Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT

PERGAMON

Pattern Recognition 38 (2005) 623–636

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

The coefficient of intrinsic dependence (feature selection using el CID)

Tailen Hsing[a], Li-Yu Liu[a], Marcel Brun[b], Edward R. Dougherty[c, d, *]

[a] *Department of Statistics, Texas A&M University, College Station, TX, USA*
[b] *Department of Biochemistry and Molecular Biology, University of Louisville, KY, USA*
[c] *Department of Electrical Engineering, Texas A&M University, 3128 TAMU, College Station, TX 77843-3128, USA*
[d] *Department of Pathology, University of Texas M. D. Anderson Cancer Center, Houston, TX, USA*

Received 20 May 2004; received in revised form 13 September 2004; accepted 13 September 2004

Abstract

Measuring the strength of dependence between two sets of random variables lies at the heart of many statistical problems, in particular, feature selection for pattern recognition. We believe that there are some basic desirable criteria for a measure of dependence not satisfied by many commonly employed measures, such as the correlation coefficient, Briefly stated, a measure of dependence should: (1) be model-free and invariant under monotone transformations of the marginals; (2) fully differentiate different levels of dependence; (3) be applicable to both continuous and categorical distributions; (4) should not have the dependence of X on Y be necessarily the same as the dependence of Y on X; (5) be readily estimated from data; and (6) be straightforwardly extended to multivariate distributions. The new measure of dependence introduced in this paper, called the *coefficient of intrinsic dependence* (CID), satisfies these criteria. The main motivating idea is that Y is strongly (weakly, resp.) dependent on X if and only if the conditional distribution of Y given X is significantly (mildly, resp.) different from the marginal distribution of Y. We measure the difference by the normalized integrated square difference distance so that the full range of dependence can be adequately reflected in the interval [0, 1]. The paper treats estimation of the CID, provides simulations and comparisons, and applies the CID to gene prediction and cancer classification based on gene-expression measurements from microarrays.

Keywords: Classification; Correlation; Dependence; Feature-selection; Microarray; Prediction

1. Introduction

Describing the relationship between two random variables is central to almost all scientific inquiries, and the strength of dependence between two sets of random quantities is at the heart of many statistical problems. In pattern recognition the dependency issue arises in the context of feature selection. In its ideal form, the feature-selection problem is to select a subset of k features from a set of n features that provides an optimal classifier with minimum error among all optimal classifiers for subsets of size k. The inherent combinatorial nature of the problem is readily seen from the fact that, in the absence of mitigating distributional prior knowledge, all k-element subsets must be checked to assure selection of the optimal k-element feature set [1].

* Corresponding author. Department of Electrical Engineering, Texas A&M University, 3128 TAMU, College Station, TX 77843-3128, USA. Tel.: +1 409 845 7441; fax: +1 409 845 6259.
E-mail address: e-dougherty@tamu.edu, edward@ee.tamu.edu (E.R. Dougherty).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.09.002

Many methods have been proposed to approach the problem suboptimally. These are often partitioned into two types of feature-selection algorithms. Proceeding according to some search procedure, *wrapper* algorithms design classifiers on feature subsets and evaluate their performances. Commonly employed wrapper algorithms include forward/backward selection and sequential forward floating selection (SFFS), where the number of features to be adjoined and deleted is not fixed, but is allowed to "float" [2]. *Filter* algorithms do not design and test subset-based classifiers but instead employ test statistics to determine a feature subset that can be expected to perform well, such as measuring the correlation between potential features and the target. When the number of potential features is very large, say in the thousands, it is commonplace for the full set of features to be first filtered to provide a smaller set on which to apply a wrapper algorithm. Moreover, when filtering features, one need not only consider bivariate relations between features and the target; one can instead consider relations between more than one feature and the target in order to avoid missing key features that only provide discrimination when used in conjunction with other features. We take the latter approach in applying the measure of dependence proposed in this paper.

When considering the worth of a measure of dependence, to be used for feature selection or otherwise, it is important to ask the following question: What constitutes some of the basic criteria of a measure of dependence that makes good statistical sense, is flexible, and is sufficiently powerful to analyze a wide variety of data? We answer with the following criteria:

(C1) The measure is model-free in the sense that no distributional or functional assumptions are placed on the variables; it is also invariant under monotone transformations of the marginals. This allows us to estimate the measure from data without having to verify model assumptions and/or make transformations. This consideration is important for data for which the distributional properties are not well understood.
(C2) The measure can fully differentiate different levels of dependence. For instance, the measure of dependence of a response variable on a predictor variable should become stronger if additional information is included in the predictor variable, or if the model is such that the response variable is functionally/stochastically more dependent on the predictor variable.
(C3) The measure applies to both continuous and categorical distributions.
(C4) The measure takes causality into consideration; in other words, the dependence of X on Y should not be necessarily the same as the dependence of Y on X.
(C5) The measure can be readily estimated from data.
(C6) The measure is straightforwardly extended to multivariate X and/or Y.

The correlation coefficient is by far the most popular measure of dependence. However, it fails (C1), (C2), (C4), and (C6); it also partially fails (C3) when the variables are genuinely categorical (i.e. qualitative). Two other well-known measures of dependence, Spearman's ρ and Kendall's τ, also fail a number of these criteria.

This paper introduces a new dependence measure, called the *coefficient of intrinsic dependence* (*CID*), that satisfies all of criteria (C1)–(C6). The motivating idea is that Y is strongly (weakly, resp.) dependent on X if and only if the conditional cumulative distribution function (cdf) of Y given X is significantly (mildly, resp.) different from the marginal cdf of Y. We measure the difference by the normalized integrated square difference distance so that the full range of dependence can be adequately reflected as numbers in [0,1].

Recently, the problem of measuring dependence has received a significant renewed interest in the context of genomic data, for instance, microarray data, where the determination of relationships between diseases and genes, and interdependence between genes, are important to functional understanding, and where the number of features is typically in the thousands. We will apply the CID for two central problems of genomics: multivariate prediction and the design of classifiers to discriminate among diseases.

2. el CID: the coefficient of intrinsic dependence

If $X_1, \ldots, X_k, Y$ are random variables on a probability space $(\Omega, \mathcal{F}, P)$, the *CID* of Y given the random vector $X = (X_1, \ldots, X_k)$ is defined by

$$\mathrm{CID}(Y|X) = \frac{\int_{-\infty}^{\infty} \mathrm{Var}(\mathrm{E}(\delta_Y(v)|X))\,\mathrm{d}F_Y(v)}{\int_{-\infty}^{\infty} \mathrm{Var}(\delta_Y(v))\,\mathrm{d}F_Y(v)}, \tag{1}$$

where F_Y is the cdf for Y, $\mathrm{E}(\delta_Y(v)|X)$ is the conditional expectation of the random variable $\delta_Y(v)$ given X, and for $y \in (-\infty, \infty)$,

$$\delta_y(u) = I(y \leqslant u), \quad u \in (-\infty, \infty). \tag{2}$$

Alternatively,

$$\mathrm{CID}(Y|X) = \frac{\int_0^1 \mathrm{Var}(\mathrm{E}(\delta_{\tilde{Y}}(u)|X))\,\mathrm{d}u}{\int_0^1 \mathrm{Var}(\delta_{\tilde{Y}}(u))\,\mathrm{d}u}, \tag{3}$$

where $\tilde{Y} = F_Y(Y)$. Observe that

$$\mathrm{Var}(\mathrm{E}(\delta_{\tilde{Y}}(u)|X)) = E\Big(P(\tilde{Y} \leqslant u|X) - P(\tilde{Y} \leqslant u)\Big)^2. \tag{4}$$

Remark. (a) The identity in Eq. (4) gives the essential motivation of the CID, namely, that if Y and X are weakly/strongly dependent, then $\mathrm{Var}(\mathrm{E}(\delta_{\tilde{Y}}(u)|X))$ will be small/large. The denominator is a normalization that constricts the CID to be in [0, 1].

