Statistical Applications in Genetics and Molecular Biology

Volume 10, Issue 1

2011

Article 12

Information Metrics in Genetic Epidemiology

David L. Tritchler*

Lara Sucheston†

Pritam Chanda‡

Murali Ramanathan**

*State University of New York at Buffalo, dtritch@rogers.com

†Roswell Park Cancer Institute, lara.sucheston@roswellpark.org

‡State University of New York at Buffalo, pchanda@buffalo.edu

**State University of New York at Buffalo, murali@buffalo.edu

Information Metrics in Genetic Epidemiology*

David L. Tritchler, Lara Sucheston, Pritam Chanda, and Murali Ramanathan

Abstract

Information-theoretic metrics have been proposed for studying gene-gene and gene-environment interactions in genetic epidemiology. Although these metrics have proven very promising, they are typically interpreted in the context of communications and information transmission, diminishing their tangibility for epidemiologists and statisticians.

In this paper, we clarify the interpretation of information-theoretic metrics. In particular, we develop the methods so that their relation to the global properties of probability models is made clear and contrast them with log-linear models for multinomial data. Hopefully, a better understanding of their properties and probabilistic implications will promote their acceptance and correct usage in genetic epidemiology. Our novel development also suggests new approaches to model search and computation.

KEYWORDS: gene-environment interaction, gene-gene interactions, K-way Interaction Index, information theory

*Thanks to Jerry Lawless for helpful discussions and key insights. This work was partly supported by NSF, NIH, NSERC and MITACS. Support from the National Multiple Sclerosis Society (RG3743 and a Pediatric MS Center of Excellence Center Grant) and the Department of Defense Multiple Sclerosis Program (MS090122) for the Ramanathan laboratory is gratefully acknowledged. Our work has benefited greatly from the informative input of reviewers.

1 Introduction

Information-theoretic metrics have been proposed for studying gene-gene and gene-environment interactions in genetic epidemiology (Sucheston et al., 2010); Chanda et al., 2007, 2008; Bush et al., 2008; Moore et al., 2006; Dong et al., 2008; Kang et al., 2008). The metrics considered in this paper have been shown to perform well in comparison to other methods in simulation experiments and case studies (Sucheston et al., 2010; Chanda et al., 2007, 2008). Although these metrics have proven very promising, they are typically interpreted in the context of communications and information transmission, diminishing their tangibility for epidemiologists and statisticians. Information concepts are typically defined as sums of entropy terms; such characterizations may be heuristically meaningful, but do not reveal their relevance to data and to global properties of probability models.

Cordell (2009) points out that the relationship of these approaches to more standard statistical methods is unclear. Clayton (2009) questions the usefulness of these methods compared to the logistic model for testing interaction and alternative methods which are based on conditional entropy. In this paper we clarify the interpretation of information-theoretic metrics, which have proven useful in the study of gene-gene and gene-environment interaction in genetic epidemiology. In particular, we develop the methods so that their relation to the global properties of probability models is made clear, and contrast them with log-linear models for multinomial data for further insight. We conclude that information-theoretic measures of statistical interaction have strong connections to more familiar approaches. However, they also differ in ways that may offer advantages and that encourage their further study. Hopefully, a better understanding of their properties and probabilistic implications will promote their acceptance as a promising approach for interaction detection in genetic epidemiology.

Section 2 of the paper, Background and Preliminaries, reviews fundamental concepts in information theory. The basic unit of information, the entropy of a distribution, is introduced. Kullback-Leibler divergence, Information concepts related to statistical independence, and key combinatorial tools are presented. Section 3 introduces and interprets information metrics specifically developed for detecting association in statistical genetic studies. Phenotype Associated Information (*PAI*) measures the association of a phenotype with a set of predictors, e.g. single nucleotide polymorphisms (SNPs). We describe it in statistical terms and give its asymptotic distribution and corresponding permutation test. Section 4 introduces the information metric of primary interest for the epidemiologic study of interaction, the K-way Interaction Index (*KWII*), and describes its properties in depth. Examples comparing *KWII* modeling with log-linear models for multinomial data give insight into the interpretation of *KWII*. The last of these examples demon-

strates that *KWII* is a more parsimonious description of interaction. The relationship between *KWII* and *PAI* is made explicit, and a decomposition of entropy in terms of *KWII* analogous to components of variance is given. The section closes with a discussion of some alternative entropy-based methods. In section 5 we address model search algorithms, briefly describe previous work on the AMBIENCE algorithm (Chanda et al., 2008) for mining interactions using information metrics, and introduce a new algorithm. A discussion section concludes the paper.

2 Background and Preliminaries

2.1 Entropy

Entropies (Cover and Thomas, 2006) are the basic building blocks for the information metrics to be discussed. The Shannon entropy of a probability distribution with density f is defined to be $H(f) = -E[\log f(x)]$ and is described as representing “uncertainty” in x or “information gained” by observing a random variable x from the distribution $f(x)$. Entropy is similar to the concept of variance. For a random variable X with density f we will often substitute the notation $H(X)$ for $H(f)$. More generally, for a true distribution f , the entropy of a different density g is $-E_f[\log g(x)] = H(g)$ where the expectation is over the true distribution.

For a discrete distribution the information gained by observing an outcome x is defined to be $-\log p(x)$. Then the information expected from an experiment is:

$$H(X) = -\sum_x p(x) \log p(x) = E[-\log p(x)]$$

The capitalization of the argument of $H(X)$ emphasizes that H depends on the distribution of the random variable X rather than an observed value, since it is an expectation. We adopt the convention that $H(\emptyset) = 0$.

For a discrete Bernoulli $p(x)$, the case $p = 1$ entails no uncertainty (and zero variance), so entropy is zero. In contrast, when outcomes are equiprobable, i.e., $p = 0.5$, the entropy is maximal because the uncertainty regarding the outcome is maximal. Note that $H = 0$ for both $p = 0$ and $p = 1$, so H is not a one-to-one transformation and it is a condensation of the detail in the probabilities.

2.2 Conditional entropy

Conditional entropy (Cover and Thomas, 2006) is defined in the natural way as:

$$H_X(Y) = E_x E_{y|x} [-\log f(y|x)]$$

where $E_{y|x} [-\log f(y|x)]$, the expectation over the conditional distribution, is entropy calculated within the slice of data with a specific x value. Note that

$$\begin{aligned} H_X(Y) &= E \left[-\log \left(\frac{f(x,y)}{f(x)} \right) \right] \\ &= H(X,Y) - H(X) \end{aligned} \quad (1)$$

This can be thought of as the entropy of y remaining after removing the entropy due to x , i.e., it is residual entropy. The joint entropy is the conditional entropy plus the entropy of the conditioning variable. Conditional entropy is distinguished from entropy of a conditional distribution $H(Y|x) = E_{y|x} [-\log f(y|x)]$, which is a function of x .

