Received: 11 September 2017 | Revised: 23 October 2017 | Accepted: 15 November 2017

DOI: 10.1002/gepi.22108

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A deeper look at two concepts of measuring gene–gene interactions: logistic regression and interaction information revisited

Jan Mielniczuk[1,2] | Paweł Teisseyre[1]

[1] Institute of Computer Science, Polish Academy of Sciences, Poland

[2] Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland

Correspondence
Jan Mielniczuk, Institute of Computer Science, Polish Academy of Sciences, 5, Jana Kazimierza, 01-248 Warsaw, Poland.
Email: miel@ipipan.waw.pl

ABSTRACT

Detection of gene–gene interactions is one of the most important challenges in genome-wide case–control studies. Besides traditional logistic regression analysis, recently the entropy-based methods attracted a significant attention. Among entropy-based methods, interaction information is one of the most promising measures having many desirable properties. Although both logistic regression and interaction information have been used in several genome-wide association studies, the relationship between them has not been thoroughly investigated theoretically. The present paper attempts to fill this gap. We show that although certain connections between the two methods exist, in general they refer two different concepts of dependence and looking for interactions in those two senses leads to different approaches to interaction detection. We introduce ordering between interaction measures and specify conditions for independent and dependent genes under which interaction information is more discriminative measure than logistic regression. Moreover, we show that for so-called perfect distributions those measures are equivalent. The numerical experiments illustrate the theoretical findings indicating that interaction information and its modified version are more universal tools for detecting various types of interaction than logistic regression and linkage disequilibrium measures.

KEYWORDS

gene–gene interactions, SNP, interaction information, mutual information, logistic regression, linkage disequilibrium

1 | INTRODUCTION

The problem of detecting gene–gene interactions in genome-wide association studies (GWAS) has attracted extensive research interest. This is motivated by the fact that most human diseases are complex, which means that they are typically caused by multiple factors, including main effects of multiple genes, well as gene–gene (G×G) interactions, gene–environment (G×E) interactions and higher order interactions (Cordell, 2002, 2009). In this paper we focus on gene–gene (G×G) interactions. The presence of gene–gene interactions has been shown in complex diseases such as breast cancer (Ritchie et al., 2001), coronary heart disease (Nelson, Kardia, Ferrell, & Sing, 2001), and Alzheimer's disease (Zubenko et al., 2001). It still remains a hotly discussed controversy to what extent the formal definition of gene–gene interaction reflects the genes' biochemical or physiological interaction (Zhao, Jin, & Xiong, 2006). Thus, there is no general agreement on how interaction should be defined. As a result, many different measures have been proposed based on different statistical and genetic assumptions (Moore & Williams, 2015). We also cite Dramiński, Kierczak, Koronacki, and Komorowski (2010) and Dramiński, Dabrowski, Diamanti, Koronacki, and Komorowski (2016) who analyze interactions using a concept of contextual dependence. Very often the competitive methods measure different effects and their detection can lead to different biological conclusions.

In this paper, we focus on two important approaches that have been widely adopted in genome-wide case–control studies: logistic regression being the state-of-the-art method and interaction information, which is an example of entropy-based methods. Logistic regression and its variants have been used by several authors (see, e.g., Hu, Wang, & Wang, 2014; Wan et al., 2010; Wu, Chen, Hastie, Sobel, & Lange, 2009). The method involves fitting two nested logistic regression models: the additive model (without interaction terms) and the non-additive model with both main effect terms and interaction terms. The value of likelihood ratio statistic is usually used as a measure of interaction strength. Recently entropy-based methods attracted significant attention, which is partly due to their nonparametric nature—they do not impose, unlike the parametric models, any particular assumptions on the data. In this group, interaction information (II) is one of the most promising measure having many desired properties. II was originally introduced by McGill (1954) who analyzed interactions in contingency tables. II is defined, using a concept of mutual information, by removing the main effect terms from the term describing the overall dependence between a pair of genes and a disease. More specifically, from the value of mutual information of the pair of genes and disease two values of mutual information between the individual genes and the disease corresponding to the main effects are subtracted (cf. equality 7). Moore et al. (2006) use II for analyzing gene–gene interactions associated with complex diseases. II is applied as a main tool to detect interactions in AMBIENCE package (Chanda et al., 2008). BOOST package uses so-called Kirkwood superposition approximation, which is closely related to II (Wan et al., 2010). Jakulin and Bratko (2003, 2004) apply II to detect interactions between variables in classification task and study how the interactions affect the performance of learning algorithms. Recently, Mielniczuk and Rdzanowski (2017) studied the properties of II and its modifications in the context of finding interactions among Single Nucleotide Polimorphism (SNPs). Some variants of II are also discussed in Fan et al. (2011). Teisseyre, Mielniczuk, and Dąbrowski (2017) developed a novel procedure for testing the positiveness of II.

The main goal of this paper is to study the relationship between the two aforementioned concepts. We show that although certain connections between II and logistic regression exist, in general they refer two different concepts of dependence and looking for interactions in those two senses leads to different approaches to interaction detection. For independent genes, we show that II is more discriminative than logistic regression in the sense that $II = 0$ implies always $\gamma_{ij} = 0$, where γ_{ij} are parameters corresponding to interaction terms in logistic regression. This is also shown for a certain classes of dependent predictors, see Theorem 5. On the other hand, we show that it is relatively easy to construct probability distributions described by additive logistic models (without interaction terms) for which the interaction defined by II does not vanish. Simulation experiments confirm that there are interactions found by II, which remain undetected when using logistic regression. In our theoretical results we use a concept of so-called perfect distributions (Darroch, 1974), which are characterized by three conditions (see (17)–(19)). We prove that for perfect distributions of the triple (gene, gene, disease), there is equivalence between $II = 0$ and $\gamma_{ij} = 0$. Moreover, we characterize the situations when $II > 0$, that is, the interaction defined by II is positive. For independent genes, it turns out that if at least one of the conditions characterizing perfect distributions is not satisfied then $II > 0$. The analogous conclusion is proved for certain class of dependent genes. In simulation experiments, we compare the performance of interaction information and its variants (Kirkwood superposition approximation and modified interaction information) with the logistic regression and, as a benchmark, also with a measure based on linkage disequilibrium (Yang, He, & Ott, 2009). The latter one is somewhat similar to II, in a sense that they both measure the difference of interloci associations between cases and controls (see Section 2 for details). We show that for the studied models, the modified II is on the whole much more powerful measure of interaction detection than the logistic regression.

The paper is organized as follows. In Section 2, we describe the two approaches for detection of gene–gene interactions: logistic regression and interaction information. We also discuss some relevant properties of interaction information. In Section 3, we state our main results on the relationship between the two approaches. Section 4 contains the results of simulation experiments. Section 5 concludes the results. Some technical issues are discussed in the Appendix.

2 | METHODS FOR DETECTING “GENE–GENE” INTERACTIONS

2.1 | Definitions and notation

Let X_1 and X_2 denote two SNPs and Y denote the class label (1 for cases and 0 for controls). SNPs are genetic markers in genome-wide case–control studies. For each SNP, there are three genotypes: the homozygous reference genotype (AA or BB), the heterozygous genotype (Aa or Bb, respectively), and the homozygous variant genotype (aa or bb). Here A and a correspond to the alleles of the first SNP (X_1), whereas B and b to the alleles of the second SNP (X_2). We denote by $p(x_i, x_j, y_k) = P(X_1 = x_i, X_2 = x_j, Y = y_k)$ the probability mass function, corresponding to the joint distribution $P_{X1,X_2,Y}$ of the triple (X_1, X_2, Y). Note that P corresponds to $3 \times 3 \times 2$ probability table. We refer to Agresti (2013) for a model-based approach for analysis of probability tables. The analogous notation is used for univariate, bivariate and conditional distributions. With a slight abuse of notion $p(x_i)$ and $p(x_j)$ will denote univariate mass functions of X_1 and X_2, respectively.

As our objective is to study how various measure of interactions are intertwined, we introduce the following definition.

Definition 1. Let I_1 and I_2 are two different measures of interaction. We say that I_1 is more discriminative (liberal) measure of interaction than I_2, denoted by $I_2 \prec I_1$, when

$$I_1 = 0 \quad \text{implies} \quad I_2 = 0$$

or equivalently

$$I_2 \neq 0 \quad \text{implies} \quad I_1 \neq 0.$$

In view of the definition, if one considers two viable interaction measures I_1 and I_2 are I_1 is more discriminative than I_2 than I_1 should be preferred. In the following, we will show among others that for independent SNPs interaction information (II) is more discriminative than the logistic regression and linkage disequilibrium (LD) measures.