(b) The definition of the CID directly extends to the case where conditioning is relative to an arbitrary σ-field $\mathscr{H} \subset \mathscr{F}$. In this case we have the CID of Y given $\mathscr{H}$, denoted by $\mathrm{CID}(Y|\mathscr{H})$, defined in the same manner as $\mathrm{CID}(Y|X)$ except that we use $\mathrm{E}(\delta_Y(v)|\mathscr{H})$, the conditional expectation of the random variable $\delta_Y(v)$ given the σ-field $\mathscr{H}$, instead of $\mathrm{E}(\delta_Y(v)|X)$. The more general definition of $\mathrm{CID}(Y|\mathscr{H})$ reduces to $\mathrm{CID}(Y|X)$ by letting $\mathscr{H}$ be the σ-field generated by X. From this perspective we observe that CID is invariant under any transformation performed on X so long as the generated σ-field remains the same.

Closed-form expressions of CID are uncommon, as is the case with other measures of dependence, including the correlation. The following example contains three theoretical computations of the CID.

Example 2.1. (a) If both X and Y are binary then $\mathrm{CID}(Y|X)$ equals the squared correlation between X and Y. This connection with the correlation breaks down even in the case of ternary variables.

(b) If (X, Y) is distributed as bivariate normal with correlation ρ, then

$$\mathrm{CID}(Y|X) = 6\sum_{k=1}^{\infty}\frac{\rho^{2k}}{k!} \times \int_{-\infty}^{\infty}\left(\phi^{(k-1)}(u)\right)^2\phi(u)\,\mathrm{d}u, \tag{5}$$

where ϕ is the standard normal pdf. Note that $\mathrm{CID}(Y|X)$ is a strictly increasing function of $|\rho|$. The derivation, given in the Appendix, is based on the series expansion of the bivariate normal pdf using the Hermite polynomials.

(c) Let X and Z be independent random variables and let $Y = \max(X, Z/c)$, $c > 0$. The influence of X on Y gets larger as c increases. It is desirable that this is reflected by $\mathrm{CID}(Y|X)$, which turns out to be the case. It is derived in the Appendix that

$$\mathrm{CID}(Y|X) = 6\int_0^1 v^2\left(\frac{1}{F_X(F_Y^{-1}(v))} - 1\right)\mathrm{d}v. \tag{6}$$

Note that since $F_Y(y) = F_X(y)F_Z(cy)$ is an increasing function of c for each fixed y, $F_Y^{-1}(v)$ is a decreasing function of c for each fixed v. Hence $\mathrm{CID}(Y|X)$ is monotone increasing in c.

We next address systematically whether the basic criteria (C1)–(C6) described in Section 1 are satisfied for the CID.

(C1) This criterion is trivially satisfied since there is no parametric assumption in the definition. Remark (b) above also gives additional support for this.

(C2) By the variance decomposition

$$\mathrm{Var}(U) = \mathrm{E}(\mathrm{Var}(U|\mathscr{H})) + \mathrm{Var}(\mathrm{E}(U|\mathscr{H})), \tag{7}$$

it is not difficult to see that for $\mathscr{H}_1 \subset \mathscr{H}_2 \subset \mathscr{F}$,

$$0 = \mathrm{CID}(Y|\mathscr{H}_0) \leqslant \mathrm{CID}(Y|\mathscr{H}_1) \leqslant \mathrm{CID}(Y|\mathscr{H}_2) \leqslant \mathrm{CID}(Y|\mathscr{F}) = 1, \tag{8}$$

where $\mathscr{H}_0$ is the trivial σ-field $\{\emptyset, \Omega\}$. Hence $\mathrm{CID}(Y|\mathscr{H}) \in [0, 1]$ and is monotone in $\mathscr{H}$. Also observe that $\mathrm{CID}(Y|\mathscr{H}) = 0/1$ iff Y and $\mathscr{H}$ are independent/completely dependent. In addition to the theoretical computations in Example 2.1 above, we will illustrate with a number of simulation results that $\mathrm{CID}(Y|X)$ increases if the model of X, Y changes in such a way that X asserts a larger influence on Y functionally or stochastically.

(C3) The CID is applicable to both continuous and discrete data. Remark (b) makes it clear that the "independent" variable X is in essence qualitative. A slightly more subtle point is that if Y is binary then the specific values of Y also do not have any relevance on the CID computation, making Y a categorical variable.

(C4) It is clear from the definition that $\mathrm{CID}(Y|X)$ is not in general symmetric in X and Y.

(C5) Estimation of CID is a key issue and will be pursued in Sections 3 and 4. We will demonstrate that CID can be estimated quite efficiently with moderate to large sample sizes. Although one cannot expect to estimate CID well for small samples, even for small samples one may still be able to use the estimates to differentiate levels of dependence. Example 4.3 below gives evidence for this.

(C6) CID can be defined when either or both of X, Y are multivariate; when Y is multivariate, replace the delta function and the cdf of Y by their multivariate versions.

3. Estimation

To estimate $\mathrm{CID}(Y|X_1, \ldots, X_k)$ from data, for simplicity of notation assume that $k = 1$ and we observe the data (X_i, Y_i), $1 \leqslant i \leqslant n$. If $k > 1$, then replace the univariate cdf's by multivariate ones. It will be convenient to work with Eq. (3). First, we estimate $F_Y(Y_i)$ by

$$\frac{1}{n}\sum_{k=1}^{n} I(Y_k \leqslant Y_i) = R_{Y,i} = \frac{1}{n} \times \text{rank of } Y_i \text{ in } Y_1, \ldots, Y_n. \tag{9}$$

The estimation of the denominator of CID is easy. In general, writing $H(v) = \frac{1}{n}\sum_{i=1}^{n}\delta_{R_{Y,i}}(v)$, we can estimate

the denominator by

$$\int_{u=0}^{1}\left[\int_{v=0}^{1}\delta_v(u)^2\,\mathrm{d}H(v)-\left(\int_{v=0}^{1}\delta_v(u)\,\mathrm{d}H(v)\right)^2\right]\mathrm{d}u$$
$$=\int_{u=0}^{1}\left[(1-H(u))-(1-H(u))^2\right]\mathrm{d}u$$
$$=\int_{u=0}^{1}H(u)(1-H(u))\,\mathrm{d}u. \tag{10}$$

If all of the Y's are distinct, then

$$\int_{u=0}^{1}H(u)(1-H(u))\,\mathrm{d}u=\frac{1}{n}\sum_{i=1}^{n-1}\frac{i}{n}\left(1-\frac{i}{n}\right)$$
$$\approx\frac{1}{6}\quad\text{for large } n. \tag{11}$$

Indeed, if Y has a continuous distribution, then $\tilde{Y}$ is distributed as uniform over [0, 1]. Hence,

$$\int_{u=0}^{1}\mathrm{Var}(\delta_{\tilde{Y}}(u))\,\mathrm{d}u$$
$$=\int_{u=0}^{1}\left[\int_{v=0}^{1}\delta_v(u)^2\,\mathrm{d}v-\left(\int_{v=0}^{1}\delta_v(u)\,\mathrm{d}v\right)^2\right]\mathrm{d}u$$
$$=\int_{u=0}^{1}\left[(1-u)-(1-u)^2\right]\mathrm{d}u=1/6. \tag{12}$$

On the other hand, if Y is discrete then the ranks of the Y_i will contain ties and the computations here will have to be modified.

The numerator of CID is more complicated. Letting A_j, $1\leqslant j\leqslant m$, be a partition of the real line and c_j be the number of $X_i\in A_j$, we estimate the numerator by

$$\int_0^1\sum_{j=1}^{m}\frac{c_j}{n}\left(\frac{1}{c_j}\sum_{X_i\in A_j}\delta_{R_{Y,i}}(u)-\frac{1}{n}\sum_{i=1}^{n}\delta_{R_{Y,i}}(u)\right)^2\mathrm{d}u. \tag{13}$$

The choice of A_j is clearly a delicate issue; a good choice of A_j balances bias and variance. Write the above as

$$\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n}I(X_i\in A_j)\int_0^1$$
$$\times\left(\frac{(1/n)\sum_{i=1}^{n}I(X_i\in A_j,R_{Y,i}\leqslant u)}{(1/n)\sum_{i=1}^{n}I(X_i\in A_j)}\right.$$
$$\left.-\frac{1}{n}\sum_{i=1}^{n}I(R_{Y,i}\leqslant u)\right)^2\mathrm{d}u. \tag{14}$$

Let us assume for convenience that

$$\sum_{i=1}^{n}I(X_i\in A_j)=\frac{n}{m}; \tag{15}$$

which means we divide the data into m bins each containing the same number of X_i's. Then the above expression becomes

$$m\sum_{j=1}^{m}\sum_{i=1}^{n}\int_0^1\left(\frac{1}{n}\sum_{i=1}^{n}I(X_i\in A_j,R_{Y,i}\leqslant u)\right.$$
$$\left.-\frac{1}{mn}\sum_{i=1}^{n}I(R_{Y,i}\leqslant u)\right)^2\mathrm{d}u. \tag{16}$$

This formula makes it possible to compute the variance and bias of the estimator and hence provides guidelines on the best choice of the number of bins. However, the computations are clearly challenging technically. In order to not diffuse the focus of this paper, we will tackle the computations as well as large sample properties in a future paper.