2.3 Kullback-Leibler divergence

Kullback-Leibler divergence (Anderson, 2008) is defined to be

$$I(f;q) = E_f \left[\log \left(\frac{f(x)}{q(x)} \right) \right]$$

where f and q are probability density functions and the expectation is with respect to the true density f . We will use f and q to represent densities throughout this paper. Note that $I(f;q) = H(q) - H(f)$ so it can be viewed as comparing the entropies (with respect to f) of the two density functions. Important properties of $I(f;q)$ include positive definiteness; i.e., $I(f;q) \geq 0$ and $I(f;q) = 0$ if and only if $f = q$. Thus the unique density that minimizes the divergence is the true density function.

In our applications of K-L divergence $q(x)$ will be a restriction of $f(x)$ representing a statistical hypothesis. The K-L divergence therefore is the expected log-likelihood ratio.

2.4 Mutual information and statistical independence

The *mutual information* $Inf(X \perp Y)$ (Cover and Thomas, 2006) describing the information in one random vector x about another random vector y is:

$$Inf(X \perp Y) = I(f(x, y); f(x)f(y))$$

This provides a measure of association or correlation between X and Y . In work with an information theory perspective, this correlation measure is denoted as “interaction”, a potential source of confusion for statistical applications; we will restrict the application of the term interaction to relations like nonadditive joint effects. If X and Y are independent, $f(x, y) = f(x)f(y)$, so $Inf(X \perp Y) = 0$. Conversely, by the positive definiteness of divergence $Inf(X \perp Y) = 0$ only when X and Y are independent. Note that:

$$\begin{aligned} Inf(X \perp Y) &= H(Y) + H(X) - H(X, Y) \\ &= H(Y) - H_X(Y) \end{aligned} \quad (2)$$

by (1). Thus, if the uncertainty in Y remaining after accounting for X is large (small) as measured by $H_X(Y)$ then the information X possesses about Y is small (large). Since $Inf(X \perp Y) \geq 0$, $H(Y) \geq H_X(Y)$ and controlling for X can never increase entropy. The role of X here is analogous to that of a blocking variable in ANOVA or a covariate in regression analysis.

Conditional mutual information (Cover and Thomas, 2006) is defined in the natural way:

$$\begin{aligned} Inf(X \perp Y | Z) &= I(f(x, y|z); f(x|z)f(y|z)) \\ &= E \left[\log \left(\frac{f(x, y, z)f(z)}{f(x, z)f(y, z)} \right) \right] \end{aligned}$$

The expectation is over all the variables, so that $Inf(X \perp Y | Z)$ is not a function of Z .

Total correlation information (Chanda et al., 2007) is defined as

$$TCI(X_1, X_2, \dots, X_K) = \sum_{i=1}^K H(X_i) - H(X_1, X_2, \dots, X_K) \quad (3)$$

For example,

$$TCI(X_1, X_2, X_3) = H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3)$$

This is easy to relate to divergence. Direct substitutions of the definitions of the entropies gives:

$$\begin{aligned} TCI(X_1, X_2, \dots, X_K) &= I(f(X_1, X_2, \dots, X_K); \prod_{i=1}^K f(X_i)) \\ &= E \left[\log \frac{f(X_1, X_2, \dots, X_K)}{\prod_{i=1}^K f(X_i)} \right] \end{aligned}$$

For example,

$$TCI(X, Y, Z) = I(f_{x,y,z}; f_x f_y f_z)$$

We see that TCI is the information against complete independence.

2.5 The Möbius transform

The Möbius transform is a key combinatorial tool in the development to follow. We will use it in section 4.1 to relate our interaction metrics to a factorization of the joint density. It will also be central to the algorithm given in section 5.2.

Let g and G be functions defined on the set of all subsets of a finite set Ω . Then the two statements i) $g(x) = \sum_{z \subseteq x} G(z)$ and ii) $G(z) = \sum_{w \subseteq z} (-1)^{|z-w|} g(w)$ are equivalent, where $|\cdot|$ is the cardinality of the argument. Thus a function defined by either expression can be inverted via the other equation of the pair. There is a Fast Möbius Transform (FMT, Thoma (1991)) for fast computation of the functions for all subsets of Ω in $O(N2^N)$ time, where $N = |\Omega|$:

To compute the Möbius transform using the FMT let initially

$$g_0(S) = G(S)$$

for all subsets S . Then iterate for j ranging over the variables in C as follows:

$$g_j(S) = \begin{cases} g_{j-1}(S) & \text{if } j \notin S \\ g_{j-1}(S - \{j\}) + g_{j-1}(S) & \text{if } j \in S \end{cases}$$

The inverse transform is computed by initializing

$$G_0(S) = g(S)$$

for all subsets S . Then iterate for j ranging over the variables in C for as follows:

$$G_j(S) = \begin{cases} G_{j-1}(S) & \text{if } j \notin S \\ -G_{j-1}(S - \{j\}) + G_{j-1}(S) & \text{if } j \in S \end{cases}$$

3 A measure of association: Phenotype Associated Information (PAI)

Since epidemiologic studies usually distinguish a response variable P (phenotype in our case) from the vector of predictors $\mathbf{x}$, a concept $PAI(X, P)$, the Phenotype Associated Information has been introduced (Chanda et al., 2007):

$$PAI(X, P) = TCI(P, X_1, X_2, \dots, X_K) - TCI(X_1, X_2, \dots, X_K)$$

This can be seen to be $I(f_{x,P}; f_{x,P}) = Inf(X \perp P)$. That is, PAI measures the association of P with the multivariate vector $\mathbf{x}$. Note that $PAI(X, P) = PAI(P, X)$ since $Inf(X \perp P) = Inf(P \perp X)$. Using (1), PAI can be characterized as:

$$\begin{aligned} PAI(X, P) &= H(X) + H(P) - H(X, P) \\ &= H(P) - H_X(P) \end{aligned} \tag{4}$$

the reduction in the entropy of P due to X , i.e., the difference between the entropy of P and the residual entropy.