2.2 | Logistic regression

One of the most popular way of defining interactions is via logistic regression models. In the additive logistic regression (with only main effect terms), logarithmic odds have the following additive form:

$$\begin{aligned}\log\left[\frac{P(Y=1|X_1,X_2)}{P(Y=0|X_1,X_2)}\right] \\ = \mu + \alpha_1 I(X_1 = Aa) + \alpha_2 I(X_1 = aa) + \beta_1 I(X_2 = Bb) \\ + \beta_2 I(X_2 = bb), \qquad (1)\end{aligned}$$

where $I(C)$ is an indicator function of set C. The general logistic regression model with both main effect terms and interaction terms has the form

$$\begin{aligned}\log\left[\frac{P(Y=1|X_1,X_2)}{P(Y=0|X_1,X_2)}\right] = \\ \mu + \alpha_1 I(X_1 = Aa) + \alpha_2 I(X_1 = aa) \\ + \beta_1 I(X_2 = Bb) + \beta_2 I(X_1 = bb) \\ + \gamma_{11} I(X_1 = Aa, X_2 = Bb) + \gamma_{12} I(X_1 = Aa, X_2 = bb) \\ + \gamma_{21} I(X_1 = aa, X_2 = Bb) + \gamma_{22} I(X_1 = aa, X_2 = bb). \qquad (2)\end{aligned}$$

In the logistic regression model, X_1 and X_2 interact when γ_{ij} is nonzero for some i, j. Thus, interactions here pertain to any nonzero coefficients γ_{ij}, and absence of interactions means that all of them are zero. Note that the number of independent parameters is 5 for model (1) and 9 for model (2). As the number of parameters in (2) equals the number of values of the odds $P(Y = 1|X_1, X_2)/P(Y = 0|X_1, X_2)$, it follows that any conditional distribution is described by this equation and that is why (2) is sometimes called saturated model. Observe that for $X_1 = AA$ and $X_2 = BB$, logarithmic odds in (1) and (2) are equal to μ and thus genotypes AA and BB correspond to the reference levels. However, the choice of the reference level may be arbitrary and it does not influence prediction in the model. Typically, likelihood ratio statistic is used as a measure of interaction strength

$$LRT(X_1; X_2; Y) := 2(L_{M1} - L_{M0}), \qquad (3)$$

where L_{M0}, L_{M1} are log-likelihood functions corresponding to the fitted models (1) and (2), respectively. Alternatively, other measures can be used, for example, interaction logistic measure IL :

$$IL(X_1; X_2; Y) := \gamma_{11}^2 + \gamma_{12}^2 + \gamma_{21}^2 + \gamma_{22}^2, \qquad (4)$$

considered by Hu et al. (2014), which equals 0 only when $\gamma_{ij} \equiv 0$. Thus, $IL > 0$ is equivalent to existence of logistic interactions.

2.3 | Interaction information

Lack of interactions as defined in additive logistic model is an appealing concept that is widely used. However, this approach has its limitations as it is based on a specific model. Our aim is now to define a model-free interaction measure that does not suffer from this drawback. To this end, we recall first some concepts developed in information theory in order to measure the strength of dependence between two qualitative variables (we refer to Cover and Thomas, 2006, for extensive treatment of the subject). Definition of interaction information, we are about to define, is based on mutual information and thus we first define mutual information and discuss its properties. Mutual information quantifies the amount of information obtained about one random variable due to the knowledge of the other random variable. It is defined as

$$I(X_1, X_2) := \sum_{i,j} p(x_i, x_j) \log\left(\frac{p(x_i, x_j)}{p(x_i)p(x_j)}\right) \qquad (5)$$

and thus can be regarded as a measure of association for a pair of discrete variables. It determines how similar the joint distribution is to the product of factored marginal distributions. Mutual information can be also defined for continuous variables and in this case it is more general measure of dependence than traditional correlation coefficient as it allows to detect nonlinear dependencies. In order to grasp the

TABLE 1 Probability mass function $p(x_i, x_j)$ corresponding to a pair (X_1, X_2), $a \in [0, 2/9]$ is a parameter

	bb	*Bb*	*BB*	
aa	$1/9 + a$	$1/9 - a/2$	$1/9 - a/2$	1/3
Aa	$1/9 - a/2$	$1/9 + a$	$1/9 - a/2$	1/3
AA	$1/9 - a/2$	$1/9 - a/2$	$1/9 + a$	1/3
	1/3	1/3	1/3	1

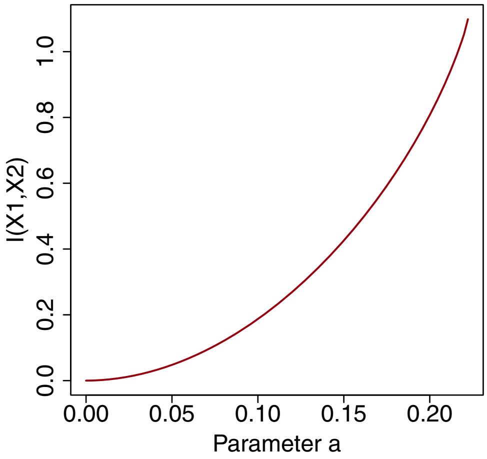


FIGURE 1 Theoretical value of $I(X_1, X_2)$ with respect to the value of parameter a

idea of mutual information, we discuss the following example. Consider a probability mass function $p(x_i, x_j)$ given in Table 1 describing the distribution of two SNPs. Parameter $a \in [0, 2/9]$ controls strength of dependence between two SNPs. The larger the value of a, the stronger the dependence between SNPs. For $a = 0$ variables X_1 and X_2 are independent, whereas for $a = 2/9$ there is a complete dependence between them, meaning that the value of the one variable determines the value of the other. Figure 1 shows how the mutual information depends on the value of a.

Mutual information can be expressed as a so-called Kullback–Leibler (KL) divergence of the joint distribution from the product of marginal distributions (see the definition in the Appendix). The mutual information equals 0 if and only if X_1 and X_2 are independent. It can be also interpreted as a decrease in the amount of uncertainty of one variable when the value of the other variable is known. Due to this, it is often called information gain. Mutual information is always nonnegative and its value reflects a strength but not a direction of dependence, that is, it does not distinguish between positive and negative association.

Analogously to (5), we define the mutual information between a pair (X_1, X_2) and Y as

$$I[(X_1, X_2); Y] := \sum_{i,j,k} p(x_i, x_j, y_k) \log \left(\frac{p(x_i, x_j, y_k)}{p(x_i, x_j)p(y_k)} \right). \tag{6}$$

Some authors use (6) for detecting interactions in GWAS (see, e.g., Leem, Jeong, Lee, Wee, & Sohn, 2014). However, one should be aware that (6) quantifies the overall dependence between Y and the pair (X_1, X_2). Thus, it contains the information related to interaction between two SNPs as well as to marginal contributions of two SNPs in predicting Y(see the definition (7)). Measure (6) is useful if we look for the pairs of genes that affect the disease, but it can be misleading if we are mainly interested in detecting interactions. For example, (6) may be large even though there is no interaction between predictors in explaining Y. This will be the case for a pair of SNPs that influence Y only individually.

Interaction information (II) (McGill, 1954; Fano, 1961) is defined as

$$II(X_1; X_2; Y) := I[(X_1, X_2); Y] - I(X_1; Y) - I(X_2, Y). \tag{7}$$

It follows from the above definition that II can be interpreted as a part of the mutual information of (X_1, X_2) and Y, which is due solely to interaction between X_1 and X_2 in predicting Y, that is, the part of $I[(X_1, X_2); Y]$ that remains after subtraction of individual information between Y and X_1 and Y and X_2. In other words, II is obtained by removing the main effects from the term describing the overall dependence between Y and the pair (X_1, X_2). Note that in contrast to IL measure defined in (4), interaction information does not depend on any specific model.