4. Simulation results and comparisons

In this section we present some examples of CID for a number of models. The first example demonstrates the manner in which the value of CID reflects the degree of dependence for various models and compares the results with the correlation, Spearman's ρ and Kendall's τ. The second example explores the performance of the CID estimator for a model with different choices of bin and sample sizes. The third example shows that an off-the-target estimate of CID could still be useful in distinguishing different strengths of dependence.

Example 4.1. We consider 4 models:

- *Model* 1: $Y=X^2+c^{-1}Z$, where X and Z are independent standard normal and $c\in[1,10]$.
- *Model* 2: $Y=X+c^{-1}Z$, where X and Z are independent standard normal and $c\in[1,10]$.
- *Model* 3: The joint cdf of $U=F_X(X)$, $V=F_Y(Y)$ is

$$C(u,v)=\exp(-[(-\ln u)^c+(-\ln v)^c]^{1/c}), \tag{17}$$

  with $c\in[1,10]$ and F_X, F_Y both equal to the standard normal. $C(u,v)$ is known as the Gumbel copula (cf. Ref. [3]).
- *Model* 4: $Y=|I(X+W>0)-I(X+Z>c)|$ where W, X, Z are independent standard normal and $c\in[1,10]$.

In all of these models, the strengths of dependence between X and Y increase with c. For Model 1 we obtained the CID curve; for Models 2 and 3 we obtained the curves for correlation, Spearman's ρ, Kendall's τ and CID; and for Model 4 the curves for $\mathrm{CID}(Y|W)$, $\mathrm{CID}(Y|X)$ and $\mathrm{CID}(Y|W,X)$ are obtained. The curves are computed by performing 10 000 simulations where $\mathrm{CID}(Y|X)$ in Models 1–3 is estimated using 20 bins for X and $\mathrm{CID}(Y|W,X)$

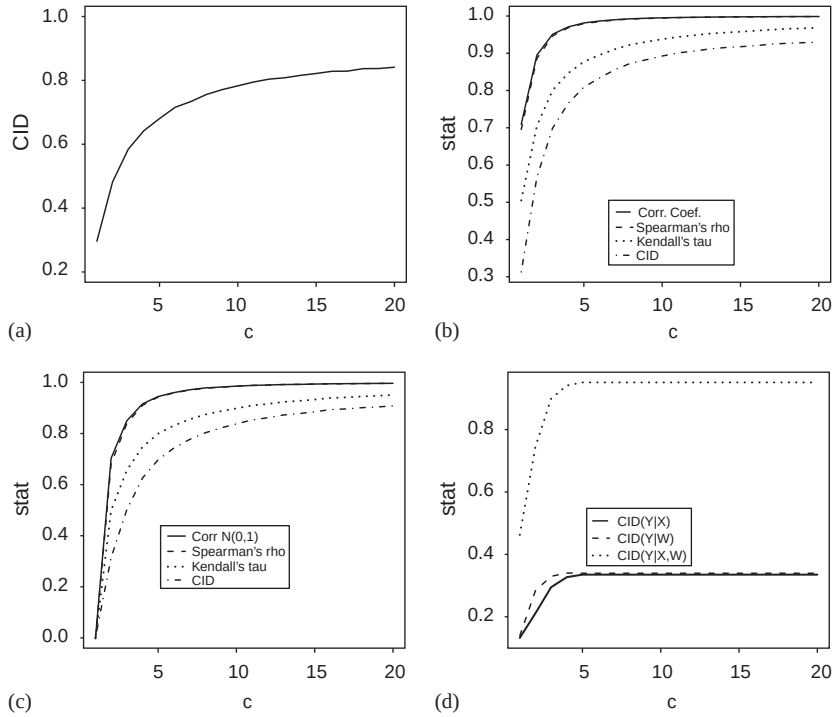

Fig. 1. The CID results are compared with various dependence measures based on simulated data, with (a)–(d) corresponding to models 1–4, respectively.

in Model 4 is estimated using 20 bins for each of the predictor variables. Recall that

- Spearman's ρ between X and $Y = \mathrm{Corr}(\tilde{X}, \tilde{Y})$,
- Kendall's τ between X and $Y = \mathrm{E}(\mathrm{sign}[(X_1 - X_2)(Y_1 - Y_2)])$, where $(X_i, Y_i), i = 1, 2$, are iid.

The results are summarized in Fig. 1. For Model 1, since correlation $=$ Spearman's $\rho =$ Kendall's $\tau = 0$ by symmetry, only CID was able to detect and differentiate different levels of dependence. For Models 2 and 3, all of the measures apparently were able to differentiate different degrees of dependence; however, the CID curves ascend most gradually, making it a superior measure of dependence in these examples. Note that all of the curves in Models 1–3 eventually reach 1 as c increases. In Model 4, $\mathrm{CID}(Y|W, X)$ is larger than $\mathrm{CID}(Y|W)$ and $\mathrm{CID}(Y|X)$.

Example 4.2. In this example, we consider the effect of bin size on CID estimation for different sample sizes. Consider the following model: $Y = X + X^2 + 0.5Z$, where X and Z are independent standard normal and $c \in [1, 10]$. Figs. 2–6 summarize the results of $\mathrm{CID}(Y|X)$ estimation together with the MSE, variance and bias for different combinations of bin and sample sizes based on 20 simulations. The set of plots are for sample sizes 1–200. One can see that the CID estimator works very well for sample sizes as small

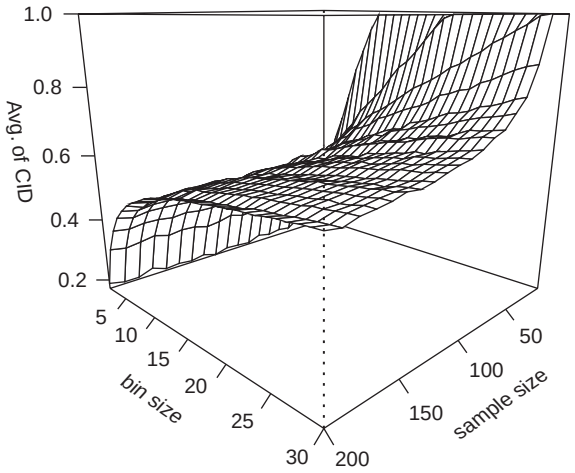

Fig. 2. CID values for different combination of bin and sample sizes from one perspective.

as 50 and a wide range of bin numbers. Based on additional simulations on a variety of models, we are comfortable that these conclusions hold quite generally so long as both Y and X are one-dimensional. If X is multi-dimensional, then larger samples are required.

Example 4.3. Even for a modest number of predictors (conditioning variables), the CID cannot be precisely estimated

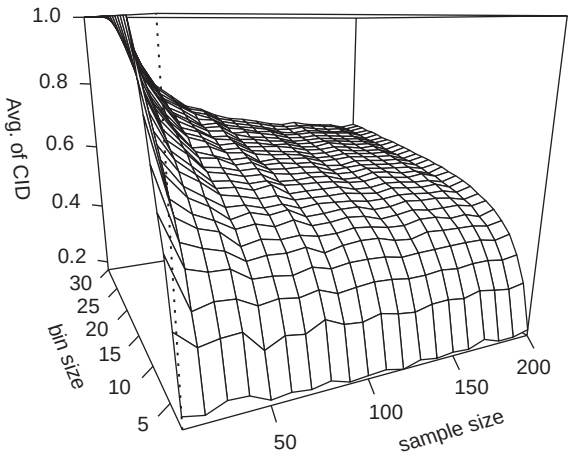

Fig. 3. CID values for different combination of bin and sample sizes from another perspective.