$PAI(X, P) = 0$ if and only if x and P are independent, so we can use properties of probabilistic independence to explicate the meaning of a null PAI . For example, by the Reduction Theorem (Whittaker, 1990), $PAI(X, P) = 0 \Leftrightarrow P \perp X \Rightarrow P \perp X_1, P \perp X_1, \dots, P \perp X_K$, so each x_i alone contributes no information about P , i.e., no marginal effect. By the Block Independence Theorem (Whittaker, 1990), when the joint density of P and X is positive $P \perp X \Leftrightarrow$ both $P \perp X_A | X_B$ and $P \perp X_B | X_A$ are true, where A, B is any partition of the variables in x . Thus the unique information regarding P possessed by any subset of the x 's is zero. For example, $PAI(X, P) = 0$ means that $P \perp X_i | X - X_i$, i.e., P is independent of each X_i given the remaining predictors. This is analogous to every regression coefficient being zero in a multiple prediction model regressing P on the combined set of predictors.

3.1 Statistical distribution of PAI

For multinomial data the sample estimate of PAI is obtained by substituting maximum likelihood estimates of the cell probabilities into the definition. The implication of this for statistical inference about PAI is given by Proposition 7.4.2 of Whittaker (1990):

Proposition: let $\hat{p}(x)$ be the unrestricted maximum likelihood estimate of the probability of cell x , i.e. the cell count $n(x)$ divided by the number of observations N . A model M is defined to be a set of restrictions on the probabilities p . Let $\hat{p}^M(x)$ be the maximum likelihood estimate under M . The deviance of M is

$$dev(M) = 2NI(\hat{p}, \hat{p}^M) = 2 \sum_x n(x) \log \frac{n(x)}{N \hat{p}^M(x)},$$

which is asymptotically chi-squared with degrees of freedom equal to the number of parameters set to zero.

A model $PAI(x, P) = 0$ implies $\hat{p}^M = \hat{p}_x \hat{p}_P$ where $\hat{p}_x$ and $\hat{p}_P$ are unrestricted maximum estimates in the marginal tables of x and P . Since $I(f_{xP}; f_x f_P) = Inf(X \perp P) = PAI(X, P)$, the empirical value of $2N PAI(X, P)$ has an asymptotic chi-squared distribution. The degrees of freedom is $(r_P - 1)(\prod r_{X_i} - 1)$, where r_A is the number of values the random variable A can take. The chi-squared distribution is only an approximation to the exact sampling distribution and is strictly valid only in large samples. Since the cell frequencies decrease as interaction order increases, larger sample sizes are required for testing interactions.

Since $PAI(X, P) = Inf(X \perp P)$, when the sample size is insufficient to permit the reliable use of asymptotic methods a permutation method can be used. Randomly permuting the phenotype assignments will produce observations from the

null distribution with X and P independent, yielding an empirical null distribution of $PAI(X, P)$.

4 A information metric for interaction: K-Way Interaction Index ($KWII$)

Sucheston et al. (2010) and Chanda et al. (2007, 2008) have explored in detail the $KWII$, which is interpreted as an interaction of K variables. The $KWII$ provides a consistent definition of statistical interaction which is of the same form for any choice of joint distribution of the variables.

The $KWII$ of a set of variables X is defined as

$$KWII(X) = - \sum_{Z \subseteq X} (-1)^{|X-Z|} H(Z)$$

where $H(\phi) = 0$.

It is clear from the definition that $KWII$ is symmetric in its variables. The $KWII$ of a single variable is negative entropy. For two variables it is mutual information (i.e. a measure of association/correlation). For three or more variables we will see that it captures the notion of statistical interaction. Note that $KWII(X)$ is an expectation over the set of random variables X , not a function of an observed value x . A characterization of $KWII$ is given by the following recursion:

Proposition: Define $KWII_Y(X) = - \sum_{Z \subseteq X} (-1)^{|X-Z|} H_Y(Z)$. Then

$$KWII(X, Y) = KWII_Y(X) - KWII(X), \quad (5)$$

where Y is a single variable, i.e. $|Y| = 1$, and X is a set of random variables not including Y .

Proof:

$$\begin{aligned} KWII(X, Y) &= - \sum_{Z \subseteq X \cup Y} (-1)^{|X \cup Y| - |Z|} H(Z), \\ &\quad \text{by definition.} \\ &= - \sum_{Z \subseteq X} (-1)^{|X| + |Y| - |Z|} H(Z) - \sum_{Z \subseteq X} (-1)^{|X| + 1 - |Z \cup Y|} H(Z, Y), \\ &\quad \text{decomposing the summation into sets excluding } Y \text{ and sets including } Y. \\ &= \sum_{Z \subseteq X} (-1)^{|X| - |Z|} H(Z) - \sum_{Z \subseteq X} (-1)^{|X| + 1 - |Z \cup Y|} H(Z, Y), \\ &\quad \text{since } |Y| = 1. \\ &= -KWII(X) - \sum_{Z \subseteq X} (-1)^{|X| + 1 - |Z \cup Y|} (H_Y(Z) + H(Y)), \\ &\quad \text{by (1) and the definition of } KWII. \\ &= -KWII(X) - \sum_{Z \subseteq X} (-1)^{|X| - |Z|} H_Y(Z) - H(Y) \sum_{Z \subseteq X} (-1)^{|X| - |Z|}, \\ &\quad \text{since } |Y| = 1. \\ &= -KWII(X) + KWII_Y(X) - H(Y) \sum_{Z \subseteq X} (-1)^{|X| - |Z|} \end{aligned}$$

Note that

$$\sum_{Z \subseteq X} (-1)^{|X|-|Z|} = \sum_{i=0}^{|X|} \sum_{Z \subseteq X, |Z|=i} 1^{|Z|} (-1)^{|X|-|Z|} = \sum_{i=0}^{|X|} \binom{|X|}{i} 1^{|Z|} (-1)^{|X|-|Z|}.$$

Thus the summation in the last term above is equal to $(1-1)^{|X|} = 0$ since it is of the form of the binomial theorem, $\sum_{i=0}^n \binom{n}{i} a^i b^{n-i} = (a+b)^n$.

Note that (5) holds no matter which variable we condition on, i.e., $KWII(X) = KWII_{X_i}(X \setminus X_i) - KWII(X \setminus X_i)$ for any i , where $A \setminus B$ is the set difference, i.e. A with the elements shared with B removed.

The following lemma gives a method for computing $KWII_Y(X)$.

Lemma: Let $KWII(X|Y=y)$ be the value of $KWII(X)$ calculated for the subset of observations taking value y for the random variable Y . Then $KWII_Y(X) = E_y KWII(X|Y=y)$

Proof:

$$\begin{aligned} KWII_Y(X) &= -\sum_{Z \subseteq X} (-1)^{|X|-|Z|} H_Y(Z) \\ &= -\sum_{Z \subseteq X} (-1)^{|X|-|Z|} E_y E_{x|y} [-\log f(Z|y)] \\ &= E_y \left[-\sum_{Z \subseteq X} (-1)^{|X|-|Z|} E_{x|y} [-\log f(z|y)] \right] \\ &= E_y KWII(X|Y=y). \end{aligned}$$

Thus if $KWII(X)$ is computed within each subset of observations with a fixed y value, then $KWII_Y(X)$ is obtained as the expectation of those $KWII$ over the values of y .