Next we state and discuss some important properties of interaction information (II), which will be used to prove our main results in the next section. First, it turns out that II is closely related to so-called Kirkwood superposition approximation (Matsuda, 2000; Wan et al., 2010). In order to see this, define the unnormalized Kirkwood superposition approximation $\tilde{P}_K$, which corresponds to a mass function

$$\tilde{p}_K(x_i, x_j, y_k) = \frac{p(x_i, x_j)p(x_i, y_k)p(x_j, y_k)}{p(x_i)p(x_j)p(y_k)}. \tag{8}$$

We note that the following properties hold:

(i) The interaction information can be written as KL divergence between the joint distribution of X_1, X_2, Y, and unnormalized Kirkwood superposition approximation:

$$II(X_1; X_2; Y) = KL(P_{X_1,X_2,Y} || \tilde{P}_K)$$
$$= \sum_{i,j,k} p(x_i, x_j, y_k) \log \left(\frac{p(x_i, x_j, y_k)}{\tilde{p}_K(x_i, x_j, y_k)} \right), \tag{9}$$

where KL is KL divergence, defined in the Appendix.

(ii) The interaction information equals the difference between the conditional mutual information $I(X_1; X_2|Y)$

and the unconditional mutual information $I(X_1;X_2)$

$$II(X_1;X_2;Y) = I(X_1;X_2|Y) - I(X_1;X_2), \qquad (10)$$

where the conditional mutual information $I(X_1;X_2|Y)$ is defined in the Appendix as the mutual information of X_1 and X_2 given $Y = y$ averaged over the values of $Y = y$.

The proofs of (9) and (10) are straightforward and can be found, for example, in Mielniczuk and Rdzanowski (2017). In connection with (9) note that $\tilde{P}_K$ is not necessarily proper probability distribution, as masses $\tilde{p}_K(x_i,x_j,y_k)$ do not necessarily sum up to 1. Define a normalizing constant η:

$$\eta = \sum_{i,j,k} \tilde{p}_K(x_i,x_j,y_k) \qquad (11)$$

and let Kirkwood superposition approximation P_K correspond to probability mass function $p_K(x_i,x_j,y_k) = \tilde{p}_K(x_i,x_j,y_k)/\eta$.

Note that η is a numerical index related to dependence structure of X_1, X_2, Y. In particular, $\eta \neq 1$ implies that X_1 and X_2 are dependent. Indeed, if X_1 and X_2 were independent then

$$\eta = \sum_{i,j,k} \frac{p(x_i,x_j)p(x_i,y_k)p(x_j,y_k)}{p(x_i)p(x_j)p(y_k)}$$
$$= \sum_{i,j,k} \frac{p(x_i,y_k)p(x_j,y_k)}{p(y_k)} = \sum_{j,k} p(x_j,y_k) = 1.$$

Moreover, note that in view of (9) we have

$$II(X_1;X_2;Y) = KL(P_{X_1,X_2,Y}||P_K) - \log(\eta).$$

As $KL(P_{X_1,X_2,Y}||P_K)$ is always nonnegative we have that $\eta \leq 1$ implies $II \geq 0$ and analogously $\eta < 1$ implies that II is strictly positive.

In connection with (ii) above, note that equality (10) indicates that II measures the influence of a variable Y on the amount of information shared between X_1 and X_2. In other words, we verify how much Y influences the dependence between X_1 and X_2. Property (10) shows that II is loosely related to the popular group of methods based on measuring the difference of interloci associations between cases and controls (Cordell, 2009; Hu et al., 2014; Kang, Yue, Cui, & Zhang, 2008). This important group is represented by the methods that compare the linkage disequilibrium (LD) in cases and controls (Yang et al., 2009), in particular LD measure defined in (23). Note that in (10) instead of comparing the strength of dependence between cases and controls one compares an averaged conditional strength over strata with an unconditional one.

Observe that it follows from (10) that in contrast to the mutual information, II can be either positive or negative. Positive value of II indicates that Y enhances the association between X_1 and X_2. In other words, the conditional dependence is stronger than the unconditional one. A negative value of II indicates that Y weakens or inhibits the dependence between X_1 and X_2. For more detailed discussion and examples of positive and negative II, we refer to Teisseyre et al. (2017). In genome-wide case–control studies, we are mainly interested in $II > 0$, as in view of (7), the positive interaction information indicates that the information contained in the pair (X_1, X_2) is larger than the sum of information due to the individual variables X_1 and X_2.

On the negative side, large sample properties of empirical counterparts of II are not known for a general distribution $P_{X_1,X_2,Y}$ from which the data are sampled (see however Han, 1980, for the case when all components are mutually independent). This creates problems when one would like to use II for testing, which can be partially overcome by using Monte-Carlo methods (see Teisseyre et al., 2017). In the numerical experiments in Section 4, we focus on ranking procedures that do not require testing.

Finally, we note that modifications of II are used for interaction detection. For example, the measure defined as KL divergence of $P_{X_1,X_2,Y}$ its from Kirkwood superposition approximation equals

$$KA(X_1;X_2;Y) := KL(P_{X_1,X_2,Y}||P_K)$$
$$= II(X_1;X_2;Y) + \log(\eta) \qquad (12)$$

can be used as a measure of interaction strength (Mielniczuk & Rdzanowski, 2017; Wan et al., 2010). Mielniczuk and Rdzanowski (2017) also proposed modified interaction information

$$IIM(X_1;X_2;Y) := \max[II(X_1;X_2;Y), KA(X_1;X_2;Y)] \qquad (13)$$

and showed experimentally that very often it outperforms standard II when testing for interaction effects. Both measures above together with II are used in our numerical experiments.

Fan et al. (2011) treated the difference between the mutual information in the affected population and that in the general population as the information gain of two genes in the presence of a disease. In consequence, they proposed to use

$$IG(X_1;X_2;Y) := I(X_1;X_2|Y = 1) - I(X_1;X_2)$$

as the measure of interaction effect between two genes. We also refer to Lee, Sjölander, and Pawitan (2016) who among others analyze IG (but not II) and study its relationship with logistic regression. Although the definition of IG and equality (10) for II seem similar, there are substantial differences between these two concepts. The advantage of II over IG is its intuitive interpretation as a difference between overall dependence term and main effect terms, see (7). Indeed, we note that the decomposition (7) is not true for IG, that is, when the left-hand side of (7) is replaced by IG and the summands on the right-hand side are replaced by their conditional analogues given $Y = 1$. Moreover, in their Example 2 Lee et al. (2016) give an example of a logistic model with binary trait Y depending on predictor X_1 only, such that for a certain X_2 dependent on X_1 but independent of Y, interaction between X_1 and X_2 in predicting Y measured by IG is positive. This is obviously counterintuitive as X_2 is not a valid predictor in the considered model, that is, $p(y|x_1, x_2) = p(y|x_1)$ and constitutes a major drawback of this measure. In contrast, in the discussed situation interaction information vanishes. Indeed, it follows from the last equality that $H(Y|X_1, X_2) = H(Y|X_1)$, where $X(Y|X)$ is the conditional entropy of Y given X defined in the Appendix and thus (cf. equality (24) in the Appendix):

$$I[(X_1, X_2); Y] = H(Y) - H(Y|X_1, X_2)$$

$$= H(Y) - H(Y|X_1) = I(X_1; Y)$$

and as $I(X_2; Y) = 0$ it follows from (7) that $II(X_1; X_2; Y) = 0$. Moreover, Lee et al. (2016) construct an example when $IG = 0$ but there is nonzero interaction in the logistic model. The example does not hold for II as it assumes that X_1 and X_2 are independent but conditionally independent given $Y = 1$ only (but not given $Y = 0$). Thus $II > 0$, whereas $IG = 0$. In the next section, we discuss the property that for independent predictors X_1 and X_1, II equals 0 is equivalent to conditional independence of predictors given Y and this entails that additive logistic model holds.

3 | MAIN RESULTS

The main goal of this paper is to study the relationship between logistic interaction and interaction information. Recall that interactions in logistic regression are described by parameters $\{\gamma_{ij}\}$ (see (2) in Section 2.2). We show that although certain connections between II and $\{\gamma_{ij}\}$ exist, in general they refer two intrinsically different concepts of dependence and looking for interactions in those two senses leads to different approaches to interaction detection. This has important implications for interpretation of biological phenomena as the interactions detected by the one method may remain undetected by the other one and vice versa.

More specifically, we prove in Theorem 1 that when X_1 and X_2 are independent, logistic interactions are zero provided X_1 and X_2 do not interact in information theoretic sense. Moreover, in Theorem 4 we show that II and IL are strictly equivalent for a certain family of $3 \times 3 \times 2$ probability distributions called perfect distributions defined by equations (17)–(19). However, outside this family, when constant η defined in (11) is not larger than 1, II is *not* in general equivalent to IL. This is shown in Theorem 2 for independent SNPs and in Theorem 3 in general case. The question what happens for nonperfect distributions when $\eta > 1$ is the subject of ongoing research.