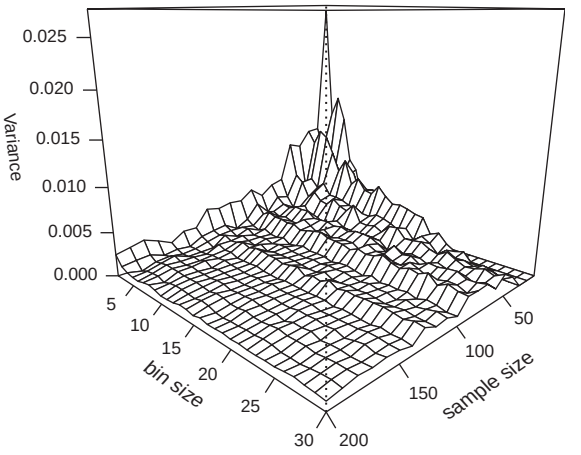

Fig. 5. Variance values for different combination of bin and sample sizes.

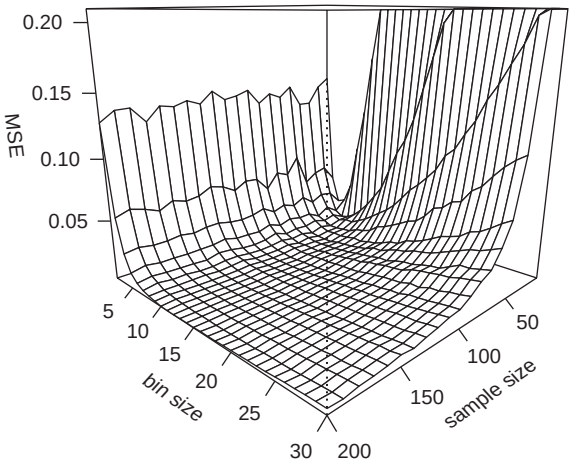

Fig. 4. MSE values for different combination of bin and sample sizes.

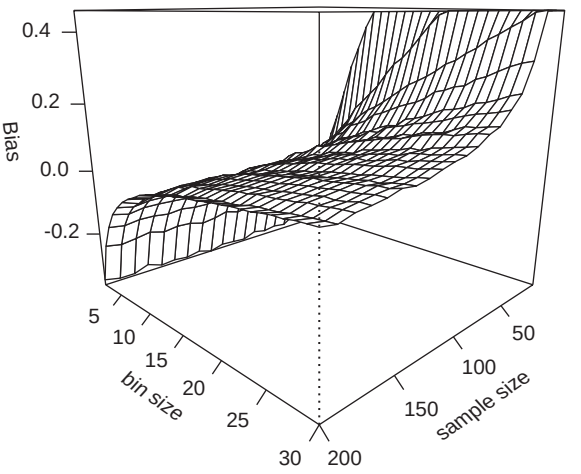

Fig. 6. Bias values for different combination of bin and sample sizes.

with a small sample. In practice, however, estimating the true CID is often not the goal; rather, the goal is to compare the strengths of dependence. Consider the model $Y = I(cX * W + (1-c)W * Z)$ where W, X, Z are independent standard normal and $c \in [0, 1]$. Clearly,

$$\mathrm{CID}(Y|X, W) > \mathrm{CID}(Y|W, Z) \quad \text{iff } c > 1/2 \tag{18}$$

and

$$\mathrm{CID}(Y|X, Z) < \max(\mathrm{CID}(Y|X, W), \mathrm{CID}(Y|W, Z)). \tag{19}$$

Suppose we have a limited number of observations with which to estimate the CID's of Y on two variables. We want to illustrate that, with high probability, we may still be able to detect which of the three pairs (X, Z), (X, W), (W, Z) has the largest influence on Y based on the estimated CID's, even though the estimates are poor. For instance, if X', W' and Z' are discretized versions of X, W, and Z, respectively, then $\mathrm{CID}(Y|X', W') > \mathrm{CID}(Y|W', Z')$ iff $c > 1/2$, and even though we cannot estimate $\mathrm{CID}(Y|X, Z)$ and $\mathrm{CID}(Y|W, Z)$ well with a small sample, it may be possible to estimate $\mathrm{CID}(Y|X', Z')$ and $\mathrm{CID}(Y|W', Z')$ well. The left graph in Fig. 7 contains the estimated CID profile curves based on a particular simulated sample of 100 observations using 3 bins. We expect them to be quite different from the true curves. On the right in Fig. 7 is the estimated probabilities of correct identification of which pair of (X, W), (W, Z) and (X, Z) has the largest influence on Y based on the estimated CID with 100 observations and 3 bins. The probabilities are obtained by performing 1000 simulations. Observe that the

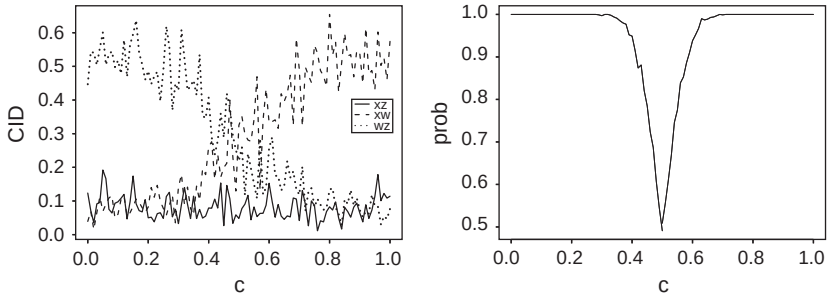

Fig. 7. The plot on the left is the CID of Y given each set of two predictors for a particular sample of size 100. The bin size is set to be 3. CID$(Y|W, Z)$ and CID$(Y|X, W)$ are expected to stand out when c is less than 0.5 and greater than 0.5, respectively. The plot on the right describes the probability that the estimated CID correctly identifies the predictor set which most strongly influences Y based on 1000 simulations.

probability of correct identification is high provided that c is not close to 0.5.

5. Application to genomics

Microarray technology facilitates large-scale surveys of gene expression in which transcript levels can be determined for thousands of genes simultaneously [4]. Since transcription control is accomplished by a method that interprets a variety of inputs, analytical tools are required for expression profile data that can detect the types of multivariate influences on decision-making produced by complex genetic networks. Application is generally directed towards tissue classification and the discovery of signaling pathways, both based on the expressed macromolecule phenotype of the cell. Because transcriptional control is accomplished by a complex method that interprets a variety of inputs [5–7], the development of analytical tools that detect multivariate influences on decision-making present in complex genetic networks is essential. Here we use the CID in two contexts:

(a) measurement of multivariate relationships among genes via prediction;
(b) expression-based classification.

We use a gene-expression data set taken from a study involving 295 breast-cancer patients, of whom 115 had a good prognosis signature and 180 had a poor prognosis signature [8]. The purpose of the study was to demonstrate the ability of expression data to predict the outcome of the disease. The expression levels of approximately 25 000 genes were determined for each patient. About 5000 significantly regulated genes were selected from the 25 000 genes on the microarray. The correlation coefficient of the expression for each gene with disease outcome was calculated and 231 genes were found to be significantly associated with disease outcome (correlation coefficient < -0.3 or > 0.3). These 231 genes were then rank-ordered on the basis of the magnitude of the correlation coefficient. The top 70 genes in the rank-ordered list were then used to classify prognosis [8,9]. We have access only to the gene-expression data of the 70 genes and therefore we will limit our analysis to these.

5.1. Multivariate prediction

We consider the problem of predicting the expression level of a target gene via the levels of 2-predictor genes. Specifically, we want to discover genes that can serve as predictors for a target gene, which is a form of feature selection. This problem has received attention in the genomics literature, with feature selection depending on predictor design, a wrapper approach. One method has been to discover associations between the expression patterns of genes via the coefficient of determination, or CoD [10–12]. The CoD measures the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the transcriptional state of a target gene relative to the best possible prediction in the absence of observations. A related approach treats the problem of finding inter-gene relations via prediction as a model-selection problem using the principle of minimum description length (MDL) [13]. Here we employ the CID to measure multivariate relationships among genes and demonstrate its performance as a feature-selection filter. Whereas both the CoD and MDL methods have been applied using coarsely quantized data, either binary or ternary, we apply CID with continuous data.