Example 3.1. Consider the 2 x 2 tables $\begin{array}{cc|c} 2 & 2 & 4 \\ 1 & 2 & 3 \\ \hline 3 & 4 & 7 \end{array}$ and $\begin{array}{cc|c} 2 & 2 & 4 \\ 2 & 1 & 3 \\ \hline 4 & 3 & 7 \end{array}$.

We consider the two tables to be conditional on a third binary variable. For example, the row and column variables G_r and G_c might represent binary indicator variables corresponding to the dominance coding of the genotypes at two different loci and the tables might be stratified according to a binary phenotype P . For the left table (corresponding to $P=0$) $H(G_r, G_c) = \frac{2}{7} \log(\frac{2}{7}) + \frac{2}{7} \log(\frac{2}{7}) + \frac{2}{7} \log(\frac{2}{7}) + \frac{1}{7} \log(\frac{1}{7}) = -1.35178$; $H(G_r) = \frac{4}{7} \log(\frac{4}{7}) + \frac{3}{7} \log(\frac{3}{7}) = -0.682908$; $H(G_c) = \frac{3}{7} \log(\frac{3}{7}) + \frac{4}{7} \log(\frac{4}{7}) = -0.682908$; $H(\phi) = 0$. $KWII(G_r, G_c|P=0)$ for that table is $-H(G_r, G_c) + H(G_r) + H(G_c) = -0.0140322$. For the right table ($P=1$) $H(G_r, G_c) = \frac{2}{7} \log(\frac{2}{7}) + \frac{2}{7} \log(\frac{2}{7}) + \frac{1}{7} \log(\frac{1}{7}) + \frac{2}{7} \log(\frac{2}{7})$; $H(G_r) = \frac{4}{7} \log(\frac{4}{7}) + \frac{3}{7} \log(\frac{3}{7})$; $H(G_c) = \frac{4}{7} \log(\frac{4}{7}) + \frac{3}{7} \log(\frac{3}{7})$; $H(\phi) = 0$. Since these entropies are the same as those of the left table $KWII(G_r, G_c|P=1) = KWII(G_r, G_c|P=0) = -0.0140322$.

The pooled table (representing the marginal distribution) is

4	4	8
3	3	6
7	7	14

In the margin $H(G_r, G_c) = \frac{4}{14} \log(\frac{4}{14}) + \frac{4}{14} \log(\frac{4}{14}) + \frac{3}{14} \log(\frac{3}{14}) + \frac{3}{14} \log(\frac{3}{14}) = -1.37606$; $H(G_r) = \frac{8}{14} \log(\frac{8}{14}) + \frac{6}{14} \log(\frac{6}{14}) = 0.116121$; $H(G_c) = \frac{7}{14} \log(\frac{7}{14}) + \frac{7}{14} \log(\frac{7}{14}) = -0.693147$; $H(\phi) = 0$. Thus $KWII(G_r, G_c)$ for the marginal table is $-H(G_r, G_c) + H(G_r) + H(G_c) = 0.799029$. By the Lemma $KWII_P(G_r, G_c) = \frac{1}{2} KWII(G_r, G_c|P=0) + \frac{1}{2} KWII(G_r, G_c|P=1) = -0.0140322$. By the Proposition $KWII(G_r, G_c, P) = KWII_P(G_r, G_c) - KWII(G_r, G_c) = -0.0140322 - 0.799029 \neq 0$, indicating $KWII$ interaction.

In contrast, odds ratio is 2 for the left table and $\frac{1}{2}$ for the right table. Interaction on the odds ratio scale is indicated by the differing odds ratios. This example highlights the fact that $KWII$ interaction is not a difference in the information across strata, but the change in information as a consequence of stratification.

4.1 Global properties of $KWII$

In general, factorizations of the density (equivalently, additive decompositions of the log density) determine the dependence relationships among the variables. A simple example is the independence of x and y , where the joint density factors into separate functions of x and y :

$$f(xy) = f(x)f(y)$$

Another example is no interaction of the x, y association with z :

$$f(xyz) = f(xy|z)f(z) = f(xy)f(z)$$

so the x, y association is constant over levels of z . Here the absence of the 3-way interaction is reflected by the fact that no factor is a function of more than two variables.

4.1.1 $KWII$ and factorization of joint density: example

A dependence model is of the form

$$f(x) = \prod_A g_A(x_A) \tag{6}$$

where A ranges over some subsets of indices, and g is a general function (i.e., not necessarily a density function). We assume that each g cannot be further factored. Such a factorization completely determines the dependencies among the variables.

We will show that a particular factorization of the density function leads to the KWII. Considering the case of three variables x ; y ; z demonstrates the general idea. Assuming strictly positive densities can write the identity:

$$\begin{aligned} f(x,y,z) &= \underbrace{f(x)}_{g_x(x)} \underbrace{f(y)}_{g_y(y)} \underbrace{f(z)}_{g_z(z)} \underbrace{\frac{f(x,y)}{f(x)f(y)}}_{g_{x,y}(x,y)} \underbrace{\frac{f(x,z)}{f(x)f(z)}}_{g_{x,z}(x,z)} \underbrace{\frac{f(y,z)}{f(y)f(z)}}_{g_{y,z}(y,z)} \underbrace{\frac{f(x,y,z)f(x)f(y)f(z)}{f(x,y)f(x,z)f(y,z)}}_{g_{x,y,z}(x,y,z)} \\ &= g_x(x) g_y(y) g_z(z) g_{x,y}(x,y) g_{x,z}(x,z) g_{y,z}(y,z) g_{x,y,z}(x,y,z) \end{aligned} \quad (7)$$

Then:

$$\begin{aligned} \log f(x,y,z) &= \log(g_x(x)) + \log(g_y(y)) + \log(g_z(z)) + \log(g_{x,y}(x,y)) + \log(g_{x,z}(x,z)) \\ &\quad + \log(g_{y,z}(y,z)) + \log(g_{x,y,z}(x,y,z)) \end{aligned} \quad (8)$$

From (7), (8), and the definition of entropy and *KWII* we see that:

$$\begin{aligned} E[\log f(x,y,z)] &= KWII(x) + KWII(y) + KWII(z) + KWII(x,y) \\ &\quad + KWII(x,z) + KWII(y,z) + KWII(x,y,z). \end{aligned} \quad (9)$$

For example, referring to (7):

$$\begin{aligned} E[\log g_{x,y,z}(x,y,z)] &= E\left[\log \left(\frac{f(x,y,z)f(x)f(y)f(z)}{f(x,y)f(x,z)f(y,z)} \right)\right] \\ &= -H(x,y,z) + H(x,y) + H(x,z) + H(y,z) - H(x) - H(y) - H(z) \\ &= KWII(x,y,z) \end{aligned}$$

In this way, we can pair each factor in (7) with a corresponding *KWII*. $KWII(x_A)$ corresponds to the expectation of $\log g_A(x_A)$, which will be zero when $g_A(x_A) \equiv 1$, i.e., there is no factor involving all the indices in A and thus no interaction of that order. Otherwise, all variables in A must be considered simultaneously to describe the probability distribution and no simplification is possible. $KWII(X_A) = 0$ when the average value of $g_A(X_A)$ is 1, and measures interaction using the expected log likelihood.

