e, \\
\frac{\partial f_{220}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{220} \log \frac{P_{220}}{P_{2..}P_{.2}P_{..0}} \right] = \frac{P_{220}}{P_{2..}} \log e, \\
\frac{\partial f_{221}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{221} \log \frac{P_{221}}{P_{2..}P_{.2}P_{..1}} \right] = \frac{P_{221}}{P_{2..}} \log e, \\
\frac{\partial f_{222}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{222} \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}} \right] = -\log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}} + \left[-1 + \frac{P_{222}}{P_{2..}} \right] \log e.
\end{aligned} \tag{A.9}$$

By relations (A.7), (A.8) and (A.9), we have

$$\begin{aligned}
\frac{\partial f}{\partial P_{122}} &= \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 \frac{\partial f_{ije}}{\partial P_{122}} \\
&= \log \frac{P_{122}}{P_{1..}P_{.2}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\frac{\partial f}{\partial P_{022}} &= \log \frac{P_{022}}{P_{0..}P_{.2}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{202}} &= \log \frac{P_{202}}{P_{2..}P_{.0}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}}, \\
\frac{\partial f}{\partial P_{212}} &= \log \frac{P_{212}}{P_{2..}P_{.1}P_{..2}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{..2}},
\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial P_{220}} &= \log \frac{P_{220}}{P_{2..}P_{.2}P_{.0}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{.2}}, \\ \frac{\partial f}{\partial P_{221}} &= \log \frac{P_{221}}{P_{2..}P_{.2}P_{.1}} - \log \frac{P_{222}}{P_{2..}P_{.2}P_{.2}}.\end{aligned}$$

Appendix B Proof of Relations (17)

B.1 The Subscripts ije Do Not Contain 2

Notice

$$\begin{aligned}\frac{\partial h_{000}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{000} \log \frac{P_{000}P_{0..}P_{.0}P_{.0}}{P_{00..}P_{0.0}P_{00}} \right] \\ &= \log \frac{P_{000}P_{0..}P_{.0}P_{.0}}{P_{00..}P_{0.0}P_{00}} + \left[1 + \frac{P_{000}}{P_{0..}} + \frac{P_{000}}{P_{.0}} + \frac{P_{000}}{P_{.0}} - \frac{P_{000}}{P_{00..}} - \frac{P_{000}}{P_{0.0}} - \frac{P_{000}}{P_{00}} \right] \log e, \\ \frac{\partial h_{001}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{001} \log \frac{P_{001}P_{0..}P_{.0}P_{.1}}{P_{00..}P_{0.1}P_{01}} \right] = \left[\frac{P_{001}}{P_{0..}} + \frac{P_{001}}{P_{.0}} - \frac{P_{001}}{P_{00..}} \right] \log e, \\ \frac{\partial h_{002}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{002} \log \frac{P_{002}P_{0..}P_{.0}P_{.2}}{P_{00..}P_{0.2}P_{02}} \right] = \left[\frac{P_{002}}{P_{0..}} + \frac{P_{002}}{P_{.0}} - \frac{P_{002}}{P_{.2}} - \frac{P_{002}}{P_{00..}} \right] \log e, \\ \frac{\partial h_{010}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{010} \log \frac{P_{010}P_{0..}P_{.1}P_{.0}}{P_{01..}P_{0.0}P_{10}} \right] = \left[\frac{P_{010}}{P_{0..}} + \frac{P_{010}}{P_{.0}} - \frac{P_{010}}{P_{0.0}} \right] \log e, \\ \frac{\partial h_{011}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{011} \log \frac{P_{011}P_{0..}P_{.1}P_{.1}}{P_{01..}P_{0.1}P_{11}} \right] = \frac{P_{011}}{P_{0..}} \log e, \\ \frac{\partial h_{012}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{012} \log \frac{P_{012}P_{0..}P_{.1}P_{.2}}{P_{01..}P_{0.2}P_{12}} \right] = \left[\frac{P_{012}}{P_{0..}} - \frac{P_{012}}{P_{.2}} \right] \log e, \\ \frac{\partial h_{020}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{020} \log \frac{P_{020}P_{0..}P_{.2}P_{.0}}{P_{02..}P_{0.0}P_{20}} \right] = \left[\frac{P_{020}}{P_{0..}} - \frac{P_{020}}{P_{.2}} + \frac{P_{020}}{P_{.0}} - \frac{P_{020}}{P_{0.0}} \right] \log e, \\ \frac{\partial h_{021}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{021} \log \frac{P_{021}P_{0..}P_{.2}P_{.1}}{P_{02..}P_{0.1}P_{21}} \right] = \left[\frac{P_{021}}{P_{0..}} - \frac{P_{021}}{P_{.2}} \right] \log e, \\ \frac{\partial h_{022}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{022} \log \frac{P_{022}P_{0..}P_{.2}P_{.2}}{P_{02..}P_{0.2}P_{22}} \right] = \left[\frac{P_{022}}{P_{0..}} - \frac{P_{022}}{P_{.2}} - \frac{P_{022}}{P_{.2}} + \frac{P_{022}}{P_{22}} \right] \log e.\end{aligned} \tag{B.1}$$