Each of 70 genes is used as a target gene and the CID computed for all possible two-gene predictor combinations from among the remaining 69 genes using the log ratio values. For each target gene we have computed the mean CID over the 2346 two-gene predictor sets and found the top six targets to be the following (with the mean CID in parenthesis): CENPA (0.3232), KIAA0175 (0.3201), PRC1 (0.2957), ORC6L (0.2936), Sequence 12 (0.2904), LOC51203 (0.2900). We have selected gene CENPA as the target and compared the CID for the 2346 predictor sets with the mean-square error (MSE) estimated for a 2-layer

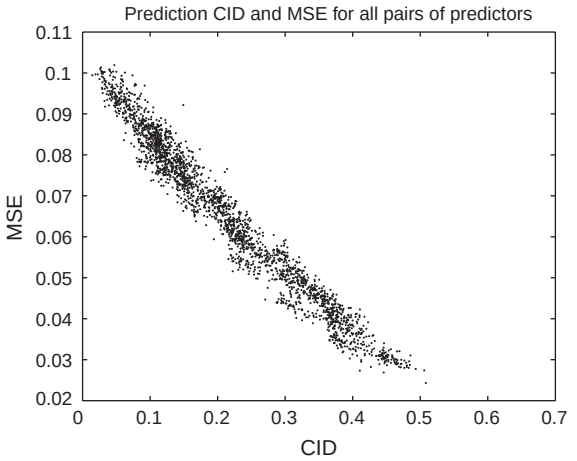

Fig. 8. The scatter plot between the CID and the MSE estimated for a neural-network predictor of target gene CENPA for all predictor sets.

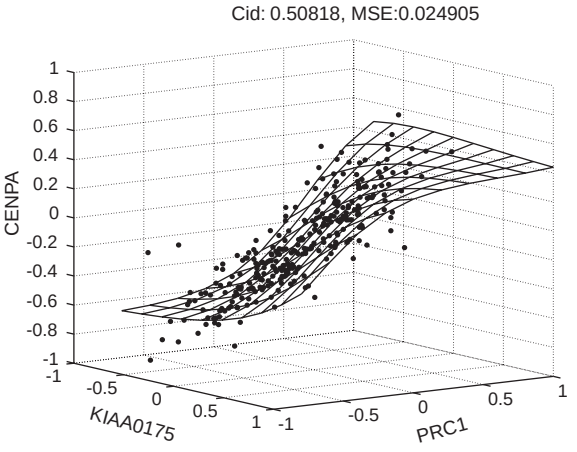

Fig. 9. Error plot obtained by the neural network predicting gene CENPA via the best set of two predictors, genes KIAA0175 and PRC1.

neural-network predictor having three neurons in the inner layer. A small number of neurons is used to avoid overfitting. All 295 data points are used for design. The MSE is estimated using resubstitution, which is close to unbiased with very small variance for 295 training examples.

Fig. 8 shows the scatter plot between the CID and the MSE computed for target gene CENPA for all 2346 predictor sets. There is a fairly tight linear relationship. For the CID, the best predictor set for gene CENPA is the predictor set consisting of gene KIAA0175 and gene PRC1, with CID 0.5082. This set also has the lowest MSE (0.0246) among all predictor sets. In terms of feature selection, choosing the feature set with maximum CID yields the lowest MSE among all feature sets. The plot in Fig. 9 shows the surface

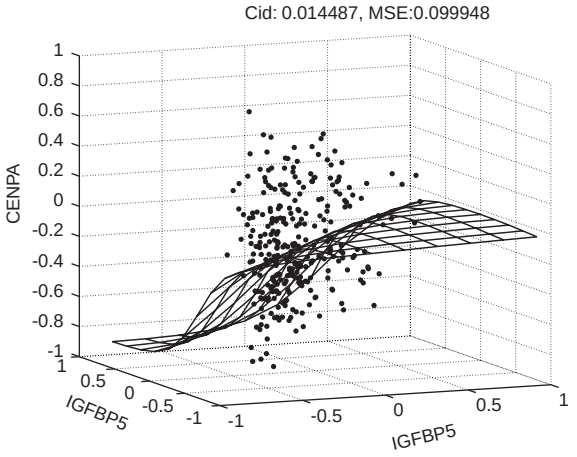

Fig. 10. Error plot obtained by the neural network predicting gene CENPA via the worst set of two predictors, genes IGFBP5a and IGFBP5b.

obtained by the neural network predicting gene CENPA via genes KIAA0175 and PRC1. There is a good fit with the data points. In contrast, the plot on Fig. 10 shows the surface resulting from the set with the lowest CID (0.0145). This set consists of genes IGFBP5a and IGFBP5b, and the MSE for the neural network is 0.0993.

5.2. Classification

An expression-based classifier provides a list of genes whose product abundance is indicative of important differences in cell state, such as healthy or diseased, or one particular type of cancer or another. Two central goals of molecular analysis of disease are to use such information to directly diagnose the presence or type of disease and to produce therapies based on the disruption or correction of the aberrant function of gene products whose activities are central to the pathology of a disease. Correction would be accomplished either by the use of drugs already known to act on these gene products or by developing new drugs targeting these gene products. Achieving these goals requires designing a classifier that takes a vector of gene-expression levels as input and outputs a class label that predicts the class containing the input vector. Classification can be between different kinds of cancer, different stages of tumor development, or many other such differences. A variety of methods has been used to exploit the class-separating power of expression data in cancer, for instance, leukemias [14], small, round, blue-cell cancers [15], hereditary breast cancer [16], colon cancer [17], breast cancer [18], melanoma [19], and glioma [20]. To illustrate use of the CID in the context of expression-based classification, we again use the 295-patient breast-cancer study, this time employing the signature associated with good and bad prognosis [8]. Specifically, the 295 sample points are split into two classes, one associated

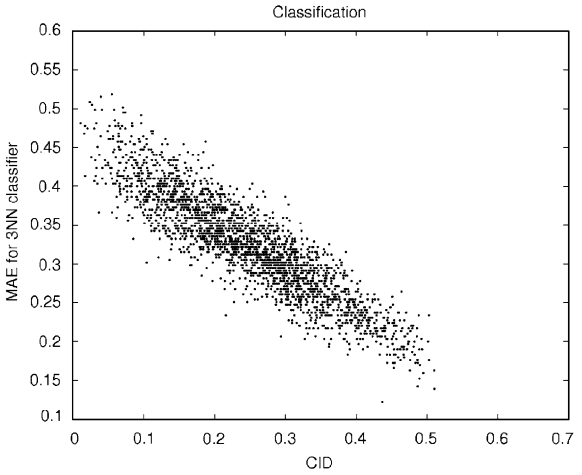

Fig. 11. The scatter plot between the CID and the MAE from 3-nearest-neighbor classification for all two-gene classifiers.

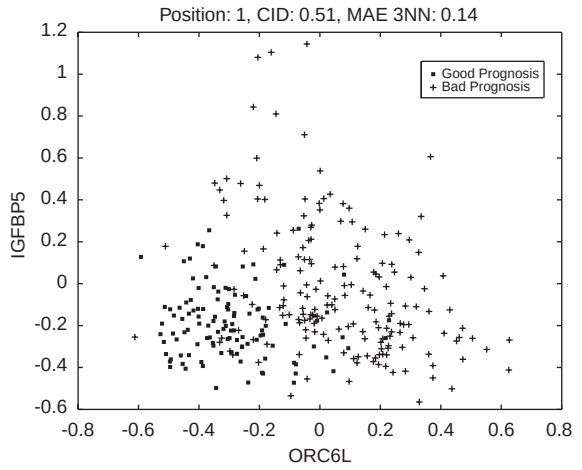

Fig. 12. Expression values of the best (genes ORC6L and IGFBP5) set of predictors for the 295 patients and their associated prognoses.

with a good prognosis and the other with a bad prognosis. Here we compute the CID for all pairs of genes as predictors of the prognosis signature. We use 3-nearest-neighbor classification to analyze the ability of each pair of genes to classify an example between good and bad prognosis, and use leave-one-out estimation to estimate the mean-absolute error (MAE) of a classifier. To be in accord with error measurements used in the related literature, we have used MSE for prediction and now use MAE for classification.