First, we consider a situation of independent X_1 and X_2. The following Theorem shows that in this case $II = 0$ implies logistic model with no interaction effects and thus II is more discriminative than IL, $II \succ IL$.

Theorem 1. *Assume that X_1 and X_2 are independent. Then $II(X_1; X_2; Y) = 0$ implies $\gamma_{ij} = 0$ for all i, j.*

Proof. When X_1 and X_2 are independent it follows from property (10) that $II(X_1; X_2; Y) = 0$ is equivalent to conditional independence of X_1 and X_2 given Y. It follows from Lemma 1 (see Appendix) that the conditional independence implies $\gamma_{ij} = 0$ for all i, j.

□

Next we discuss why the converse statement is not true. More specifically, we give an example of a class of probability distributions $p(x_i, x_j, y_k)$, which are represented by additive logistic regression (without interaction terms) and for which $II > 0$. Consider first the case when X_1 and X_2 are independent. Then $I(X_1; X_2) = 0$ and it follows from (10) that $II(X_1; X_2; Y) \geq 0$ and, moreover, $II(X_1; X_2; Y) > 0$ is equivalent to conditional dependence of X_1 and X_2. Thus in this case, II is a measure of strength of conditional dependence of X_1 and X_2 given Y. This corresponds to frequent approaches to interaction detection, which consist in checking how much conditional distributions of (X_1, X_2) on strata $Y = 0$ and $Y = 1$ differ. Now we check whether conditional dependence is reflected by logistic regression.

Note that when X_1 and X_2 are conditionally independent given Y we have that

$$p(x_i, x_j, y_k) = p(x_i, x_j|y_k)p(y_k) = p(x_i|y_k)p(x_j|y_k)p(y_k)$$

and thus

$$\log p(x_i, x_j, y_k) = \lambda_k + \lambda_{ik}^{X_1,Y} + \lambda_{jk}^{X_2,Y}, \tag{14}$$

where $\lambda_k = \log p(y_k)$, $\lambda_{ik}^{X_1,Y} = \log p(x_i|y_k)$ and $\lambda_{jk}^{X_2,Y} = \log p(x_j|y_k)$. We consider now a special type of conditional dependence of X_1 and X_2 given Y by introducing the term $\lambda_{ij}^{X_1,X_2}$ in equation (14). Namely, we assume that

$$\log p(x_i, x_j, y_k) = \lambda_k + \lambda_{ij}^{X_1,X_2} + \lambda_{ik}^{X_1,Y} + \lambda_{jk}^{X_2,Y}. \tag{15}$$

Then logarithmic odds of $Y = 1$ given X_1 and X_2 equal

$$\log\left[\frac{P(Y=1|x_i,x_j)}{P(Y=0|x_i,x_j)}\right] = \log\left[\frac{p(x_i,x_j,1)}{p(x_i,x_j,0)}\right]$$

$$= (\lambda_1 - \lambda_0) + (\lambda_{i1}^{X_1,Y} - \lambda_{i0}^{X_1,Y}) + (\lambda_{j1}^{X_2,Y} - \lambda_{j0}^{X_2,Y})$$

$$= \mu + \alpha_i + \beta_j. \tag{16}$$

Note that the term $\lambda_{ij}^{X_1,X_2}$ cancels out and we obtain additive logistic regression representation. Thus in the case of (15), the probability table $p(x_i, x_j, y_k)$ with conditionally *dependent* X_1 and X_2 yields additive logistic model and the strength of the conditional dependence cannot be learnt by detection of logistic interactions in fitted logistic regression. In general, conditional dependence is also reflected by appearance of second-order interaction term $\lambda_{ijk}^{X_1,X_2,Y}$ in (15) and then, obviously, interaction terms

$$\lambda_{ij1}^{X_1,X_2,Y} - \lambda_{ij0}^{X_1,X_2,Y}$$

will be present in (16). However, this example indicates the extent to which meaning of II and $\{\gamma_{ij}\}$ may differ. Namely, it shows that for model (15) conditional dependence of X_1 and X_2 will be detected by considering II whereas it will not be detectable by logistic regression.

We define now a certain regular family of $3 \times 3 \times 2$ distributions called perfect distributions (or perfect tables, see Darroch, 1974). Namely members of the family satisfy the following conditions:

$$\sum_i \frac{p(x_i, x_j)p(x_i, y_k)}{p(x_i)} = p(x_j)p(y_k) \quad (\text{all } j, k), \tag{17}$$

$$\sum_j \frac{p(x_j, y_k)p(x_i, x_j)}{p(x_j)} = p(x_i)p(y_k) \quad (\text{all } i, k), \tag{18}$$

$$\sum_k \frac{p(x_i, y_k)p(x_j, y_k)}{p(y_k)} = p(x_i)p(x_j) \quad (\text{all } i, j). \tag{19}$$

We note the following.

(i) If X_1 and X_2 are independent, conditions and (17) and (18) are automatically satisfied. If additionally X_1 or X_2 are independent of Y then (19) holds, that is, the corresponding distribution is perfect.

In order to see this, consider, for example, (18). For any i, k we have

$$\sum_j \frac{p(x_j, y_k)p(x_i, x_j)}{p(x_j)} = \sum_j p(x_i)p(x_j, y_k) = p(x_i)p(y_k).$$

Additionally, if, for example, X_1 is independent of Y we have

$$\sum_k \frac{p(x_i, y_k)p(x_j, y_k)}{p(y_k)} = \sum_k p(x_i)p(x_j, y_k) = p(x_i)p(x_j)\,(\text{all } i, j)\,, \tag{20}$$

that is, equality (19) holds.

(ii) If X_1 and X_2 are independent and additionally they are conditionally independent given Y then (19) is satisfied and the corresponding distribution is perfect.

This follows by direct calculation (see the proof of Theorem 2). Note that conditional independence of X_1 and X_2 given Y does not imply independence of X_1 and X_2 as example: $X_1 = W + Y, X_2 = Z + Y$, W, Z, Y-independent, shows. It is also easily seen that

(iii) if one of the conditions (17)–(19) holds then $\eta = 1$;

(iv) if (17)–(19) hold, then P_K has the same marginal bivariate distributions as $P_{X_1,X_2,Y}$.

Indeed, in view of (iii) we have $\eta = 1$ and using (17) we have

$$p_K(x_i, x_j) = \sum_k \frac{p(x_i, x_j)p(x_i, y_k)p(x_j, y_k)}{p(x_i)p(x_j)p(y_k)}$$
$$= \frac{p(x_i, x_j)p(x_i)p(x_j)}{p(x_i)p(x_j)} = p(x_i, x_j).$$

Equality of two other bivariate marginals is checked in the same way. We will show in Theorem 4 that for such distributions equivalence between $II = 0$ and $\gamma_{ij} \equiv 0$ holds. On the other hand, outside this family, when $\eta \leq 1$ the two conditions are not equivalent in the sense that for any nonperfect distribution that conforms to the additive logistic model, we have $\gamma_{ij} \equiv 0$ and $II > 0$ (cf. Theorems 2 and 3). First, we discuss the following result that shows the relevance of (19) in studying the positivity of II when X_1 and X_2 are independent. Moreover, it indicates that for independent genes, II and IL are not equivalent when the underlying distribution is not perfect.

Theorem 2. *Assume that X_1 and X_2 are independent. If condition* (19) *is not satisfied then $II > 0$.*

Note that the theorem holds regardless whether the logistic model has nonzero logistic interactions or not. Thus, if $\gamma_{ij} \equiv 0$ and (19) does not hold, we still have $II > 0$. In Section 4, we will consider some specific cases of such situation.

Proof. In view of the discussion above it is sufficient to show that X_1 and X_2 are conditionally dependent given Y if (19) is not satisfied. Note that if the opposite were true we would have for any i, j, k

$$p(x_i, x_j, y_k) = \frac{p(x_i, y_k)p(x_j, y_k)}{p(y_k)}.$$

Summing the above equality over k we obtain

$$\sum_k \frac{p(x_i, y_k)p(x_j, y_k)}{p(y_k)}$$

$$= \sum_k p(x_i, x_j, y_k) = p(x_i, x_j) = p(x_i)p(x_j),$$

where the last equality follows from independence of predictors. But this means that (19) is satisfied.