In addition, we have

$$\begin{aligned}\frac{\partial h_{100}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{100} \log \frac{P_{100}P_{1..}P_{.0}P_{.0}}{P_{10..}P_{1.0}P_{00}} \right] = \left[\frac{P_{100}}{P_{.0}} + \frac{P_{100}}{P_{.0}} - \frac{P_{100}}{P_{00}} \right] \log e, \\ \frac{\partial h_{101}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{101} \log \frac{P_{101}P_{1..}P_{.0}P_{.1}}{P_{10..}P_{1.1}P_{01}} \right] = \frac{P_{101}}{P_{.0}} \log e, \\ \frac{\partial h_{102}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{102} \log \frac{P_{102}P_{1..}P_{.0}P_{.2}}{P_{10..}P_{1.2}P_{02}} \right] = \left[\frac{P_{102}}{P_{.0}} - \frac{P_{102}}{P_{.2}} \right] \log e, \\ \frac{\partial h_{110}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{110} \log \frac{P_{110}P_{1..}P_{.1}P_{.0}}{P_{11..}P_{1.0}P_{10}} \right] = \frac{P_{110}}{P_{.0}} \log e,\end{aligned}$$

$$\begin{aligned}
\frac{\partial h_{111}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{111} \log \frac{P_{111} P_{1..} P_{.1} P_{..1}}{P_{11.} P_{1.1} P_{.11}} \right] = 0, \\
\frac{\partial h_{112}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{112} \log \frac{P_{112} P_{1..} P_{.1} P_{..2}}{P_{11.} P_{1.2} P_{.12}} \right] = -\frac{P_{112}}{P_{..2}} \log e, \\
\frac{\partial h_{120}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{120} \log \frac{P_{120} P_{1..} P_{.2} P_{..0}}{P_{12.} P_{1.0} P_{.20}} \right] = \left[-\frac{P_{120}}{P_{..2}} + \frac{P_{120}}{P_{..0}} \right] \log e, \\
\frac{\partial h_{121}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{121} \log \frac{P_{121} P_{1..} P_{.2} P_{..1}}{P_{12.} P_{1.1} P_{.21}} \right] = -\frac{P_{121}}{P_{..2}} \log e, \\
\frac{\partial h_{122}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{122} \log \frac{P_{122} P_{1..} P_{.2} P_{..2}}{P_{12.} P_{1.2} P_{.22}} \right] = -\left[\frac{P_{122}}{P_{..2}} + \frac{P_{122}}{P_{..2}} + \frac{P_{122}}{P_{..2}} \right] \log e,
\end{aligned} \tag{B.2}$$

