Submitted to the Annals of Applied Statistics

INTEGRATIVE EXPLORATION OF LARGE HIGH-DIMENSIONAL DATASETS

By Christopher Pardy[†], Sally Galbraith, and Susan R Wilson[§]

University of New South Wales [†§] and Australian National University [§]

Large, high dimensional datasets containing different types of variables are becoming increasingly common. For exploring such data there is a need for integrated methods. For example, a single genomic experiment can contain large quantities of different types of data (including clinical data) that make it a challenge to coherently describe the patterns of variability within and between the interrelated datasets. Mutual information (MI) is a widely used information theoretic dependency measure that also can identify nonlinear and non-monotonic associations. First we develop a computationally efficient implementation of MI between a discrete and a continuous variable. This implementation allows us to apply a coherent approach to all comparisons arising from continuous and categorical data. As commonly applied, MI can have high levels of bias. So we present a novel development of mutual information (MI) that reduces the bias, and that we term bias corrected mutual information (BCMI). Further, BCMI is useful as an association measure that can be incorporated in subsequent analyses such as clustering and visualisation procedures.

To demonstrate our approach a genomic dataset is re-examined. This dataset contains single nucleotide polymorphisms (SNPs, a discrete variable), gene expression levels and clinical data (all continuous variables). Our approach allows us to integrate these different types of data by exploring associations both within and between these types of variables.

1. Introduction. As large and complex datasets become increasingly available there is a growing need for exploratory methods that can identify novel associations between variables. For example, integrative genomic experiments containing diverse types of measurements on the same subjects give rise to multiple datasets potentially containing many novel interactions. If we knew in sufficient detail the underlying biological processes we could design experiments or build models accordingly. However there are many cases where experiments are essentially exploratory with data being used to generate further hypotheses.

[†]This research has been supported by the Australian National Health and Medical Research Council grant 525453.

One such experiment provides a motivation for the methods described in this paper. The liver tissue dataset we use to demonstrate our approach arose from an F2 intercross mouse model of obesity. The data consist of clinical measurements (continuous variables), gene expression levels (also continuous) and single nucleotide polymorphism genotypes (categorical variables). Our procedure enables exploration of associations both within and between the continuous and categorical variables, and is in this sense integrative.

It can be difficult to assess any parametric assumptions that may be made during the analysis of large genomic datasets. Often the most effective technique for doing so, plotting the variables, is infeasible due to the combinatorial explosion of possible pairwise comparisons. However, when a selection of variables is examined it often becomes clear that the data are not well described by the most commonly used parametric models. It is particularly clear that Gaussian assumptions are frequently inappropriate due to the presence of skew. It is also often the case that high leverage points exist that may result in association measures such as Pearson correlation giving spuriously strong results. Thus we seek an association measure that is valid for potentially skewed data and is relatively insensitive to the occasional outlying value.

Mutual information (MI) is an information theoretic measure of dependency that has been widely used to identify nonlinear associations. MI has been used as an association measure in bioinformatics and its estimation by kernel density approaches is well known (for example Steuer et al. (2002)). It is relatively straightforward to use MI to assess dependency between pairs of continuous variables and pairs of discrete variables. However for comparisons between a discrete and a continuous variable MI has been deemed to pose too many computational problems to be able to be applied automatically as is required for exploring large datasets (Dawy et al., 2006).

Therefore, there is a need for easily implementable methods based on a measure of association that can accommodate comparisons between all types of variables. Such methods would enable the researcher to scan through the data, regardless of whether variables are discrete or continuous, and determine which variables may be associated. The most common current approaches include either to simply discretize the continuous variable and calculate a cross-tabulation based score (for example, as done by Dawy et al. (2006)), or to assign numeric coding to the categorical variables and analyze them as if they were continuous (for example, Chu et al. (2009)). Here we propose an alternative approach that does not require such awkward manipulation of the data. This involves an implementation of MI using discrete and continuous variables that is supported by statistical rigor, and that can

be applied automatically with good accuracy.

The accuracy of our association measure is improved by the use of a nonparametric correction for estimation bias using the jackknife. Although the jackknife has proved poor for the purposes of statistical inference, the original intention of the jackknife as a bias correction (Quenouille, 1956) remains valid and will be shown by simulation to work well for our purposes. We refer to the application of the jacknife bias correction to MI estimation as Bias Corrected Mutual Information (BCMI). In addition we show that BCMI is less affected by outlying high leverage points than Pearson correlation.