□

We note that Theorem 2 can be generalized to the case of dependent predictors. Recall that η is a normalizing constant from Kirkwood superposition approximation, defined in (8) and also that $\eta \neq 1$ implies that X_1 and X_2 are dependent. In the following theorem, we allow for dependent predictors assuming more generally that $\eta \leq 1$.

Theorem 3. *If* $\eta \leq 1$ *and at least one of conditions* (17)–(19) *is not satisfied then* $II > 0$.

Proof. Note that when $\eta \leq 1$ then $II \geq 0$. When $\eta < 1$ then $II = KL(P_{X_1,X_2,Y}||P_K) - \log \eta > 0$ as KL divergence is nonnegative. When $\eta = 1$ then $II = KL(P_{X_1,X_2,Y}||P_K)$. Assume by contradiction that $II = 0$. Then because P_K is probability distribution we have that $P = P_K$ and thus

$$p(x_i, x_j, y_k) = \frac{p(x_i, x_j)p(x_i, y_k)p(x_j, y_k)}{p(x_i)p(x_j)p(y_k)}. \tag{21}$$

Summing over k of both sides of (21) yields

$$p(x_i, x_j) = \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \sum_k \frac{p(x_i, y_k)p(x_j, y_k)}{p(y_k)},$$

which after division by $p(x_i, x_j)$ implies that (19) holds. Analogously, summing over i or j in (21) shows that (17) or (18) holds.

□

Now we prove a positive result stating that for perfect distributions satisfying (17)–(19), conditions $II = 0$ and $\gamma_{ij} \equiv 0$ are equally discriminative, that is, $II \prec IL$ and $IL \prec II$. Note that if one of conditions (17)–(19) hold, then in view of (iii) above $\eta = 1$. This will be used in the following proof.

Theorem 4. *Assume that* (17)–(19) *hold. Then* $II = 0$ *if and only if* $\gamma_{ij} = 0$ *for any* i, j.

Proof. Assume that $\gamma_{ij} = 0$ for any i, j, that is, additive logistic model is satisfied. Then $p(x_i, x_j, y_k)$ can be represented in the following multiplicative form:

$$p(x_i, x_j, y_k) = \phi_{ij}\psi_{jk}\theta_{ik} \tag{22}$$

for some ϕ_{ij}, ψ_{jk} and θ_{ik} (cf., e.g., Mielniczuk and Rdzanowski, 2007, Proposition 7(i)). Note that Kirkwood approximation P_K is also representable in this form. Moreover, as $\eta = 1$, it follows from (iv) above that if (17)–(19) are satisfied then P_K has the same bivariate marginals as $P_{X_1,X_2,Y}$. We will use now an important result stating that for any distribution $p(x_i, x_j, y_k)$ there exists exactly one probability distribution $\bar{p}(x_i, x_j, y_k)$ having the same bivariate marginals as $p(x_i, x_j, y_k)$ and representable in the multiplicative form (cf. Birch, 1963). As both p and p_K have these properties, they must coincide,

$$p(x_i, x_j, y_k) = \bar{p}(x_i, x_j, y_k) = p_K(x_i, x_j, y_k).$$

Then it follows that $II = KL(P_{X_1,X_2,Y}||\tilde{P}_K) = KL(P_{X_1,X_2,Y}||P_K) = 0$, taking into account that $P_K = \tilde{P}_K$ as $\eta = 1$. Assume now that for some i, j $\gamma_{ij} \neq 0$. As above we show that P_K is probability distribution having the same bivariate marginals as $P_{X_1,X_2,Y}$. Because P_K conforms with logistic additive model and $P_{X_1,X_2,Y}$ does not, they are different probability distributions and thus $II = D(P_{X_1,X_2,Y}||P_K) > 0$.

□

We observe that one part of the last theorem, as well as Theorem 2, may be strengthened in the analogous way Theorem 3 strengthens Theorem 2. Namely, it holds the following theorem.

Theorem 5. *If* $\eta \leq 1$ *and* $II(X_1; X_2; Y) = 0$ *then* $\gamma_{ij} \equiv 0$, *that is,* $IL \prec II$.

Proof. This follows by reasoning similar to the proof of Theorem 3 as $II = 0$ implies $\eta = 1$ and thus $II = KL(P_{X_1,X_2,Y}||P_K) = 0$. Consequently, $P_{X_1,X_2,Y} = P_K$. This means, however, that (22) holds and interactions of the logistic model are all zero.

□

The converse result is not true as introductory example in this section indicates.

4 | SIMULATION STUDY

The aim of the simulation study is to compare the performance of the discussed measures and empirically confirm the main conclusion of theoretical part of this work, namely that logistic regression and II may detect different types of interactions. We compare the following measures: interaction information (II) defined in (7), the measure based on Kirkwood approximation (KA) defined in (12), modified interaction information (IIM) defined in (13), and likelihood ratio (LRT) statistic defined in (3). We also investigated the performance of IL (4), but as it turns out to be inferior to LRT we do not discuss its behavior here. As a benchmark we also use an index measuring strength of linkage disequilibrium, proposed by Yang et al. (2009). It is defined as

$$LD(X_1; X_2; Y) := \sum_{i,j} \frac{(\delta_{ij}^{(1)} - \delta_{ij}^{(0)})^2}{p(x_i, x_j)}, \tag{23}$$

TABLE 2 The odds of disease for two-locus models. Parameter γ corresponds to the baseline odds and parameter θ is a genotypic effect

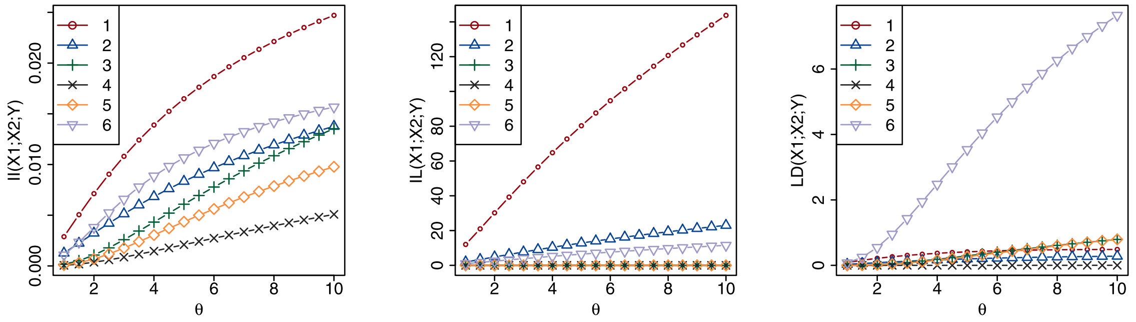
	Model 1				**Model 2**		
	bb	***Bb***	***BB***		***bb***	***Bb***	***BB***
aa	$\gamma(1+\theta)^4$	$\gamma(1+\theta)^2$	γ	aa	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
Aa	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$	γ	Aa	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
AA	γ	γ	γ	AA	γ	γ	γ
	Model 3				**Model 4**		
	bb	***Bb***	***BB***		***bb***	***Bb***	***BB***
aa	$\gamma(1+\theta)^4$	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$	aa	$\gamma(1+\theta)^2$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$
Aa	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$	Aa	$\gamma(1+\theta)^2$	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$
AA	$\gamma(1+\theta)^2$	$\gamma(1+\theta)$	γ	AA	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ
	Model 5				**Model 6**		
	bb	***Bb***	***BB***		***bb***	***Bb***	***BB***
aa	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$	aa	$\gamma(1+\theta)^4$	$\gamma(1+\theta)^3$	$\gamma(1+\theta)^2$
Aa	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ	Aa	$\gamma(1+\theta)^3$	γ	γ
AA	$\gamma(1+\theta)$	$\gamma(1+\theta)$	γ	AA	$\gamma 1+\theta)^2$	γ	γ

where $\delta_{ij}^{(1)} := p(x_i, x_j|1) - p(x_i|1)p(x_j|1)$, $\delta_{ij}^{(0)} := p(x_i, x_j|0) - p(x_i|0)p(x_j|0)$. LD is a representative of a large family of measures that are based on comparing the interloci associations between cases and controls. More examples of LD-based measures can be found in Hu et al. (2014) and Zhao et al. (2006). Observe that for independent X_1 and X_2, $II(X_1;X_2;Y) = 0$ implies $LD(X_1;X_2;Y) = 0$. Indeed, it then follows from (10) that $II(X_1;X_2;Y) = 0$ is equivalent to conditional independence of X_1 and X_2 given Y. This in turn implies that $\delta_{ij}^{(1)} = \delta_{ij}^{(0)} = 0$ and thus $LD(X_1;X_2;Y) = 0$. Whence for independent predictors we have that $LD \prec II$. In simulation experiments, we consider empirical counterparts of the above quantities.