4.1.2 *KWII* and factorization of joint density: general result

The derivation of the general result corresponding to (9) relies on the Möbius transform defined in section 2.5. For positive densities define

$$\Gamma(z) = \sum_{w \subseteq z} (-1)^{|z-w|} \log f(w) \quad (10)$$

so that:

$$E[\Gamma(z)] = \sum_{w \subseteq z} (-1)^{|z-w|} E[\log f(w)] = - \sum_{w \subseteq z} (-1)^{|z-w|} H(w) = KWII(z)$$

We note that (10) is a Möbius transform, thus the inverse transform gives $\log f(x) = \sum_{z \subseteq x} \Gamma(z)$ so the desired factorization of the joint density is $f(x) = \prod_{z \subseteq x} e^{\Gamma(z)}$. Finally,

$$-H(X) = E[\log f(x)] = \sum_{z \subseteq x} E[\Gamma(z)] = \sum_{z \subseteq x} KWII(z). \quad (11)$$

The above equation can be viewed as substituting the factorization of $e^{-H(X)}$ for the factorization of the density function.

4.2 Dependency structure

The sum of all non-singleton $KWII$ accounts for the total correlation information since by (3) and (11) $TCI(X) = \sum_{Z \subseteq X, |Z| \geq 2} KWII(Z)$ and the $KWII$ of a singleton is the negative entropy. Sets of null $KWII$ reflect global dependency statements. For example, note that if $g_{x,y,z}(x,y,z) \equiv 1$ and $g_{x,z}(x,z) \equiv 1$ in (7), then the density has a factorization $h(x,y)h(y,z)$ so that X and Z are conditionally independent given Y . This implies $KWII(X,Y,Z) = KWII(X,Y) = 0$. Likewise, if Y and Z are conditionally independent then $KWII(X,Y,Z) = KWII(Y,Z) = 0$. When the three variables are independent $KWII(X,Y,Z) = KWII(X,Y) = KWII(Y,Z) = KWII(X,Z) = 0$.

4.3 Relation to log-linear statistical models

A decomposition of the form (6) but different from (7) is the log-linear model for contingency tables, e.g.,

$$\log f(x) = u_1(x_1) + u_2(x_2) + u_3(x_3) + u_{1,2}(x_1, x_2) + \dots$$

The interpretation of the log-linear model relies on the implied factorization of the density in the same way as (7). To gain insight into $KWII$ we will compare it with log-linear modeling in three examples.

Example 4.1: Two binary variables

The joint and marginal probabilities are presented in the following table.

p_{11}	p_{10}	$p_{1\cdot}$
p_{01}	p_{00}	$p_{0\cdot}$
$p_{\cdot 1}$	$p_{\cdot 0}$	

The log-linear model factors the density such that for (x_1, x_2) in $\{0, 1\}^2$:

$$\log f(x_1, x_2) = \mu_0 + \mu_1 x_1 + \mu_2 x_2 + \mu_{12} x_1 x_2 \quad (12)$$

where $\mu_0 = \log p_{00}$, $\mu_1 = \log \frac{p_{10}}{p_{00}}$, $\mu_2 = \log \frac{p_{01}}{p_{00}}$ and $\mu_{12} = \log \frac{p_{11}p_{00}}{p_{01}p_{10}}$. μ_{12} is the log cross-product ratio, where the cross-product ratio of a 2x2 table is defined as $cpr = \frac{p_{11}p_{00}}{p_{10}p_{01}}$. The parameter μ_{12} is a measure of the association x_1 and x_2 , which is *KWII* interaction for the simple case of the 2x2 table.

With information metrics, to get a factorization like (7) we write:

$$\begin{aligned} \log f(x_1, x_2) &= \log f(x_1) + \log f(x_2) + \log \frac{f(x_1, x_2)}{f(x_1)f(x_2)} \\ &= \log p_{1\cdot}^{x_1} p_{0\cdot}^{1-x_1} + \log p_{\cdot 1}^{x_2} p_{\cdot 0}^{1-x_2} + \log \frac{f(x_1, x_2)}{f(x_1)f(x_2)} \\ &= x_1 \log p_{1\cdot} + (1 - x_1) \log p_{0\cdot} + x_2 \log p_{\cdot 1} + (1 - x_2) \log p_{\cdot 0} + \log \frac{f(x_1, x_2)}{f(x_1)f(x_2)}, \end{aligned}$$

so that

$$E[\log f(x_1, x_2)] = -H(x_1) - H(x_2) + Inf(x_1 \perp x_2)$$

KWII interaction (i.e. association/correlation in this case) is measured by $Inf(x_1 \perp x_2)$, which equals 0 iff x_1 and x_2 are independent. An alternate description of $Inf(x_1 \perp x_2) = 0$ which better generalizes to interaction among more than two variables is $H(X_1) = H_{X_2}(X_1)$ by (2).

It is informative to inspect the last term above, $Inf(x_1 \perp x_2) = KWII(x_1, x_2) = E[\log \frac{f(x_1, x_2)}{f(x_1)f(x_2)}]$ in detail:

$$E[\log \frac{f(x_1, x_2)}{f(x_1)f(x_2)}] = p_{11} \log \frac{p_{11}}{p_{1\cdot} p_{\cdot 1}} + p_{10} \log \frac{p_{10}}{p_{1\cdot} p_{\cdot 0}} + p_{00} \log \frac{p_{00}}{p_{0\cdot} p_{\cdot 0}} + p_{01} \log \frac{p_{01}}{p_{0\cdot} p_{\cdot 1}}$$

Each term corresponds to a cell of the 2 x 2 table going clockwise from the upper left, and is a measure of independence in that cell weighted by probability of that cell. The sum is thus a weighted average of independence metrics. This is reminiscent of the chi-squared test for independence in a 2 x 2 table, which combines within-cell measures of deviation from independence.