and

$$\begin{aligned}
\frac{\partial h_{200}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{200} \log \frac{P_{200} P_{2..} P_{.0} P_{..0}}{P_{20.} P_{2.0} P_{.00}} \right] = \left[-\frac{P_{200}}{P_{..2}} + \frac{P_{200}}{P_{..0}} + \frac{P_{200}}{P_{..0}} - \frac{P_{200}}{P_{..0}} \right] \log e, \\
\frac{\partial h_{201}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{201} \log \frac{P_{201} P_{2..} P_{.0} P_{..1}}{P_{20.} P_{2.1} P_{.01}} \right] = \left[-\frac{P_{201}}{P_{..2}} + \frac{P_{201}}{P_{..0}} \right] \log e, \\
\frac{\partial h_{202}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{202} \log \frac{P_{202} P_{2..} P_{.0} P_{..2}}{P_{20.} P_{2.2} P_{.02}} \right] = \left[-\frac{P_{202}}{P_{..2}} + \frac{P_{202}}{P_{..0}} - \frac{P_{202}}{P_{..2}} + \frac{P_{202}}{P_{..2}} \right] \log e, \\
\frac{\partial h_{210}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{210} \log \frac{P_{210} P_{2..} P_{.1} P_{..0}}{P_{21.} P_{2.0} P_{.10}} \right] = \left[-\frac{P_{210}}{P_{..2}} + \frac{P_{210}}{P_{..0}} \right] \log e, \\
\frac{\partial h_{211}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{211} \log \frac{P_{211} P_{2..} P_{.1} P_{..1}}{P_{21.} P_{2.1} P_{.11}} \right] = -\frac{P_{211}}{P_{..2}} \log e, \\
\frac{\partial h_{212}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{212} \log \frac{P_{212} P_{2..} P_{.1} P_{..2}}{P_{21.} P_{2.2} P_{.12}} \right] = \left[-\frac{P_{212}}{P_{..2}} - \frac{P_{212}}{P_{..2}} + \frac{P_{212}}{P_{..2}} \right] \log e, \\
\frac{\partial h_{220}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{220} \log \frac{P_{220} P_{2..} P_{.2} P_{..0}}{P_{22.} P_{2.0} P_{.20}} \right] = \left[-\frac{P_{220}}{P_{..2}} - \frac{P_{220}}{P_{..2}} + \frac{P_{220}}{P_{..0}} + \frac{P_{220}}{P_{..2}} \right] \log e, \\
\frac{\partial h_{221}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{221} \log \frac{P_{221} P_{2..} P_{.2} P_{..1}}{P_{22.} P_{2.1} P_{.21}} \right] = \left[-\frac{P_{221}}{P_{..2}} - \frac{P_{221}}{P_{..2}} + \frac{P_{221}}{P_{..2}} \right] \log e, \\
\frac{\partial h_{222}}{\partial P_{000}} &= \frac{\partial}{\partial P_{000}} \left[P_{222} \log \frac{P_{222} P_{2..} P_{.2} P_{..2}}{P_{22.} P_{2.2} P_{.22}} \right] \\
&= -\log \frac{P_{222} P_{2..} P_{.2} P_{..2}}{P_{22.} P_{2.2} P_{.22}} + \left[-1 - \frac{P_{222}}{P_{..2}} - \frac{P_{222}}{P_{..2}} - \frac{P_{222}}{P_{..2}} + \frac{P_{222}}{P_{..2}} + \frac{P_{222}}{P_{..2}} + \frac{P_{222}}{P_{..2}} \right] \log e.
\end{aligned} \tag{B.3}$$

By relations (B.1), (B.2) and (B.3), we have

$$\begin{aligned}
\frac{\partial h}{\partial P_{000}} &= \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 \frac{\partial h_{ije}}{\partial P_{000}} \\
&= \log \frac{P_{000} P_{0..} P_{.0} P_{..0}}{P_{00.} P_{0.0} P_{.00}} - \log \frac{P_{222} P_{2..} P_{.2} P_{..2}}{P_{22.} P_{2.2} P_{.22}}.
\end{aligned}$$

Similarly, we have for $i, j, e = 0, 1$

$$\frac{\partial h}{\partial P_{ije}} = \log \frac{P_{ije} P_{i..} P_{.j} P_{..e}}{P_{ij} P_{i.e} P_{.je}} - \log \frac{P_{222} P_{2..} P_{.2} P_{..2}}{P_{22} P_{2.2} P_{.22}}.$$