TABLE 3 The odds of disease for two-locus models. Parameter γ corresponds to the baseline odds and parameter θ is a genotypic effect. Symbols $\#a$ and $\#b$ denote the number of disease alleles at the first and the second disease locus, respectively

Simulation model	Odds of disease
Model 1	$\gamma(1+\theta)^{\#a\times\#b}$
Model 2	$\gamma(1+\theta)^{I\{\#a>0\}\times I\{\#b>0\}}$
Model 3	$\gamma(1+\theta)^{\#a+\#b}$
Model 4	$\gamma(1+\theta)^{I(\#a>0)+I(\#b>0)}$
Model 5	$\gamma(1+\theta)^{2I(X_1=aa)+I(X_2=bb)+I(X_2=Bb)}$
Model 6	$\gamma(1+\theta)^{I(\max(\#a,\#b)\geq 2)\times(\#a+\#b)}$

We consider six two-locus disease models. The odds of disease corresponding to the models are given in Table 2. The odds depend on two parameters: γ, which is the baseline odds, and θ, which is the genotypic effect. In our experiments, we determine the value of γ that matches the prevalence $\pi_1 = P(Y = 1)$. Models 1 and 2 were considered previously by Kang et al. (2008) and Wang, Liu, Feng, and Wong (2011). These models have different biological interpretations. For example, Model 1 is referred to as a two-locus multiplicative. In this model, the odds of disease have a baseline value γ and increase multiplicatively once there is at least one disease allele (denoted by small letter: a or b) at each disease locus, that is, the odds of disease are equal to $\gamma(1+\theta)^{\#a\times\#b}$, where $\#a$ and $\#b$ are the number of disease alleles at the first and the second disease locus, respectively. Model 2 refers to a two-locus threshold model. Here, the odds of disease also have a baseline value γ unless the disease allele is present at each locus. Once this threshold condition is satisfied, the odds of disease increase to $\gamma(1+\theta)$, that is, the odds are $\gamma(1+\theta)^{I(\#a>0)\times I(\#b>0)}$, where $I(A)$ is an indicator of event A. In Model 3 (Wang et al., 2011) the odds increase additively once there is at least one disease allele, that is, the odds of disease are equal to $\gamma(1+\theta)^{\#a+\#b}$. Analogously, Model 4 is an additive analogue of Model 2, that is, the odds are $\gamma(1+\theta)^{I(\#a>0)+I(\#b>0)}$ and the odds in Model 6 are $\gamma(1+\theta)^{I(\max(\#a,\#b)\geq 2)\times(\#a+\#b)}$. Model 5 is an asymmetric model in the sense that presence of disease alleles influences odds differently. Namely, odds are equal $\gamma(1+\theta)^{2I(X_1=aa)+I(X_2=bb)+I(X_2=Bb)}$. Table 3 summarizes the considered models. Models 1 and 2 do not contain logistic main effects. Figure 2 shows how the theoretical values of measures of $I(X_1;X_2;Y)$, $IL(X_1;X_2;Y) := \sum_{i,j}\gamma_{ij}^2$, and $LD(X_1;X_2;Y)$ depend on genotypic effect θ. We use $IL(X_1;X_2;Y)$ as a proxy for $LRT(X_1;X_2;Y)$. Generally, the larger the value of θ the larger the value of the considered measures. Models 3, 4, and 5 are additive logistic models (without interaction terms) and thus the logistic interactions γ_{ij} are zero. Interestingly, the rankings corresponding to the considered measures are not consistent across models. For example, interactions in Model 1 seem to be most easily detectable using II, whereas for Model 6 LD yields the largest values among considered interaction measures.



(a) (b) (c)

FIGURE 2 Theoretical values of $I(X_1; X_2; Y)$ (a), $IL(X_1; X_2; Y) := \sum_{i,j} \gamma_{ij}^2$ (b) and $LD(X_1; X_2; Y)$ (c) with respect to value of genotypic effect θ. Prevalence $P(Y = 1) = 0.1$, $MAF = 0.2$

For each setting 100 datasets are generated. In each dataset there are 50 SNPs, which results in $50 \times 49/2 = 1,225$ pairs of SNPs. Among all pairs, we embed one pair of SNPs with interaction generated according to one of Models 1–6. The remaining variables are generated independently of Y. Like most data simulation strategies, we adopt the assumption of Hardy–Weinberg equilibrium for all variables. In data generation algorithm, we control the values of prevalence $\pi_1 = P(Y = 1)$ and minor allele frequency (MAF) $q = P(X_1 = aa) = P(X_2 = bb)$. Detailed description of data generation algorithm is given in the Appendix. Thus, for each dataset we have 1,225 pairs of SNPs, among which there is only one true interaction.

In our experiments power is defined as the fraction of the 100 datasets for which the pair with rank 1, that is, having the largest interaction matches the true interaction. The similar experimental setup was used, for example, by Wang et al. (2011) or Leem et al. (2014). This seems to be useful measure for evaluation of ranking procedures when pairs of genes are ranked according to their interactions strengths. Note that this setup is more challenging than a frequently used scheme that uses power of the corresponding statistical test.

Figure 3 shows how power depends on the genotypic effect θ for Models 1–6, whereas Figure 4 shows how power depends on the sample size n. For Models 3, 4, and 5, power of the logistic regression (LRT) is close to zero, which is concordant with lack of logistic interactions in these models. Interestingly, IIM usually outperforms II, which is consistent with the conclusions of Mielniczuk and Rdzanowski (2017) obtained for performance of statistical tests based on II and IIM. LD is the winner for Model 6, for the remaining models it performs worse than II-based methods. Note that for Model 4, its power is close to 0 similarly as in case of LRT. The best performing method overall is IIM, which is superior for the first five models and it is only slightly inferior to LD for Model 6.

Numerical experiments support the view that there are types of interactions of interest between genes in explaining disease prevalence, which remain undetected by logistic regression methods. Interaction information and especially its variant, modified interaction information, seem to be more universal tool for detecting various types of interactions than the logistic regression.

5 | CONCLUSIONS

In this paper, we discuss the relationship between two important concepts of measuring gene–gene interactions: logistic regression and interaction information. Although both approaches have been used in several GWAS, the relationship between them has not been thoroughly investigated theoretically. In the paper, we contributed to filling this gap. We show that these two measures refer to two different concepts of dependence. This is due to the fact that logistic interaction measure is model-based and which refers to specific parameters of the logistic regression, whereas II is model-free. We stress that how interactions are measured strongly influences our conclusions on their existence and strength. This has significant consequences in genome-wide case–control studies as the interactions detected by the one method may be undetected by the other one and vice versa. We introduced an ordering between interaction measures of being more or less discriminative in detection of interaction. We study properties of normalization constant η of Kirkwood superposition approximation, which plays a role of the dependence index for (X_1, X_2, Y). We show that if $\eta \leq 1$, in particular if genes are independent, then II is more discriminative than the logistic interaction. Moreover, for so-called perfect distributions those measures are equivalent. Furthermore, outside this family, when $\eta \leq 1$ it is established that these two measures of interactions are not equivalent. In particular, we give an example of distributions described by additive logistic models (without interaction terms) for which the interaction defined by II does not vanish. We also characterize situations in which interaction information is positive. In numerical experiments,