Fig. 11 shows the scatter plot between the CID and the MAE for all two-gene classifiers. The two-gene set possessing the highest CID (0.5104) for prognosis classification consists of genes ORC6L and IGFBP5. Its classification MAE is 0.1390. This is the second lowest MAE among all two-gene classifiers. As in the case of prediction, CID has selected an excellent feature set. The plot on Fig. 12 shows

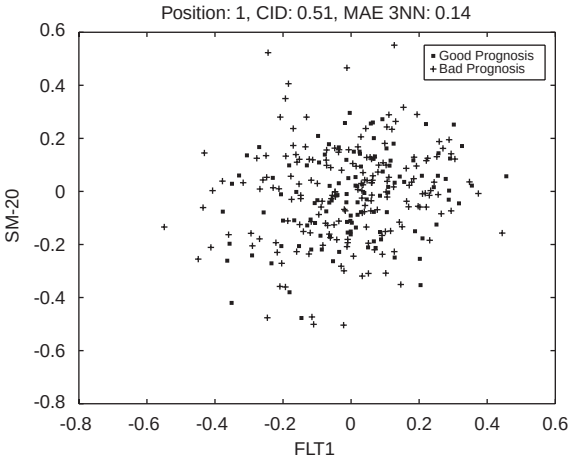

Fig. 13. Expression values of the worst (genes FTL1 and SM-20) set of predictors for the 295 patients and their associated prognoses.

the values of genes ORC6L and IGFBP5 for the 295 patients and their associated prognoses. The lowest CID, 0.0107, is attained by the set consisting of genes FTL1 and SM-20, and it has the worst classification MAE, 0.48136. The plot for these genes is shown in Fig. 13.

6. Feature selection

We now apply the CID in filter-style feature selection. As noted in the Introduction, when there is a very large set of potential features, it is common to apply filter-style feature selection to pre-filter the features down to a manageable set upon which to apply a wrapper algorithm. Since our intent here is to demonstrate the efficacy of the CID, and since we want to both avoid error estimation and to compare the CID relative to ground truth, we will assume a modest number of features and progressively filter the features by both correlation and the CID. We assume the distribution to be known and will compute the Bayes error for each feature set. The result will be two decreasing error curves, one for correlation filtering and the other for CID filtering. So as to remain in the domain of genomics and prediction, we will consider gene prediction in the framework of genetic regulatory networks. Using a probabilistic Boolean network (PBN) for the regulatory network model [21], we obtain the steady-state distribution of its corresponding Markov chain by running the network a large number of times [22] and consider multivariate gene prediction in the steady-state distribution. Since we have in hand the steady-state distribution, we can compute actual prediction errors. Prediction in the Boolean context means deciding whether a particular target gene is expressed, an event denoted by a binary variable Y. Available are the binary gene expressions, $X_1, X_2, \ldots, X_M$, of M other genes, and feature sets composed of these are to

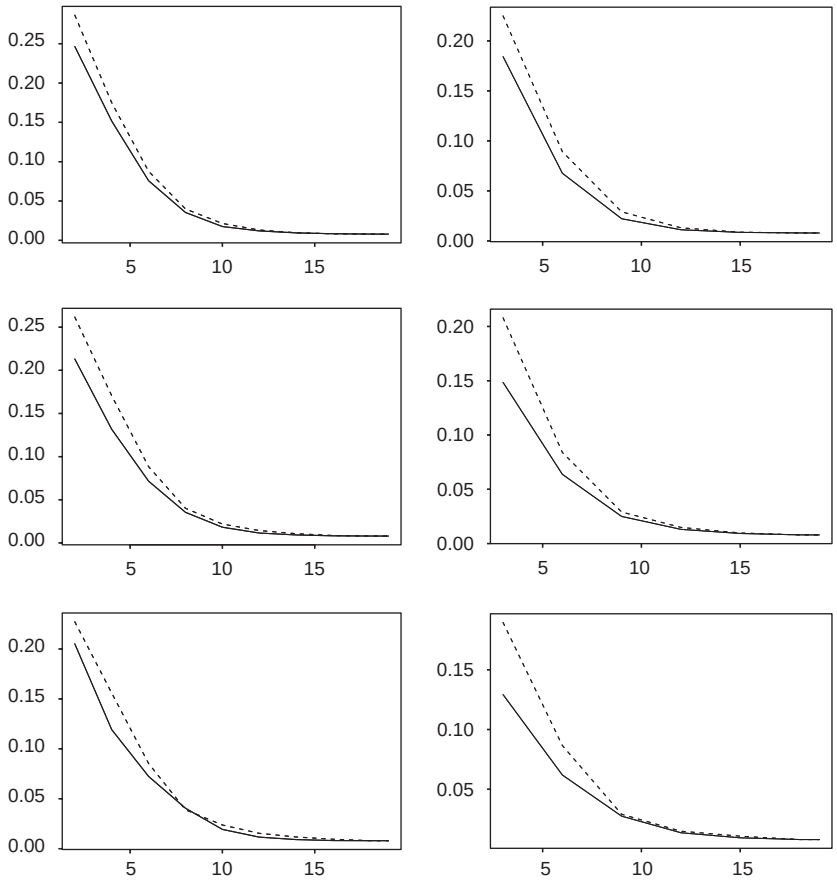

Fig. 14. Bayes errors for feature selection based on CID (solid lines) and on correlation (dashed lines). The *x*-axis indicates the number of genes being selected; the *y*-axis denotes the Bayes errors. The genes are taken 2 at a time on the left panel and 3 on the right. From top to bottom, the sample size are 30, 100, and 300, respectively.

be selected. The approach of using PBN's provides a natural means for examining prediction issues via simulated data whose distributional properties reflect expression data derived from genetic regulatory networks [23]. We present the feature-selection results in this section and defer to the Appendix for a description of the PBN model.

After estimating the steady-state distribution of a PBN consisting of 20 genes, we have chosen the one having the largest error when predicted by its own mean to be the target. This insures that good prediction does not simply result from the gene possessing little variation in the steady state.We randomly select 100 samples of sizes $n = 30$, 100, and 300 from the steady-state distribution (the results for $n = 500$ being omitted since they are essentially the same as for $n = 300$). The CID values of the target gene given all combinations of $k = 2$ and 3 variables are computed, along with the correlation between the target and each variable. We do not consider $k = 1$ for the CID because in that case the CID is equal to the squared correlation. The best genes based on the values of the CID and on the correlation are selected. Since the CID considers more than a single gene at a time, a protocol must be defined to make the CID list. For $k = 2$, the genes occur in pairs when the best 2-predictor CID values are taken. For $k = 3$, they occur in triples when the best 3-predictor CID values are taken. In each case we list in descending order the sets (pairs or triples) and make the CID gene list by taking the genes as they appear in the list. The results are given in Fig. 14, where the x and y axes give the number of genes and the Bayes error, respectively. The genes are taken 2 and 3 at a time from the $k = 2$ and 3 CID lists, respectively, and similarly from the correlation list. Besides the fact that the CID provides better feature selection than the correlation, two salient points should be noted. First, the CID performance advantage is better for $k = 3$ than $k = 2$, reflecting the fact that $k = 3$ takes more multivariate effects into account. Second, the performance advantage for the CID for $k = 3$ increases for increasing n, reflecting improved estimation of the CID with larger samples. Finally, note that even for samples as small as $n = 30$, where CID estimation is substantially poorer than

correlation estimation, the CID's ability to measure multivariate nonlinear interaction still provides better feature selection.

7. Conclusion

This paper has introduced a new measure of dependence between random variables that satisfies a number of desirable properties for such measurements. It does so by comparing the conditional distribution of Y given X with the marginal distribution of Y. Basic issues for application to feature selection have been investigated. The measure has been applied to gene prediction and tissue classification based on continuous microarray data.