B.2 The Subscripts ije Contain One 2

Notice

$$\begin{aligned} \frac{\partial h_{000}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{000} \log \frac{P_{000} P_{0..} P_{.0} P_{..0}}{P_{00} P_{0.0} P_{.00}} \right] = \left[\frac{P_{000}}{P_{0..}} + \frac{P_{000}}{P_{.0}} - \frac{P_{000}}{P_{00}} \right] \log e, \\ \frac{\partial h_{001}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{001} \log \frac{P_{001} P_{0..} P_{.0} P_{..1}}{P_{00} P_{0.1} P_{.01}} \right] = \left[\frac{P_{001}}{P_{0..}} + \frac{P_{001}}{P_{.0}} - \frac{P_{001}}{P_{00}} \right] \log e, \\ \frac{\partial h_{002}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{002} \log \frac{P_{002} P_{0..} P_{.0} P_{..2}}{P_{00} P_{0.2} P_{.02}} \right] \\ &= \log \frac{P_{002} P_{0..} P_{.0} P_{..2}}{P_{00} P_{0.2} P_{.02}} + \left[1 + \frac{P_{002}}{P_{0..}} + \frac{P_{002}}{P_{.0}} - \frac{P_{002}}{P_{00}} - \frac{P_{002}}{P_{0.2}} - \frac{P_{002}}{P_{.02}} \right] \log e, \\ \frac{\partial h_{010}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{010} \log \frac{P_{010} P_{0..} P_{.1} P_{..0}}{P_{01} P_{0.0} P_{.10}} \right] = \frac{P_{010}}{P_{0..}} \log e, \\ \frac{\partial h_{011}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{011} \log \frac{P_{011} P_{0..} P_{.1} P_{..1}}{P_{01} P_{0.1} P_{.11}} \right] = \frac{P_{011}}{P_{0..}} \log e, \\ \frac{\partial h_{012}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{012} \log \frac{P_{012} P_{0..} P_{.1} P_{..2}}{P_{01} P_{0.2} P_{.12}} \right] = \left[\frac{P_{012}}{P_{0..}} - \frac{P_{012}}{P_{0.2}} \right] \log e, \\ \frac{\partial h_{020}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{020} \log \frac{P_{020} P_{0..} P_{.2} P_{..0}}{P_{02} P_{0.0} P_{.20}} \right] = \left[\frac{P_{020}}{P_{0..}} - \frac{P_{020}}{P_{.2}} \right] \log e, \\ \frac{\partial h_{021}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{021} \log \frac{P_{021} P_{0..} P_{.2} P_{..1}}{P_{02} P_{0.1} P_{.21}} \right] = \left[\frac{P_{021}}{P_{0..}} - \frac{P_{021}}{P_{.2}} \right] \log e, \\ \frac{\partial h_{022}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{022} \log \frac{P_{022} P_{0..} P_{.2} P_{..2}}{P_{02} P_{0.2} P_{.22}} \right] = \left[\frac{P_{022}}{P_{0..}} - \frac{P_{022}}{P_{.2}} - \frac{P_{022}}{P_{0.2}} + \frac{P_{022}}{P_{.22}} \right] \log e. \end{aligned} \tag{B.4}$$

In addition, we have

$$\begin{aligned} \frac{\partial h_{100}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{100} \log \frac{P_{100} P_{1..} P_{.0} P_{..0}}{P_{10} P_{1.0} P_{.00}} \right] = \frac{P_{100}}{P_{.0}} \log e, \\ \frac{\partial h_{101}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{101} \log \frac{P_{101} P_{1..} P_{.0} P_{..1}}{P_{10} P_{1.1} P_{.01}} \right] = \frac{P_{101}}{P_{.0}} \log e, \\ \frac{\partial h_{102}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{102} \log \frac{P_{102} P_{1..} P_{.0} P_{..2}}{P_{10} P_{1.2} P_{.02}} \right] = \left[\frac{P_{102}}{P_{.0}} - \frac{P_{102}}{P_{.02}} \right] \log e, \\ \frac{\partial h_{110}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{110} \log \frac{P_{110} P_{1..} P_{.1} P_{..0}}{P_{11} P_{1.0} P_{.10}} \right] = 0, \\ \frac{\partial h_{111}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{111} \log \frac{P_{111} P_{1..} P_{.1} P_{..1}}{P_{11} P_{1.1} P_{.11}} \right] = 0, \\ \frac{\partial h_{112}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{112} \log \frac{P_{112} P_{1..} P_{.1} P_{..2}}{P_{11} P_{1.2} P_{.12}} \right] = 0, \end{aligned} \tag{B.5}$$