To compare the log-linear interaction parameter μ_{12} and $KWII(x_1, x_2)$ let us assume that the log-linear model holds so that (12) implies that $KWII(x_1, x_2)$ is the expectation of

$$\log \frac{f(x_1, x_2)}{f(x_1)f(x_2)} = \log f(x_1, x_2) - \log f(x_1) - \log f(x_2)$$

$$\begin{aligned}
 &= \mu_0 + \mu_1 x_1 + \mu_2 x_2 + \mu_{12} x_1 x_2 \\
 &\quad - x_1 \log p_{1\cdot} - (1 - x_1) \log p_{0\cdot} - x_2 \log p_{\cdot 1} - (1 - x_2) \log p_{\cdot 0} \\
 &= \mu_0 - \log p_{0\cdot} - \log p_{\cdot 0} + x_1 (\mu_1 - \log p_{1\cdot} + \log p_{0\cdot}) \\
 &\quad + x_2 (\mu_2 - \log p_{\cdot 1} + \log p_{\cdot 0}) + x_1 x_2 \mu_{12} \\
 &= x_1 x_2 \mu_{12} + \log \frac{p_{00}}{p_{0\cdot} p_{\cdot 0}} + x_1 \log \frac{p_{10}/p_{00}}{p_{1\cdot}/p_{0\cdot}} + x_2 \log \frac{p_{01}/p_{00}}{p_{\cdot 1}/p_{\cdot 0}}
 \end{aligned}$$

We see that *KWII* includes the log-linear parameter $\mu_{12} = \log \frac{p_{11}p_{00}}{p_{01}p_{10}}$ as a component, plus three additional components which seem to also reflect the notion of interaction. For example, nonzero $\frac{p_{00}}{p_{0\cdot} p_{\cdot 0}}$ is a deviation from the statistical independence of x_1 and x_2 , while $\frac{p_{10}/p_{00}}{p_{1\cdot}/p_{0\cdot}}$ contrasts the odds of x_1 when $x_2 = 0$ to the odds of x_1 ignoring x_2 to measure the heterogeneity of the x_1 effect with respect to x_2 . A similar interpretation holds for $\frac{p_{01}/p_{00}}{p_{\cdot 1}/p_{\cdot 0}}$.

When row and column sums are fixed, μ_{12} is unaffected by marginal probabilities and is known to capture all of the conditional information about the x_1, x_2 association, ie. it is the sufficient statistic with respect to the conditional distribution. The *KWII* is affected by the marginal probabilities and is not conditional on the margins. This is further discussed in section 4.7.

Example 4.2: Three binary variables

The log-linear model factors the density such that:

$$\log f(x_1, x_2, x_3) = \mu_0 + \mu_1 x_1 + \mu_2 x_2 + \mu_3 x_3 + \mu_{12} x_1 x_2 + \mu_{13} x_1 x_3 + \mu_{23} x_2 x_3 + \mu_{123} x_1 x_2 x_3$$

The 2-way -terms are calculated from 2 x 2 marginal tables as in the previous example. The 3-way -term is $\log \frac{cpr(x_1, x_2 | x_3=1)}{cpr(x_1, x_2 | x_3=0)}$, where *cpr* is the cross product ratio of the corresponding marginal 2x2 table. It measures the effect of x_3 on the x_1, x_2 association.

With information metrics we write the expected log density as:

$$E \left[\log \left(f(x_1) f(x_2) f(x_3) \frac{f(x_1, x_2)}{f(x_1) f(x_2)} \frac{f(x_1, x_3)}{f(x_1) f(x_3)} \frac{f(x_2, x_3)}{f(x_2) f(x_3)} \frac{f(x_1, x_2, x_3) f(x_1) f(x_2) f(x_3)}{f(x_1, x_2) f(x_1, x_3) f(x_2, x_3)} \right) \right]$$

The 3-way interaction derives from the last factor:

$$\begin{aligned}
 KWII(x_1, x_2, x_3) &= E \left[\log \frac{f(x_1, x_2, x_3) f(x_1) f(x_2) f(x_3)}{f(x_1, x_2) f(x_1, x_3) f(x_2, x_3)} \right] \\
 &= E \left[\log \left(\frac{f(x_1, x_2, x_3) f(x_3)}{f(x_1, x_3) f(x_2, x_3)} \right) \left(\frac{f(x_1) f(x_2)}{f(x_1, x_2)} \right) \right] \\
 &= Inf(x_1 \perp x_2 | x_3) - Inf(x_1 \perp x_2)
 \end{aligned}$$

Similarly to the log-linear model the 3-way term represents the effect of effect of x_3 on the x_1, x_2 association.

Example 4.3: Two trinomial genotype variables

Here x_1 and x_2 take values in $\{0, 1, 2\}$ representing genotype. In this case the log-linear model uses four parameters to model the x_1, x_2 interaction, $\mu_{12}(1, 1)$, $\mu_{12}(1, 2)$, $\mu_{12}(2, 1)$, $\mu_{12}(2, 2)$. In contrast, only the single metric $KWII(x_1, x_2)$ is required. This reduction of degrees of freedom is different from that resulting from recoding genotypes to model restricted genetic models such as dominant or recessive, as it makes no genetic assumptions and relies only on the distribution of genotypes. In general, for a k th-order interaction of trinomial genotype variables the log-linear model is expressed by 2^k parameters, as opposed to a single $KWII$. So the information-based model is more parsimonious, especially for high order interactions. There is a loss of detail, but since the number of parameters poses a huge challenge when screening genome-wide data for interactions the principled reduction of parameters inherent in the information approach is advantageous. For example, if we are analyzing haplotypes with r variants, a k^{th} order interaction is expressed by $(r - 1)^k$ interaction parameters using the log-linear parameterization, but only a single $KWII$.

4.4 Relation of $KWII$ and PAI

By (4) and (11),

$$\begin{aligned} PAI(X, P) &= H(X) + H(P) - H(X, P) \\ &= -\sum_{Z \subseteq X} KWII(Z) + H(P) + \sum_{Z \subseteq X \cup P} KWII(Z) \\ &= \sum_{Z \subseteq X} KWII(Z, P) + H(P) \end{aligned} \quad (13)$$

PAI is an imperfect but useful tool for screening potential interactions. In the AMBIENCE algorithm to be described later an interaction term $KWII(Z)$ is considered whenever $PAI(Z)$ is large. However a $KWII$ can be either positive or negative, so non-zero terms in (13) could conceivably cancel each other to make PAI small when there are interactions present. However Han (1980) shows that $KWII$ is nonnegative when the distribution is sufficiently close to complete independence so cancelation may be considered unlikely, supporting AMBIENCE's use of PAI as a screening tool for interactions.