Because microarray-based genomic studies constitute a motivating factor for the development of the CID, before closing, some comments on the applicability of the CID to such studies are warranted, in particular, regarding the issue of sample size. As we have noted, estimation of the CID cannot be accomplished with samples that are small relative to the number of variables. The database used in our example is large compared to those of most microarrays studies, where small-sample estimation is a common impediment [24]. Two matters point to the beneficial use of the CID for microarray-based quantification of gene interaction. First, good prediction (owing to regulation) will most often be achieved with a small number of genes; moreover, an important consideration is that the number of predictor or classifier genes should be sufficiently small to be potentially useful as candidates for functional analysis to determine whether they can serve as useful targets for therapy. Second, regarding sample size, as the technology becomes more cost effective and placed into the service of large-scale operations, larger studies become feasible, and are indeed underway. For instance, the Project for Oncology of the International Genomic Consortium aims at integrating longitudinal clinical annotation with gene-expression data to develop diagnostic markers, prognostic indicators, and therapeutic targets. A current 3-year project is in the process of creating a database of the gene-expression profiles of 2500 human tumor specimens and 500 normal tissues collected under standardized conditions, clinically annotated, and de-identified for public access. It is the advent of such large-scale expression databases that has motivated the development of the CID.

Appendix A

A.1. Some derivations of CID

We first derive the CID of the bivariate normal distribution with correlation ρ. Denote by ϕ_ρ be the pdf of the bivariate normal with correlation ρ and standardized marginals, and ϕ the standard normal pdf. By Fourier inversion,

$$\begin{aligned}
&\phi_\rho(u,v)\\
&\quad=\frac{1}{4\pi^2}\int\int \exp\left\{\mathrm{i}tu+\mathrm{i}sv-\frac{1}{2}[t^2+s^2+2ts\rho]\right\}\mathrm{d}t\,\mathrm{d}s\\
&\quad=\sum_{k=0}^{\infty}\frac{(-\rho)^k}{k!}\frac{1}{4\pi^2}\int\int s^k t^k\,\exp\left\{\mathrm{i}tu+\mathrm{i}sv\right.\\
&\qquad\left.-\frac{1}{2}[t^2+s^2]\right\}\mathrm{d}t\,\mathrm{d}s\\
&\quad=\sum_{k=0}^{\infty}\frac{\rho^k}{k!}\frac{\mathrm{d}^k}{\mathrm{d}u^k}\left[\frac{1}{2\pi}\int \mathrm{e}^{\mathrm{i}tu-(1/2)t^2}\,\mathrm{d}t\right]\frac{\mathrm{d}^k}{\mathrm{d}v^k}\\
&\qquad\times\left[\frac{1}{2\pi}\int \mathrm{e}^{\mathrm{i}sv-(1/2)s^2}\,\mathrm{d}s\right]\\
&\quad=\sum_{k=0}^{\infty}\frac{\rho^k}{k!}H_k(u)\phi(u)H_k(v)\phi(v),
\end{aligned}\tag{A.1}$$

where

$$H_k(u)=(-1)^k\mathrm{e}^{u^2/2}\frac{\mathrm{d}^k}{\mathrm{d}u^k}\mathrm{e}^{-u^2/2}\tag{A.2}$$

is the kth Hermite polynomial. Since $H_0(x)=1$, we have

$$\begin{aligned}
&P(Y\leqslant u|X=x)-P(Y\leqslant u)\\
&\quad=\frac{1}{\phi(x)}\int_{-\infty}^{u}[\phi_\rho(w,x)-\phi(w)\phi(x)]\,\mathrm{d}w\\
&\quad=\frac{1}{\phi(x)}\int_{-\infty}^{u}\sum_{k=1}^{\infty}\frac{\rho^k}{k!}H_k(w)H_k(x)\phi(w)\phi(x)\,\mathrm{d}w\\
&\quad=\sum_{k=1}^{\infty}\frac{\rho^k}{k!}H_k(x)\int_{-\infty}^{u}H_k(w)\phi(w)\,\mathrm{d}w,
\end{aligned}\tag{A.3}$$

where, by (A.2),

$$\int_{-\infty}^{u}H_k(w)\phi(w)\,\mathrm{d}w=-\phi(u)H_{k-1}(u).\tag{A.4}$$

Since the Hermite polynomials are orthogonal with respect the normal distribution, we have

$$\begin{aligned}
&E[P(Y\leqslant u|X=x)-P(Y\leqslant u)]^2\\
&\quad=\sum_{k=1}^{\infty}\frac{\rho^{2k}}{k!}\big(\phi(u)H_{k-1}(u)\big)^2.
\end{aligned}\tag{A.5}$$

Hence the numerator of $\mathrm{CID}(Y|X)$ is

$$\sum_{k=1}^{\infty}\frac{\rho^{2k}}{k!}\int_{-\infty}^{\infty}\big(\phi(u)H_{k-1}(u)\big)^2\phi(u)\,\mathrm{d}u,\tag{A.6}$$

where, by (A.2),

$$\big(\phi(u)H_{k-1}(u)\big)^2\phi(u)=\left[\phi^{(k-1)}(u)\right]^2\phi(u).\tag{A.7}$$

The denominator of CID is 1/6 as explained in Section 3. This completes the derivation.

We next derive $\mathrm{CID}(Y|X)$ where $Y=\max(X, Z/c)$, $c>0$, with X, Z being independent. Observe that

$$P(Y\leqslant u) = P(\max(X, Z/c)\leqslant u) = F_X(u)F_Z(cu) \quad (A.8)$$

and also that

$$P(Y\leqslant u|X=x) = P(\max(x, Z/c)\leqslant u) = 0\cdot I(x>u) + F_Z(cu)I(x\leqslant u). \quad (A.9)$$

Hence

$$P(Y\leqslant u|X=x) - P(Y\leqslant u) = F_Z(cu)[I(x\leqslant u) - F_X(u)], \quad (A.10)$$

$$\begin{aligned}\int_{-\infty}^{\infty} & E[P(Y\leqslant u|X=x) - P(Y\leqslant u)]^2\,\mathrm{d}F_Y(u)\\ &= \int_{-\infty}^{\infty} F_Z^2(cu)[F_X(u) - F_X^2(u)]\,\mathrm{d}F_Y(u)\\ &= \int_{-\infty}^{\infty} F_Y^2(u)\left(\frac{1}{F_X(u)} - 1\right)\mathrm{d}F_Y(u) \quad (A.11)\end{aligned}$$

by (A.8). Making a variable change with $v = F_Y(u)$ in the last integral, we have

$$\begin{aligned}\int_{-\infty}^{\infty} & E[P(Y\leqslant u|X=x) - P(Y\leqslant u)]^2\,\mathrm{d}F_Y(u)\\ &= \int_0^1 v^2\left(\frac{1}{F_X(F_Y^{-1}(v))} - 1\right)\mathrm{d}v. \quad (A.12)\end{aligned}$$

Hence

$$\mathrm{CID}(Y|X) = 6\int_0^1 v^2\left(\frac{1}{F_X(F_Y^{-1}(v))} - 1\right)\mathrm{d}v. \quad (A.13)$$

Appendix B

B.1. Probabilistic Boolean networks

A probabilistic Boolean network (PBN) is defined by a set of binary-valued nodes $\{X_1, X_2, \ldots, X_n\}$ and a list $F=\{F_1, F_2, \ldots, F_n\}$ of sets $F_i=\{f_1^{(i)}, f_2^{(i)}, \ldots, f_{l(i)}^{(i)}\}$ of Boolean functions. Each node $X_i \in \{0, 1\}$ represents the state (expression) of gene i, where $X_i = 1$ means that gene i is expressed and $X_i = 0$ means that it is not expressed. The set F_i contains the possible rules of regulatory interactions for gene i. For $j = 1, 2, \ldots, l(i)$, $f_j^{(i)}$ is a possible Boolean function determining the value of x_i in terms of some other gene states. The functions are called *predictors*. All genes (nodes) are updated synchronously and repeatedly in accordance with the functions assigned to them. A realization of a PBN at a given time is determined by a vector $\mathbf{f}$ of Boolean functions. If there are N possible realizations, then there are N vector functions $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$ of the form $\mathbf{f}_k=(f_{k_1}^{(1)}, f_{k_2}^{(2)}, \ldots, f_{k_n}^{(n)})$, for $k=1, 2, \ldots, N$, $1\leqslant k_i\leqslant l(i)$, and $f_{k_i}^{(i)} \in F_i$ $(i = 1, 2, \ldots, n)$. The *network function* $\mathbf{f}_k$ acts as a transition mapping representing a possible realization of the entire PBN. Each predictor function usually has many fictitious variables, which means there are only a few input genes that actually regulate X_i at any given time. At each time point, the expression vector $\mathbf{X}=(X_1, X_2, \ldots, X_n)$ is called a *gene activity pattern*.