$$\begin{aligned}
\frac{\partial h_{120}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{120} \log \frac{P_{120} P_{1..} P_{.2} P_{..0}}{P_{12.} P_{1.0} P_{.20}} \right] = -\frac{P_{120}}{P_{.2}} \log e, \\
\frac{\partial h_{121}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{121} \log \frac{P_{121} P_{1..} P_{.2} P_{..1}}{P_{12.} P_{1.1} P_{.21}} \right] = -\frac{P_{121}}{P_{.2}} \log e, \\
\frac{\partial h_{122}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{122} \log \frac{P_{122} P_{1..} P_{.2} P_{..2}}{P_{12.} P_{1.2} P_{.22}} \right] = \left[-\frac{P_{122}}{P_{.2}} + \frac{P_{122}}{P_{.22}} \right] \log e,
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial h_{200}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{200} \log \frac{P_{200} P_{2..} P_{.0} P_{..0}}{P_{20.} P_{2.0} P_{.00}} \right] = \left[-\frac{P_{200}}{P_{2..}} + \frac{P_{200}}{P_{.0}} \right] \log e, \\
\frac{\partial h_{201}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{201} \log \frac{P_{201} P_{2..} P_{.0} P_{..1}}{P_{20.} P_{2.1} P_{.01}} \right] = \left[-\frac{P_{201}}{P_{2..}} + \frac{P_{201}}{P_{.0}} \right] \log e, \\
\frac{\partial h_{202}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{202} \log \frac{P_{202} P_{2..} P_{.0} P_{..2}}{P_{20.} P_{2.2} P_{.02}} \right] = \left[-\frac{P_{202}}{P_{2..}} + \frac{P_{202}}{P_{.0}} - \frac{P_{202}}{P_{.02}} + \frac{P_{202}}{P_{2.2}} \right] \log e, \\
\frac{\partial h_{210}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{210} \log \frac{P_{210} P_{2..} P_{.1} P_{..0}}{P_{21.} P_{2.0} P_{.10}} \right] = -\frac{P_{210}}{P_{2..}} \log e, \\
\frac{\partial h_{211}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{211} \log \frac{P_{211} P_{2..} P_{.1} P_{..1}}{P_{21.} P_{2.1} P_{.11}} \right] = -\frac{P_{211}}{P_{2..}} \log e, \\
\frac{\partial h_{212}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{212} \log \frac{P_{212} P_{2..} P_{.1} P_{..2}}{P_{21.} P_{2.2} P_{.12}} \right] = \left[-\frac{P_{212}}{P_{2..}} + \frac{P_{212}}{P_{2.2}} \right] \log e, \\
\frac{\partial h_{220}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{220} \log \frac{P_{220} P_{2..} P_{.2} P_{..0}}{P_{22.} P_{2.0} P_{.20}} \right] = \left[-\frac{P_{220}}{P_{2..}} - \frac{P_{220}}{P_{.2}} + \frac{P_{220}}{P_{22.}} \right] \log e, \\
\frac{\partial h_{221}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{221} \log \frac{P_{221} P_{2..} P_{.2} P_{..1}}{P_{22.} P_{2.1} P_{.21}} \right] = \left[-\frac{P_{221}}{P_{2..}} - \frac{P_{221}}{P_{.2}} + \frac{P_{221}}{P_{22.}} \right] \log e, \\
\frac{\partial h_{222}}{\partial P_{002}} &= \frac{\partial}{\partial P_{002}} \left[P_{222} \log \frac{P_{222} P_{2..} P_{.2} P_{..2}}{P_{22.} P_{2.2} P_{.22}} \right] \\
&= -\log \frac{P_{222} P_{2..} P_{.2} P_{..2}}{P_{22.} P_{2.2} P_{.22}} + \left[-1 - \frac{P_{222}}{P_{2..}} - \frac{P_{222}}{P_{.2}} + \frac{P_{222}}{P_{22.}} + \frac{P_{222}}{P_{2.2}} + \frac{P_{222}}{P_{.22}} \right] \log e.
\end{aligned} \tag{B.6}$$

By relations (B.4), (B.5) and (B.6), we have

$$\begin{aligned}
\frac{\partial h}{\partial P_{002}} &= \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 \frac{\partial h_{ije}}{\partial P_{002}} \\
&= \log \frac{P_{002} P_{0..} P_{.0} P_{..2}}{P_{00.} P_{0.2} P_{.02}} - \log \frac{P_{222} P_{2..} P_{.2} P_{..2}}{P_{22.} P_{2.2} P_{.22}}.
\end{aligned}$$

Similarly, we can show relations (17) when the subscripts ije contain one 2.

B.3 The Subscripts ije Contain Two 2

Notice

$$\begin{aligned}
\frac{\partial h_{000}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{000} \log \frac{P_{000} P_{0..} P_{.0.} P_{..0}}{P_{00.} P_{0.0} P_{.00}} \right] = 0, \\
\frac{\partial h_{001}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{001} \log \frac{P_{001} P_{0..} P_{.0.} P_{..1}}{P_{00.} P_{0.1} P_{.01}} \right] = 0, \\
\frac{\partial h_{002}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{002} \log \frac{P_{002} P_{0..} P_{.0.} P_{..2}}{P_{00.} P_{0.2} P_{.02}} \right] = 0, \\
\frac{\partial h_{010}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{010} \log \frac{P_{010} P_{0..} P_{.1.} P_{..0}}{P_{01.} P_{0.0} P_{.10}} \right] = 0, \\
\frac{\partial h_{011}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{011} \log \frac{P_{011} P_{0..} P_{.1.} P_{..1}}{P_{01.} P_{0.1} P_{.11}} \right] = 0, \\
\frac{\partial h_{012}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{012} \log \frac{P_{012} P_{0..} P_{.1.} P_{..2}}{P_{01.} P_{0.2} P_{.12}} \right] = 0, \\
\frac{\partial h_{020}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{020} \log \frac{P_{020} P_{0..} P_{.2.} P_{..0}}{P_{02.} P_{0.0} P_{.20}} \right] = 0, \\
\frac{\partial h_{021}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{021} \log \frac{P_{021} P_{0..} P_{.2.} P_{..1}}{P_{02.} P_{0.1} P_{.21}} \right] = 0, \\
\frac{\partial h_{022}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{022} \log \frac{P_{022} P_{0..} P_{.2.} P_{..2}}{P_{02.} P_{0.2} P_{.22}} \right] = 0.
\end{aligned} \tag{B.7}$$