There is a converse to (13). Since by (13) $PAI(X, P) - H(P) = \sum_{Z \subseteq X} KWII(Z, P)$, Möbius inversion gives

$$\begin{aligned} KWII(X, P) &= \sum_{Z \subseteq X} (-1)^{|X-Z|} (PAI(Z, P) - H(P)) \\ &= \sum_{Z \subseteq X} (-1)^{|X-Z|} PAI(Z, P) \end{aligned}$$

since $\sum_{Z \subseteq X} (-1)^{|X-Z|} H(P) = H(P) \sum_{Z \subseteq X} (-1)^{|X-Z|} (-1)^{|X-Z|} = 0$.

4.5 Similarity to components of variance

We can write a decomposition of the entropy of the phenotype P . By (4) and (13)

$$H(P) = \sum_{\emptyset \subset z \subseteq x} KWII(Z, P) + H_X(P)$$

Since the last term above is a residual, the expression is of a form similar to components of variance, with the distinction that these components can be positive or negative.

4.6 Statistical distribution of $KWII$

As was the case for PAI , for multinomial data the sample estimate of $KWII$ is obtained by substituting maximum likelihood estimates of the cell probabilities into the definition. Han (1980) shows that if a condition called semi-independence holds then $2N KWII(X)$ converges in distribution asymptotically to a chi-squared statistic with degrees of freedom $\prod (r_{Xi} - 1)$, where r_A is the number of values the random variable A can take. Further, if $X \neq Y$, $2N KWII(X)$ and $2N KWII(Y)$ asymptotically approach independent chi-squared distributions.

Han's results are based on a Taylor series expansion around the distribution for which the variables are completely independent, and so are accurate for cases close to that condition. Since we wish to employ our methods when strong correlations exist the practical usefulness of the asymptotic results needs to be established. Also, the asymptotic relation of semi-independence conditions to the factorization of the density needs to be investigated. Further study is required to justify the use of these results in applications.

4.7 Related information approaches

Clayton (2009) gives a stimulating critique of $KWII$, which he terms "synergy". He advocates an alternative information metric for interaction based on the uncertainty remaining if all lower order relationships were known. Good(1963) derived the associated metric to be the maximum entropy over all possible distributions constrained by fixing the lower order margins, and showed that this corresponds to interaction in a logistic regression model. Our discussion of example 4.1 supports Clayton's observation that $KWII$ is influenced by the marginal probabilities. In the simple two variable case the argument for the specification of fixed margins in the analysis of 2 x 2 tables is not conclusive. In that case the odds ratio with fixed margins (a logistic model parameter) and unconditional inference (e.g. via

a chi-squared test) are both acceptable. This suggests the comparison of *KWII* and Good's maximum entropy may be analogous to the contrast between conditional and unconditional approaches to statistical inference for contingency tables. Agresti (2002) discusses the relative merits of those two approaches. To summarize, there is information about association in the table margins (Agresti, 2002; Zhu and Reid, 1994), but it is a small amount. Future work will consider this issue for genetic applications and general dimensionality. Conditional inference methods for multinomial data tend to be computationally complex, and this may be an important consideration.

Clayton also shows the value of *KWII* is not invariant under case-control sampling. The seriousness of this concern will be addressed in future work. The important question is whether *inferences* are rendered invalid. *KWII* is not really an interpretable parameter in the way that odds ratio contrasts are; it is used screening rather than for interpreting the strength of effects. So this criticism, while valid for the magnitude of the metric, may not discourage the use of *KWII* for case-control studies. This is an important question which needs further study.

Reconstructability Analysis (RA) (Zwick, 1994; Shervais et al., 2010) is an information-based approach related to maximum entropy. This approach reduces the dimension from the saturated model by projecting the joint distribution onto a set of margins. The margins specify the reduced model. The marginal distributions are then composed to form a more parsimonious estimate of the joint distribution whose fit to the data will be tested using Good's maximum entropy method to evaluate the model. Thus the comments of the previous paragraph also apply to RA.

Maximum entropy and RA are closely related to log-linear and logistic modeling. Importantly, they require iterative computational methods, unlike *KWII*. Thus *KWII* may be the only feasible choice for high-dimensional model search.

5 Model search

The non-iterative computation required by the information metrics of this paper makes them candidates for the large-scale screening of genetic interactions, when many other methods (e.g. logistic regression, reconstructability analysis) are infeasible. However, large problems are still daunting. Using the definition to compute *KWII* for all possible interactions (i.e. all possible combinations of variables) is of complexity $O(3^N)$, where N is the number of variables. Thus efficient algorithms and heuristics must be used. The AMBIENCE algorithm is a heuristic search strategy for interaction detection.

5.1 The AMBIENCE algorithm

The AMBIENCE algorithm (Chanda et al., 2008) screens to find sets of variables with significant PAI of order $\leq k$, where the significance is determined by the chi-squared test or the permutation method given in section 3.1. It does this greedily, building candidate i^{th} -order PAI by adding a new variable to a significant $(i-1)^{th}$ -order PAI . Finally for each PAI selected the $KWII$ with the same arguments is tested. Note that a significant PAI could be due to any of the $KWII$ given by (13), not just the highest order one. AMBIENCE accounts for this by its greedy choice of PAI , which nests the interaction orders. A i^{th} -order PAI is not tested unless it contains an $(i-1)^{th}$ -order PAI , which contains a significant $(i-2)^{th}$ -order PAI , and so on. For example, this means that a two-way interaction between SNPs is not tested unless there is significant marginal association between at least one of the SNPs and the disease phenotype. Testing the associated nested $KWII$ can be viewed as a heuristic approximation to testing all the $KWII$ implied by (13) for a significant PAI .

5.2 The Möby Quick algorithm

Thoma (1991) gave an efficient algorithm, the Fast Möbius Transform (FMT), for computing $KWII(X)$ for all subsets X . As presented, the FMT calculates all $KWII$ up to and including order N in $O(N2^N)$ time, only a factor of N greater than the N^{th} order interaction alone. As it stands, the FMT is not of use to us since it is infeasible to consider very high order interactions orders and we limit the order to $k \ll N$. Happily, the basic FMT is easily modified to efficiently calculate $KWII$ only up to order k . The modified algorithm, of complexity $O(kN^{k+1})$, is given in the Appendix.

We will usually wish to restrict the interactions so that every one includes the same fixed size c set of variables, For example, in our application every interaction of interest involves the phenotype P , so $c = 1$. We may also wish to fix other variables, e.g., gender, so that every interaction involves both phenotype and gender so $c = 2$. The new algorithm enables this to be done, and the complexity is reduced to $O(kcN^k + k(N-c)^{k+1})$. Although the complexity is polynomial in N of order $k + 1$, N can be huge. It will be necessary to reduce N , possibly by selecting those variables found to play a role in significant PAI .