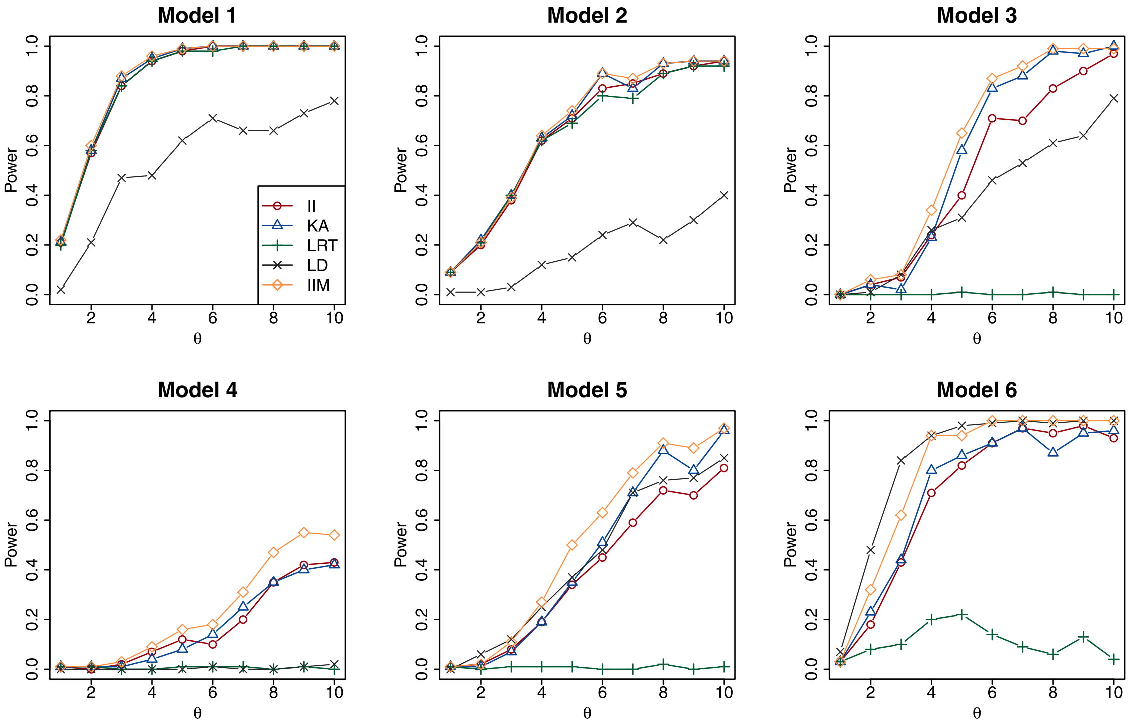


FIGURE 3 Power with respect to genotypic effect θ for simulation models $1-6$. Sample size $n = 1,000$, prevalence $P(Y = 1) = 0.1$, $MAF = 0.2$

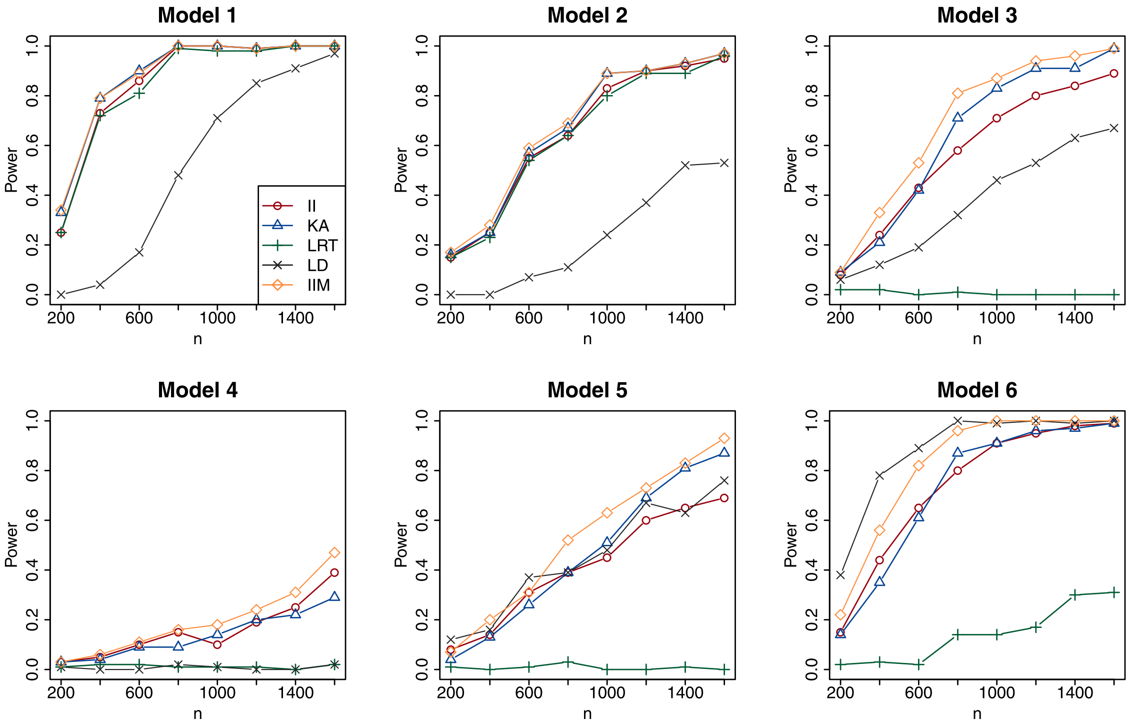


FIGURE 4 Power with respect to sample size n for simulation models 1–6. Genotypic effect $\theta = 6$, prevalence $P(Y = 1) = 0.1$, $MAF = 0.2$

we study usefulness of empirical variant of II and its modifications to rank pairs of genes according to the strength of their interactions when compared with methods based on logistic regression and linkage disequilibrium. It is established that for majority of models, the ranker using modified interaction information is more adequate than other rankers and the effect is especially pronounced for studied additive logistic regression models. Our theoretical findings as well as numerical experiments indicate that interaction information and its modified versions are more universal tools for detecting various types of interaction than logistic regression and linkage disequilibrium measures. In particular, as we have shown that for independent genes II is more discriminative than logistic interaction measure, it may happen that we will miss important interactions by solely using logistic regression.

Some conclusions of this paper can be extended to the case when one or both of the predictors are quantitative. This in particular covers an important case of gene–environment interaction (Hunter, 2005). Theorem 1 holds true, where a general logistic model with quantitative predictor Z and gene X_1 means

$$\log\left[\frac{P(Y=1|X_1,Z)}{P(Y=0|X_1,Z)}\right] = f(Z)+\alpha_1 I(X_1=Aa)+\alpha_2 I(X_1=aa) + \gamma_1(z)I(X_1=Aa)+\gamma_2(z)I(X_1=aa),$$

for certain functions $f(z)$ and $\gamma_i(z)$ and absence of interactions signifies $\gamma_1(z)\equiv\gamma_2(z)\equiv 0$. Validity of other results are still an open research problem.

We also note that constant η that plays an important role in our results can be estimated by the usual plug-in estimator $\hat{\eta}$ and the crucial condition $\eta\leq 1$ can be tested in principle using $\hat{\eta}$. This however would involve determination in its approximate distribution and it is not clear how to derive it. Moreover, the question how II and IL are ordered in the case when $\eta>1$, remains open. Note that in this case II may take either positive or negative values.

ACKNOWLEDGMENTS

We appreciate insightful comments of two reviewers, which helped to improve significantly the presentation of the contribution.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Jan Mielniczuk http://orcid.org/0000-0003-2621-2303
Paweł Teisseyre http://orcid.org/0000-0002-4296-9819

REFERENCES

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). *Wiley series in probability and statistics*. Hoboken, NJ: Wiley-Interscience.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B*, *25*(1), 220–233.

Chanda, P., Sucheston, L., Zhang, A., Brazeau, D., Freudenheim, J., Ambrosone, C., & Ramanathan, M. (2008). AMBIENCE: A novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics*, *180*, 1191–1210.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*(20), 2463–2468.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Review Genetics*, *10*(20), 392–404.

Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd ed.). New York: Wiley.

Darroch, J. (1974). Multiplicative and additive interaction in contingency tables. *Biometrika*, *9*, 207–214.

Dramińsski, M., Dabrowski, M., Diamanti, K., Koronacki, J., & Komorowski, J. (2016). *Discovering networks of interdependent features in high-dimensional problems* (pp. 285–304). Cham: Springer.

Dramiński, M., Kierczak, M., Koronacki, J., & Komorowski, J. (2010). Monte Carlo feature selection and interdependency discovery in supervised classification. In J. Koronacki, Z. W. Raś, S. T. Wierzchoń, & J. Kacprzyk (Eds.), *Advances in machine learning II, studies in computational intelligence* (Vol. *263*, pp. 371–385). Berlin, Heidelberg: Springer.

Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M., … Moore, J. H. (2011). Entropy-based information gain approaches to detect and to characterize gene–gene and gene–environment interactions/correlations of complex diseases. *Genetic Epidemiology*, *35*(7), 706–721.