In addition, we have

$$\begin{aligned}
\frac{\partial h_{100}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{100} \log \frac{P_{100} P_{1..} P_{.0.} P_{..0}}{P_{10.} P_{1.0} P_{.00}} \right] = \frac{P_{100}}{P_{1..}} \log e, \\
\frac{\partial h_{101}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{101} \log \frac{P_{101} P_{1..} P_{.0.} P_{..1}}{P_{10.} P_{1.1} P_{.01}} \right] = \frac{P_{101}}{P_{1..}} \log e, \\
\frac{\partial h_{102}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{102} \log \frac{P_{102} P_{1..} P_{.0.} P_{..2}}{P_{10.} P_{1.2} P_{.02}} \right] = \left[\frac{P_{102}}{P_{1..}} - \frac{P_{102}}{P_{1.2}} \right] \log e, \\
\frac{\partial h_{110}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{110} \log \frac{P_{110} P_{1..} P_{.1.} P_{..0}}{P_{11.} P_{1.0} P_{.10}} \right] = \frac{P_{110}}{P_{1..}} \log e, \\
\frac{\partial h_{111}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{111} \log \frac{P_{111} P_{1..} P_{.1.} P_{..1}}{P_{11.} P_{1.1} P_{.11}} \right] = \frac{P_{111}}{P_{1..}} \log e, \\
\frac{\partial h_{112}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{112} \log \frac{P_{112} P_{1..} P_{.1.} P_{..2}}{P_{11.} P_{1.2} P_{.12}} \right] = \left[\frac{P_{112}}{P_{1..}} - \frac{P_{112}}{P_{1.2}} \right] \log e, \\
\frac{\partial h_{120}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{120} \log \frac{P_{120} P_{1..} P_{.2.} P_{..0}}{P_{12.} P_{1.0} P_{.20}} \right] = \left[\frac{P_{120}}{P_{1..}} - \frac{P_{120}}{P_{1.2}} \right] \log e, \\
\frac{\partial h_{121}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{121} \log \frac{P_{121} P_{1..} P_{.2.} P_{..1}}{P_{12.} P_{1.1} P_{.21}} \right] = \left[\frac{P_{121}}{P_{1..}} - \frac{P_{121}}{P_{1.2}} \right] \log e, \\
\frac{\partial h_{122}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{122} \log \frac{P_{122} P_{1..} P_{.2.} P_{..2}}{P_{12.} P_{1.2} P_{.22}} \right] \\
&= \log \frac{P_{122} P_{1..} P_{.2.} P_{..2}}{P_{12.} P_{1.2} P_{.22}} + \left[1 + \frac{P_{122}}{P_{1..}} - \frac{P_{122}}{P_{1.2}} - \frac{P_{122}}{P_{1.2}} \right] \log e,
\end{aligned} \tag{B.8}$$