Not only must the PBN transition be between gene activity patterns, it must also transit between network functions. A binary random variable Γ, with $\mathrm{pr}(\Gamma = 1) = \alpha$, governs whether or not there is a change of network function at each time instance. There is a network-function change if and only if $\Gamma = 1$. Γ is independent of the state of the network. Given a network change $(\Gamma = 1)$, there are selection probabilities $c_1, c_2, \ldots, c_N$ determining which of the network functions $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$ will govern the network until the next switch. The PBN model also allows for random perturbations. For each gene, there is a small probability β that it will flip its value, from 0 to 1 or from 1 to 0. Hence, there is a binary random variable Θ, independent of the network state and Γ, with $\mathrm{pr}(\Theta=0)=(1-\beta)^n$, such that when $\Theta=0$ the transition from one state to another occurs as usual via a network function, and when $\Theta = 1$ the state will change due to random bit permutation.

A PBN induces a homogeneous Markov chain whose states are pairs $(\mathbf{X}, \mathbf{f})$. The chain is ergodic and possesses a steady-state distribution. The steady-state distribution for the expression values is the marginal distribution of $\mathbf{X}$ (of the genes) relative to the steady-state distribution for the Markov chain.

We consider a PBN that has been randomly generated, which means random generation of the network functions. We fix the number of network functions at 5 and the number of predictor variables at 3, the latter being typical of the PBNs generated in practice. Each network function is of the form $\mathbf{f}_k=(f_{k_1}^{(1)}, f_{k_2}^{(2)}, \ldots, f_{k_n}^{(n)})$, for $k=1, 2, \ldots, 5$, and for $i = 1, 2, \ldots, 20$, the total number of genes. The component function $f_{k_i}^{(i)}$ is generated in two steps: (1) randomly select 3 genes from among $\{X_1, X_2, \ldots, X_{20}\}$ to be the variables for $f_{k_i}^{(i)}$; and (2) using these variables as entries in a truth table, uniformly randomly assign the 2^3 values of the truth table. We let the network functions have equal probability, $c_k = 0.2$, and we set $\alpha = 0.001$ and $\beta = 0.00001$. We have examined the 20 genes to see which one has the largest error when predicted by its own mean, without using any information from the other genes. This insures that good prediction does not result from the gene possessing little variation in the steady state.

References

[1] T. Cover, J. Van Campenhout, On the possible orderings in the measurement selection problem, IEEE Trans. Systems Man Cybernet. 7 (1977) 657–661.

[2] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Lett. 15 (1994) 1119–1125.
[3] H. Joe, Multivariate Models and Dependence Concepts, Chapman & Hall, London, 1997.
[4] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.
[5] G. Evan, T. Littlewood, A matter of life and cell death, Science 281 (1998) 1317–1322.
[6] H.H. McAdams, L. Shapiro, Circuit simulation of genetic networks, Science 269 (1995) 650–656.
[7] C.-H. Yuh, H. Bolouri, E.H. Davidson, Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene, Science 279 (1998) 1896–1902.
[8] M.J. van de Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A.M. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Astma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend, R. Bernards, A gene-expression signature as a predictor of survival in breast cancer, N. Engl. J. Med. 347 (2002) 1999–2009.
[9] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van de Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bemards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.
[10] E.R. Dougherty, M.L. Bittner, Y. Chen, S. Kim, K. Sivakumar, J. Barrera, P. Meltzer, J.M. Trent, Nonlinear Filters in Genomic Control, Proc. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, 1999.
[11] S. Kim, E.R. Dougherty, M.L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer, J.M. Trent, A general framework for the analysis of multivariate gene interaction via expression arrays, Biomed. Opt. 5 (2000) 411–424.
[12] R. Hashimoto, E.R. Dougherty, M. Brun, Z. Zhou, M.L. Bittner, J.M. Trent, Efficient selection of feature-sets possessing high coefficients of determination based on incremental determinations, Signal Process. 83 (4) (2003) 695–712.
[13] I. Tabus, J. Astola, On the use of MDL principle in gene expression prediction, Appl. Signal Process. 4 (2001) 297–303.
[14] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[15] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Natl. Med. 7 (2001) 673–679.
[16] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvverger, N. Loman, O. Johannsson, H. Olsson, B. Wifond, G. Sauter, O.P. Kallioniemi, A. Borg, J. Trent, Gene expression profiles distinguish hereditary breast cancers, N. Engl. J. Med. 34 (2001) 539–548.
[17] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA 96 (1999) 6745–6750.
[18] C.M. Perou, T. Sorlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lonning, A.L. Borresen-Dale, P.O. Brown, D. Botstein, Molecular portraits of human breast tumors, Nature 406 (2000) 747–752.
[19] M. Bittner, P. Meltzer, J. Khan, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, E. Gillanders, A. Leja, K. Dietrich, C. Beaudry, M. Berrens, D. Alberts, V. Sondak, N. Hayward, J. Trent, Molecular classification of cutaneous malignant melanoma by gene expression profiling, Nature 406 (2000) 536–540.
[20] S. Kim, E.R. Dougherty, I. Shmulevich, K.R. Hess, S.R. Hamilton, J.M. Trent, G.N. Fuller, W. Zhang, Identification of combination gene sets for glioma classification, Mol. Cancer Therapeutics 1 (2002) 1229–1236.
[21] I. Shmulevich, E.R. Dougherty, S. Kim, W. Zhang, Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks, Bioinformatics 18 (2002) 261–274.
[22] I. Shmulevich, I. Gluhovsky, R. Hashimoto, E.R. Dougherty, W. Zhang, Steady-state analysis of genetic regulatory networks modeled by probabilistic Boolean networks, Comp. Funct. Genom. 4 (2003) 601–608.
[23] U.M. Braga-Neto, R. Hashimoto, E.R. Dougherty, D.V. Nguyen, R.J. Carroll, Is cross-validation better than resubstitution for ranking genes, Bioinformatics 20 (2) (2004) 253–258.
[24] E.R. Dougherty, Small sample issues for microarray-based classification, Comp. Funct. Genom. 2 (2001) 28–34.

About the Author—TAILEN HSING is a professor of statistics at Texas A&M University in College Station, Texas. He received a Ph.D. degree from the University of North Carolina at Chapel Hill. He is a fellow of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. Professor Hsing's current research interests include bioinformatics and functional data analysis.

About the Author—LI-YU LIU received the B.S. degree in 1998 and M.S. degree in 2000, both in Agronomy from National Taiwan University, Taiwan. She is currently a Ph.D.-level graduate student under the supervision of Dr. Tailen Hsing in the Department of Statistics at Texas A&M University in College Station. Her recent research interests include bioinformatics and feature selection.

About the Author—MARCEL BRUN received his Ph.D. in Computer Sciences from the University of Sao Paulo, Brazil. He was involved in research in genomic signal processing at the Electrical Engineering Department, at Texas A&M University, and the Department of Biochemistry and Molecular Biology, at the University of Louisville, from 2000 to 2004. Currently he is an Associate Investigator at the Translational Genomics Research Institute, Arizona, with research focusing on computational biology, centered in design and simulation of genetic networks and analysis of large-scale biological data.

About the Author—EDWARD DOUGHERTY is a professor in the Department of Electrical Engineering at Texas A&M University in College Station. He holds a Ph.D. in mathematics from Rutgers University and an M.S. in Computer Science from Stevens Institute of Technology. He is author of twelve books, editor of four others, and author of more than one hundred and fifty journal papers. He is an SPIE fellow, is a recipient of the SPIE President's Award, and has served as editor of the Journal of Electronic Imaging for six years. Prof. Dougherty has contributed extensively to the statistical design of nonlinear operators for image processing and the consequent application of pattern recognition theory to nonlinear image processing. His current research is focused in genomic signal processing, with the central goal being to model genomic regulatory mechanisms. He is head of the Genomic Signal Processing Laboratory at Texas A&M University.