Fano, F. (1961). *Transmission of information: Statistical theory of communication*. Cambridge: MIT Press.

Han, T. S. (1980). Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, *46*(1), 26–45.

Hu, J., Wang, X., & Wang, P. (2014). Testing gene–gene interactions in genome wide association studies. *Genetic Epidemiology*, *38*(2), 123–134.

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics*, *6*(4), 287–298.

Jakulin, A., & Bratko, I. (2003). Analyzing attribute dependencies. *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*, ECMLPKDD'16 (pp. 229–240). Berlin: Springer.

Jakulin, A., & Bratko, I. (2004). Testing the significance of attribute interactions. *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04 (pp. 52–58). Banff, Canada.

Kang, G., Yue, W. J. Z., Cui, Y. Y. Z., & Zhang, D. (2008). An entropy-based approach for testing genetic epistasis underlying complex diseases. *Journal of Theoretical Biology*, *250*, 362–374.

Lee, W., Sjölander, A., & Pawitan, Y. (2016). A critical look at entropy-based gene–gene interaction measures. *Genetic Epidemiology*, *5*(40), 416–424.

Leem, S., Jeong, H., Lee, J., Wee, K., & Sohn, K. (2014). Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure. *Computational Biology and Chemistry*, *50*, 19–28.

Matsuda, H. (2000). Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 62(3A), 3096–3102.

McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, *19*(2), 97–116.

Mielniczuk, J., & Rdzanowski, M. (2017). Use of information measures and their approximations to detect predictive gene–gene interaction. *Entropy*, *19*, 1–23.

Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N., & 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*(2), 256–261.

Moore, J. H., & Williams, S., (Eds.). (2015). *Epistasis. methods and protocols*. New York: Humana Press.

Nelson, M. R., Kardia, S., Ferrell, R., & Sing, C. F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, *11*, 458–470.

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. *American Journal of Human Genetics*, *69*, 138–147.

Teisseyre, P., Mielniczuk, J., & Dąbrowski, J. M. (2017). Detection of hidden associations and interactions in biomedical data using interaction information. Manuscript..

Wan, X., Yang, C., Yang, Q., Xue, T., Fan, X., Tang, N., & Yu, W. (2010). Boost: A fast approach to detecting gene–gene interactions in genome-wide case–control studies. *American Journal of Human Genetics*, *87*(3), 325–340.

Wang, Y., Liu, G., Feng, M., & Wong, L. (2011). An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics*, *27*(21), 2936–2943.

Wu, T., Chen, Y., Hastie, T., Sobel, E., & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, *25*(6), 714–721.

Yang, Y., He, C., & Ott, J. (2009). Testing association with interactions by partitioning chi-squares. *Annals of Human Genetics*, *73*(1), 109–117.

Zhao, J., Jin, L., & Xiong, M. (2006). Test for interaction between two unlinked loci. *American Journal of Human Genetics*, *79*(5), 831–845.

Zubenko, G. S., Hughes, H. B., & Stiffler, J. S. (2001). D10S1423 identifies a susceptibility locus for Alzheimer's disease in a prospective, longitudinal, double-blind study of asymptomatic individuals. *Molecular Psychiatry*, *6*(4), 413–419.

How to cite this article: Mielniczuk J, Teisseyre P. A deeper look at two concepts of measuring gene-gene interactions: logistic regression and interaction information revisited. *Genet Epidemiol*. 2018;42:187–200. https://doi.org/10.1002/gepi.22108

APPENDIX

AUXILIARY FACTS AND DEFINITIONS

Kullback–Leibler divergence between two measures P_1 and P_2 with corresponding positive discrete mass functions p_1 and p_2 is defined as

$$KL(P_1||P_2) = \sum_x p_1(x)\log\left(\frac{p_1(x)}{p_2(x)}\right),$$

and x ranges over all possible values of both measures. We define now the conditional mutual information $I(X_1;X_2|Y)$, which is used in property (ii), in Section 2.3. Namely,

$$I(X_1;X_2|Y) := \sum_k p(y_k)\sum_{i,j} p(x_i,x_j|y_k)\log\left(\frac{p(x_i,x_j|y_k)}{p(x_i|y_k)p(x_j|y_k)}\right).$$

Note that the conditional mutual information is mutual information of X_1 and X_2 given Y averaged over values of Y. In the same vain as for information gain, $I(X_1;X_2|Y)$ can be interpreted as the expected decrease in the amount of uncertainty of (X_1,X_2) when Y is known. The conditional mutual information equals 0 if and only if two SNPs are conditionally independent given the disease label Y. Analogously, if $H(Y)=\sum_k p(y_k)\log p(y_k)$ denotes entropy of Y, we define conditional entropy of Y given X_1 by

$$H(Y|X_1) = \sum_i p(x_i)\sum_k p(y_k|x_i)\log p(y_k|x_i)$$

and note that the following equality holds (cf. Cover and Thomas, 2006)

$$I(X_1;Y) = H(Y) - H(Y|X_1). \tag{24}$$

Lemma 1. *Assume that X_1 and X_2 are conditionally independent given Y. Then $\gamma_{ij}=0$ for all i,j (i.e., we have logistic model with no interaction terms).*

Proof. Using Bayes theorem and the conditional-independence assumption we have

$$p(y_k|x_i,x_j) = \frac{p(x_i,x_j|y_k)p(y_k)}{p(x_i,x_j)} = \frac{p(x_i|y_k)p(x_j|y_k)p(y_k)}{p(x_i,x_j)} = \frac{p(y_k|x_i)p(y_k|x_j)p(x_i)p(x_j)}{p(x_i,x_j)p(y_k)}.$$

After simple algebraic manipulations, we have

$$\log\left[\frac{P(Y=1|x_i,x_j)}{P(Y=0|x_i,x_j)}\right] = \log\left[\frac{P(Y=0)}{P(Y=1)}\right] + \log\left[\frac{P(Y=1|x_i)}{P(Y=0|x_i)}\right] + \log\left[\frac{P(Y=1|x_j)}{P(Y=0|x_j)}\right] = \mu+\alpha_i+\beta_j,$$

which corresponds to a logistic model with no interaction terms.

□

DATA GENERATION ALGORITHM

In this section, we describe how to generate X_1, X_2, Y, for a given genotypic effect θ, prevalence $\pi_1 = P(Y = 1)$ and MAF $q = P(X_1 = aa) = P(X_2 = bb)$. The odds $P(Y = 1|x_i, x_j)/P(Y = 0|x_i, x_j) = c_{ij}\gamma$ are given in Table 2. Note that c_{ij} depends only on θ. The odds in Table 2 are given up to constant γ, which corresponds to intercept in the logistic regression. In the following algorithm, we determine the value of γ associated with the prevalence π_1.

1. First we determine $p(x_i, x_j)$. We assume Hardy–Weinberg equilibrium for both genes and thus $P(X_1 = aa) = q^2$, $P(X_1 = Aa) = 2q(1 - q)$ and $P(X_1 = AA) = (1 - q)^2$ for a certain $q \in (0, 1)$ and analogously for X_2. As SNPs are generated independently $p(x_i, x_j) = p(x_i)p(x_j)$.
2. Next we calculate γ. From the law of total probability we have
$$\pi_1 = P(Y = 1) = \sum_{i,j} P(Y = 1|x_i, x_j)p(x_i, x_j)$$
$$= \sum_{i,j} \frac{c_{ij}\gamma}{1 + c_{ij}\gamma} p(x_i, x_j).$$
We calculate γ_{opt} by solving numerically the above equality on γ.
3. We adjust $P(Y = 1|x_i, x_j) = c_{ij}\gamma_{opt}/(1 + c_{ij}\gamma_{opt})$, $P(Y = 0|x_i, x_j) = 1 - P(Y = 1|x_i, x_j)$.
4. Using Bayes theorem, we calculate $p(x_i, x_j|Y = 1) = P(Y = 1|x_i, x_j)p(x_i, x_j)/\pi_1$ and $p(x_i, x_j|Y = 0) = P(Y = 0|x_i, x_j)p(x_i, x_j)/(1 - \pi_1)$.
5. We generate Y from Bernoulli distribution with success probability π_1. For $Y = 1$, we generate X_1, X_2 from $p(x_i, x_j|Y = 1)$ and for $Y = 0$ we generate X_1, X_2 from $p(x_i, x_j|Y = 0)$.