and

$$\begin{aligned}
\frac{\partial h_{200}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{200} \log \frac{P_{200} P_{2..} P_{.0.} P_{..0}}{P_{20.} P_{2.0} P_{.00}} \right] = -\frac{P_{200}}{P_{2..}} \log e, \\
\frac{\partial h_{201}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{201} \log \frac{P_{201} P_{2..} P_{.0.} P_{..1}}{P_{20.} P_{2.1} P_{.01}} \right] = -\frac{P_{201}}{P_{2..}} \log e, \\
\frac{\partial h_{202}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{202} \log \frac{P_{202} P_{2..} P_{.0.} P_{..2}}{P_{20.} P_{2.2} P_{.02}} \right] = \left[-\frac{P_{202}}{P_{2..}} + \frac{P_{202}}{P_{2.2}} \right] \log e, \\
\frac{\partial h_{210}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{210} \log \frac{P_{210} P_{2..} P_{.1.} P_{..0}}{P_{21.} P_{2.0} P_{.10}} \right] = -\frac{P_{210}}{P_{2..}} \log e, \\
\frac{\partial h_{211}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{211} \log \frac{P_{211} P_{2..} P_{.1.} P_{..1}}{P_{21.} P_{2.1} P_{.11}} \right] = -\frac{P_{211}}{P_{2..}} \log e, \\
\frac{\partial h_{212}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{212} \log \frac{P_{212} P_{2..} P_{.1.} P_{..2}}{P_{21.} P_{2.2} P_{.12}} \right] = \left[-\frac{P_{212}}{P_{2..}} + \frac{P_{212}}{P_{2.2}} \right] \log e, \\
\frac{\partial h_{220}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{220} \log \frac{P_{220} P_{2..} P_{.2.} P_{..0}}{P_{22.} P_{2.0} P_{.20}} \right] = \left[-\frac{P_{220}}{P_{2..}} + \frac{P_{220}}{P_{2.2}} \right] \log e, \\
\frac{\partial h_{221}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{221} \log \frac{P_{221} P_{2..} P_{.2.} P_{..1}}{P_{22.} P_{2.1} P_{.21}} \right] = \left[-\frac{P_{221}}{P_{2..}} + \frac{P_{221}}{P_{2.2}} \right] \log e, \\
\frac{\partial h_{222}}{\partial P_{122}} &= \frac{\partial}{\partial P_{122}} \left[P_{222} \log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}} \right] \\
&= -\log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}} + \left[-1 - \frac{P_{222}}{P_{2..}} + \frac{P_{222}}{P_{2.2}} + \frac{P_{222}}{P_{2.2}} \right] \log e.
\end{aligned} \tag{B.9}$$

By relations (B.7), (B.8) and (B.9), we have

$$\begin{aligned}
\frac{\partial h}{\partial P_{122}} &= \sum_{i=0}^2 \sum_{j=0}^2 \sum_{e=0}^2 \frac{\partial h_{ije}}{\partial P_{122}} \\
&= \log \frac{P_{122} P_{1..} P_{.2.} P_{..2}}{P_{12.} P_{1.2} P_{.22}} - \log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\frac{\partial h}{\partial P_{022}} &= \log \frac{P_{022} P_{0..} P_{.2.} P_{..2}}{P_{02.} P_{0.2} P_{.22}} - \log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}}, \\
\frac{\partial h}{\partial P_{202}} &= \log \frac{P_{202} P_{2..} P_{.0.} P_{..2}}{P_{20.} P_{2.2} P_{.02}} - \log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}}, \\
\frac{\partial h}{\partial P_{212}} &= \log \frac{P_{212} P_{2..} P_{.1.} P_{..2}}{P_{21.} P_{2.2} P_{.12}} - \log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}}, \\
\frac{\partial h}{\partial P_{220}} &= \log \frac{P_{220} P_{2..} P_{.2.} P_{..0}}{P_{22.} P_{2.0} P_{.20}} - \log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}}, \\
\frac{\partial h}{\partial P_{221}} &= \log \frac{P_{221} P_{2..} P_{.2.} P_{..1}}{P_{22.} P_{2.1} P_{.21}} - \log \frac{P_{222} P_{2..} P_{.2.} P_{..2}}{P_{22.} P_{2.2} P_{.22}}.
\end{aligned}$$

Appendix C Proof of Relations (19) and (20)

Notice

$$\begin{aligned}
\frac{\partial h}{\partial P_{0\dots 0}} &= \sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 \frac{\partial h_{a_1\dots a_K}}{\partial P_{0\dots 0}} \\
&= \sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 \frac{\partial}{\partial P_{0\dots 0}} \left[P_{a_1\dots a_K} \log \frac{P_{a_1\dots a_K} \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}} \right] \\
&= \log \frac{P_{0\dots 0} \prod_{\substack{\bar{s} \subset (0, \dots, 0) \\ |(0, \dots, 0) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (0, \dots, 0) \\ |(0, \dots, 0) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}} - \log \frac{P_{2\dots 2} \prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}} \\
&\quad + \sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 P_{a_1\dots a_K} \frac{\partial}{\partial P_{0\dots 0}} \left[\log \frac{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}} \right]. \tag{C.1}
\end{aligned}$$

It can be showed that

$$\sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 P_{a_1\dots a_K} \frac{\partial}{\partial P_{0\dots 0}} \left[\log \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|=i}} P_{\bar{s}} \right] = 0, \quad i = 1, 2, \dots, K-1.$$

Therefore, we have

$$\begin{aligned}
&\sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 P_{a_1\dots a_K} \frac{\partial}{\partial P_{0\dots 0}} \left[\log \frac{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}} \right] \\
&= \sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 P_{a_1\dots a_K} \frac{\partial}{\partial P_{0\dots 0}} \left[\log \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}} \right] \\
&\quad - \sum_{a_1=0}^2 \cdots \sum_{a_K=0}^2 P_{a_1\dots a_K} \frac{\partial}{\partial P_{0\dots 0}} \left[\log \prod_{\substack{\bar{s} \subset (a_1, \dots, a_K) \\ |(a_1, \dots, a_K) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}} \right] \\
&= 0.
\end{aligned}$$