Scientists are increasingly aware of the need to identify associations within their data. Our approach is exploratory, with minimal assumptions as well as being robust compared with other approaches. Further, there is software readily available that is easy to use so researchers can readily explore the interactions in their data. Thus BCMI can be useful during the analysis process in a wide range of scenarios, particularly with large quantities of data.

This paper is organized as follows. Section 2 introduces the liver tissue dataset which provides a motivating example for our approach. Section 3 presents our methods by describing the information theoretic concepts used and their nonparametric estimation. In addition we describe a nonparametric bias correction that can be shown to reduce error. Section 4 presents simulation studies for comparisons between continuous and categorical variables. Section 5 describes the application of BCMI to the liver tissue dataset. Section 6 presents discussion and conclusions.

2. Liver tissue data. We demonstrate the usefulness of BCMI using a publicly available F2 intercross dataset containing SNPs and gene expression levels in female mouse liver tissue. We use the same dataset analyzed in Ghazalpour et al. (2006) and Fuller et al. (2007) containing 135 mice, 3421 genes, 20 clinical measurements and 1065 SNPs. These data are available from `http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/MouseWeight/`. The experiment is described in detail in Wang et al. (2006), and briefly in Ghazalpour et al. (2006). The F2 intercross is used to create a population with variability in the outcome of interest. The strains C57BL/6J and C3H/HeJ (both ApoE$^{-/-}$) are chosen to have phenotypes compatible with the metabolic syndrome and to be particularly susceptible to weight-related clinical outcomes. They were fed on a high-fat 'western' diet from 8 to 24 weeks of age when they were sacrificed, at which point clinical outcomes were measured. Genotyping and microarray analysis of gene expression took place after sacrifice.

The dataset contains clinical outcome variables that are continuous in nature, categorical SNP data obtained from genotyping, and continuous gene expression measurements obtained from microarrays. Our approach can consistently measure associations amongst all of these variables, enabling an overall view of the association structure.

3. Method.

3.1. *Information measures.* Consider a discrete random variable X with $P(X = x) = p(x)$. In the context of the liver tissue dataset, X might represent the genotype obtained from the SNP data, taking one of three possible values depending on whether it is heterozygous (denoted Aa or H) or one of the two possible homozygous types (AA or A, and aa or B). The entropy of X (Shannon's entropy) is given by

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

Note that we use natural logarithms throughout. All information theoretic quantities in this section can be applied for continuous variables (such as gene expression data) by replacing sums with integrals.

For two discrete random variables X and Y with joint probability mass function $P(X = x, Y = y) = p(x, y)$ the mutual information (MI) is

$$I(X,Y) = \sum_x \sum_y p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3.1}$$

$$= E_{(X,Y)}\left[\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right]. \tag{3.2}$$

By defining appropriate probability measures we can apply (3.2) in the case where one variable is discrete (such as genotype) and the other is continuous (such as gene expression). This approach is described in the next subsection.

The joint entropy of X and Y is

$$\begin{aligned} H(X,Y) &= -\sum_x \sum_y p(x,y)\log\left(p(x,y)\right) \\ &= -E_{(X,Y)}\left[\log\left(p(x,y)\right)\right], \end{aligned}$$

and is related to MI by

$$I(X,Y) = H(X) + H(Y) - H(X,Y).$$

For an example of MI between two continuous variables (Cover and Thomas, 2006), consider bivariate random normal variables with mean zero, common

variance and correlation ρ. The resulting MI is $-\frac{1}{2}\log(1-\rho^2)$, equivalently, $\rho = \pm\sqrt{1-e^{-2\mathrm{MI}}}$. Note that MI has the same value for positive or negative correlations of the same magnitude. The relationship between MI and correlation for normal variables is given in Table 1. This can be a useful guide for interpreting the strength of association indicated by various MI values.

MI	0.00005	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
$\lvert\rho\rvert$	0.01	0.14	0.31	0.43	0.51	0.57	0.63	0.67	0.71	0.74
MI	0.45	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.95
$\lvert\rho\rvert$	0.77	0.80	0.84	0.87	0.89	0.91	0.93	0.94	0.95	0.99

TABLE 1

MI vs $|\rho|$ for the bivariate normal distribution for selected MI values.