In this paper we propose an algorithm, Möby Quick, which is complementary to AMBIENCE, based on the modification of the FMT given in the Appendix. After significant PAI are identified by AMBIENCE it is usually the case that a small subset Θ of variables are seen to be participating in the interactions. Rather than

computing *KWII* individually as in *AMBIENCE*, Möby Quick computes all possible interactions up to order k involving variables in Θ in $O(k|\Theta|^{k+1})$ time. The algorithm casts a wider net than *AMBIENCE*, and if $|\Theta|$ is small it is computationally feasible.

6 Discussion

In this paper we presented a likelihood factorization which identifies quantities which capture the notion of statistical interaction. The information theoretic interaction metrics arise from the expectation of the log-likelihood and sample estimates of all the metrics we have considered arise from maximum likelihood estimates of the entropies. An advantage of the information approach is that the meaning of the metrics and their connection to the statistical distribution of the data is clear, as opposed to many machine learning algorithms for interaction detection.

For categorical variables with more than two levels the expected log-likelihood reduces the number of parameters compared to log-linear models and logistic regression. This gives a much more parsimonious description of the interactions among the variables. Some detail is lost, but parsimony may be a more important consideration when searching for interactions in high dimensional data. In addition, the computations required are non-iterative, which greatly extends their feasibility compared with methods requiring numerical optimization such as logistic regression and reconstructability analysis.

The calculus of the information metrics facilitates a variety of re-expressions of the metrics, which suggest different approaches to model search and computation. The potential algorithms suggested in this paper will be explored in future work.

The goal of this paper was to connect information-theoretic metrics to probability models so that they can be better understood in the context of data analysis. Along the way we also uncovered some differences. For genetic markers with many levels there are many fewer parameters produced in the information approach. This may be critical in the case of haplotype analysis, where for haplotypes with r variants, a k^{th} order interaction is expressed by $(r-1)^k$ interaction parameters using the log-linear parameterization, but only a single *KWII*. The way the metrics are defined also makes transformations among them easy using powerful algorithmic tools.

We have also highlighted areas in need of further development. The distribution of *KWII* needs further study. Clayton (2009) has shown that the influence of the sampling plan and the influence of marginal probabilities will be key question in future comparisons of interaction detection methods. Also, we will explore

improved modeling strategies using these metrics. In most of the algorithms quantities are not refined based on previous modeling steps. That is, a *KWII* at the end of AMBIENCE is the same number as at the beginning. The AMBIENCE algorithm just tells you that you should look at it. Our future work will explore approaches where previous decisions do simplify the computation of *KWII*, essentially using fewer parameters.

APPENDIX: Modified FMT

The following algorithm calculates *KWII* for all variable subsets of orders = k , each including all of the variables in a fixed set C . Let Ω be all size = k subsets of the complete set of N variables. For a fixed set C of the variables let Ω^- represent all the variable sets in Ω which contain the variables in C . Let initially

$$f_0(S) = H(S)$$

for all $S \in \Omega$. Iterate for j ranging over the variables in C for $S \in \Omega$ as follows:

$$f_j(S) = \begin{cases} f_{j-1}(S) & \text{if } j \notin S \\ -f_{j-1}(S - \{j\}) + f_{j-1}(S) & \text{if } j \in S \end{cases} \quad (14)$$

Then iterate 14 for j ranging over the variables in $\bar{C}$ for $S \in \Omega^-$. Finally, after step N when all variables have been iterated over, set $KWII(S) = f_N(S)$, $S \in \Omega^-$.

Proof: If $k = N$ and $C = \phi$, the algorithm is the original FMT. If $k < N$ and $C = \phi$, a iteration 14 for a set S depends only on subsets of S and thus is unaffected. If $C \neq \phi$, the initial iteration over C is the same as the basic FMT. For the second stage iterating over Ω^- , a step of (14) for a set S containing C depends only on lower-order subsets containing C , which have been computed.

REFERENCES

- Agresti A (2002) Categorical Data Analysis. Wiley, New York
 Anderson D (2008) Model Based Inference in the Life Sciences. Springer, New York
 Bush W, Edwards T, Dudek S, McKinney B, Ritchie M (2008) Alternative contingency measures improve the power and detection of multifactor dimensionality reduction. BMC Bioinformatics 9:238-255
 Chanda P, Zhang A, Brazeau D, Sucheston L, Freudenheim JL, Ambrosone C, Ramanathan M (2007) Information-theoretic metrics for visualizing gene-environment interactions. Am J Hum Genet 81:939-963

- Chanda P, Sucheston L, Zhang A, Brazeau D, Freudenheim JL, 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
- Clayton, D (2009) Prediction and interaction in complex disease genetics: Experience in type 1 diabetes. *PLoS Genetics* 5:e1000540
- Cordell H (2009) Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics* 10:392-404
- Cover T, Thomas J (2006) *Elements of Information Theory*. Wiley, New York
- Dong C, Chu X, Wang Y, Jin L, Shi T, Huang W, Li Y (2008) Exploration of gene-gene interaction effects using entropy-based methods. *Eur J Hum Genet* 16:229-235
- Good I (1963) Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics* 34:911-934
- Han TS (1980) Multiple mutual information and multiple interactions in frequency data. *Information and Control* 46:26-45
- Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* 241:252-261
- Kang, G et al. (2008) An entropy-based approach for testing genetic epistasis underlying complex diseases. *Journal of Theoretical Biology* 250:362-374
- Shervais S, Kramer P, Westaway S, Cox N, Zwick M. (2010) Reconstructability analysis as a tool for identifying gene-gene interactions in studies of human diseases. *Statistical Applications in Genetics and Molecular Biology* 9(1), Article 18
- Sucheston L, Chanda P, Zhang A, Trichtler D, Ramanathan M. (2010) Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity. *BMC Genomics*, 11:487
- Thoma, H M (1991) Belief function computations. In *Conditional Logic in Expert Systems*, I R Googman, M MK Gupta, H T Nguyen, G S Rogers, Eds. Amsterdam: North-Holland, pp. 269-308
- Whittaker J (1990) *Graphical Models in Applied Multivariate Analysis*. Wiley, New York
- Zhu Y, Reid N (1994) Information, ancillarity, and sufficiency in the presence of nuisance parameters. *Canadian Journal of Statistics* 22:111-123
- Zwick M (2004) An overview of reconstructability analysis. *Kybernetes* 33:877-905