Thus, the equation (C.1) implies

$$\frac{\partial h}{\partial P_{0\dots 0}} = \log \frac{P_{0\dots 0} \prod_{\substack{\bar{s} \subset (0, \dots, 0) \\ |(0, \dots, 0) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (0, \dots, 0) \\ |(0, \dots, 0) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}} - \log \frac{P_{2\dots 2} \prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=0}} P_{\bar{s}}}{\prod_{\substack{\bar{s} \subset (2, \dots, 2) \\ |(2, \dots, 2) \setminus \bar{s}|_{\text{mod}(2)}=1}} P_{\bar{s}}}.$$

Similarly, we may show the relations (19) and (20).

Appendix D Three SNP Combinations Used to Calculate the Type I Error Rates of Test Statistic T_{IIG}

In Table D.1, we present the three SNP combinations used to calculate the empirical type I error rates of test statistic T_{IIG} . The three SNPs in each combination are strongly correlated to each other. Consider a combination of three SNPs A , B , and C . By significantly correlated to each other for the three SNPs, we mean all the four null hypotheses,

$$H_1 : P(G_A, G_B, G_C) = P(G_A, G_B)P(G_C),$$

$$H_2 : P(G_A, G_B, G_C) = P(G_A)P(G_B, G_C),$$

$$H_3 : P(G_A, G_B, G_C) = P(G_B)P(G_A, G_C),$$

$$H_4 : P(G_A, G_B, G_C) = P(G_A)P(G_B)P(G_C),$$

all unlikely to be true. We use the Pearson χ^2 test to screen the three SNP combinations. The empirical type I error rates are reported in Table 2.

Table D.1: Three SNP combinations used to calculate the empirical type I error rates of test T_{IIIG} . In each combination, the three SNPs are strongly correlated to each other. In the Table, $P(A, B, C) = P(G_A, G_B, G_C)$ means the joint genotype probability of three SNPs A, B , and C , etc.

SNPs			The Null Hypothesis H_0 , Test Values and the Related P-values of Pearson χ^2 Statistic							
A	B	C	$P(A, B, C) = P(A, B)P(C)$		$P(A, B, C) = P(A)P(B, C)$		$P(A, B, C) = P(B)P(A, C)$		$P(A, B, C) = P(A)P(B)P(C)$	
			χ^2	P-value	χ^2	P-value	χ^2	P-value	χ^2	P-value
APE1	XRCC1_399	XRCC1_194	27.72382	0.03410582	27.02576	0.04119771	40.75471	0.0006034846	32.56543	0.03763361
Xpd_751	Xpd_312	XRCC1_194	26.23323	0.05083355	27.22967	0.03899995	38.98301	0.001093773	34.63468	0.02213853
XRCC3_241	APE1	XRCC1_194	42.08023	0.0003838726	30.2027	0.01697913	75.70076	9.810734e-10	53.44553	6.98728e-05
XRCC3_241	APE1	XRCC1_399	29.35457	0.021652	24.76337	0.074107	32.45247	0.008726808	34.3638	0.02376652
XRCC3_241	XRCC1_399	XRCC1_194	25.38762	0.0632772	26.1351	0.05215666	42.28494	0.0003577711	31.26234	0.05182107

Appendix E Extra Type I Error Rates

In Table E.1, we presented the type I error rates of 1-way entropy loss test statistic T_{EL} . The test was reasonably robust.

Table E.1: Type I error rates of 1-way test statistic T_{EL} at a nominal level $\alpha = 0.01$. One di-allelic marker A is used in the simulation. Each of the entries was based on 100,000 simulations.

Frequency	Sample Sizes $M = N$								
	100	150	200	250	300	400	500	600	700
0.1	0.01632	0.01428	0.01410	0.01318	0.01271	0.01226	0.01115	0.01124	0.01116
0.2	0.01932	0.01430	0.01280	0.01243	0.01171	0.01137	0.01105	0.01063	0.01073
0.3	0.01479	0.01315	0.01232	0.01083	0.01135	0.01061	0.01007	0.01060	0.01072
0.4	0.00856	0.00797	0.00912	0.00858	0.00963	0.00920	0.00945	0.00967	0.00967
0.5	0.00366	0.00487	0.00609	0.00684	0.00713	0.00765	0.00825	0.00874	0.00853