3.2. *A mixture distribution model for the relationship between a categorical variable and a continuous variable.* A novel aspect of our approach is the calculation of MI values for comparisons between one discrete and one continuous variable. We model the distribution of the continuous variable as a mixture (that is, linear combination) of conditional distributions for each level of the categorical variable, as follows. Let X be a categorical random variable with g possible values $x_1, \ldots, x_g$ representing g groups, for example the genotype of a SNP usually has $g = 3$. Write $P(X = x_i) = p_i$ for $i \in \{1, \ldots, g\}$. Let Y be a continuous random variable (such as a gene expression level) with density and distribution functions $f_Y(y)$ and $F_Y(y)$ (i.e., $Y : \Omega_Y \to \Re$). The joint distribution of X and Y is $F_{(X,Y)}(x, y)$ with corresponding density $f_{(X,Y)}(x, y)$. For each x_i we have a conditional distribution for $Y|X = x_i$ with continuous density function $f_{Y|X=x_i}(y)$. For example, if X represents genotype, with three possible values, and Y represents gene expression, we envisage three potentially different distributions for gene expression, depending on the genotype. For notational simplicity, we shorten $f_{Y|X=x_i}(y)$ to $f_i(y)$, and omit subscripts where confusion is unlikely. We assume all integrals exist. This setup gives rise to the unusual probability space characterized by the joint distribution $F(X, Y) : (\Omega_X \times \Omega_Y) \to (\chi \times \Re)$ where χ is the discrete set of values taken by X; $F(X, Y)$ is best thought of as g separate continuous (conditional) distributions.

We write the density of Y as

$$f(y) = \sum_{i=1}^{g} p_i f_i(y). \tag{3.3}$$

By writing $f_X(x) = \sum_{i=1}^{g} p_i \delta_{x_i}(x)$ and $f(y|x) = \sum_{i=1}^{g} f_i(y)\delta_{x_i}(x)$ we ex-

press the joint density as

$$f(x,y) = \sum_{i=1}^{g} p_i f_i(y) \delta_{x_i}(x),$$

where $\delta_{x_i}(x)$ is an indicator function taking value 1 if $x = x_i$ and 0 otherwise.

PROPOSITION 3.1. *The mutual information (based on (3.2)) under this model is given as*

$$I(X,Y) = \sum_{i=1}^{g} p_i \int_{y \in \Re} f_i(y) \log\left(\frac{f_i(y)}{f(y)}\right) dy. \tag{3.4}$$

The proof is given in Section S1 of the supplementary material.

This main result was derived independently by Dawy et al. (2006); equation (17) of their paper can be shown to be equivalent, although it was arrived at by a different argument and without making it clear that $f(y)$ is a mixture of the conditional distributions $f_i(y)$. In addition our computational approach solves issues raised by Dawy et al. (2006) regarding the practical application of these procedures (that is, the estimation approach we describe in the following can be automated and potentially difficult numerical integrations are replaced by the application of the law of large numbers (LLN)). Equation (3.4) can be interpreted as

$$I(X,Y) = D(f(x,y)||f(x)f(y)) = \sum_{i=1}^{g} p_i D(f_i(y)||f(y)), \tag{3.5}$$

where

$$D(f_X||f_Y) = \int_{\Omega} \log\left(\frac{f_X(u)}{f_Y(u)}\right) dF_X(u) = \int_{-\infty}^{\infty} f_X(u) \log\left(\frac{f_X(u)}{f_Y(u)}\right) du$$

is the Kullback–Leibler divergence from the distribution of random variable X to the distribution of Y. We note that it is straightforward to extend this result to other information measures described by Principe (2010), namely (i) Renyi entropy, (ii) a measure based on the Cauchy-Schwarz inequality, and (iii) a measure based on Euclidean distances (details are given in Section S2 of the supplementary material).

3.3. *Nonparametric estimation.* *Kernel density estimation* is a nonparametric approach to the estimation of probability distributions (Wand and Jones, 1995) that requires the choice of a *smoothing parameter* (also called the *bandwidth*). We use kernel approaches as a basis for calculating MI values, and automate the procedure by using a data-driven 'Direct plug-in' estimator to estimate an optimal bandwidth (as proposed by Sheather and Jones (1991)). An evaluation of R's default `density()` function with associated bandwidth estimator `bw.SJ()` often resulted in over-smoothing, so we prefer to use the `dpik()` function from the R package `KernSmooth`.

3.4. *Comparisons between two continuous variables.* Qiu, Gentles and Plevritis (2009) propose a non-parametric Gaussian kernel smoother for comparisons where both variables are continuous. For a sample $\mathbf{z} = z_1, \ldots, z_n$, where n is the number of observations in $\mathbf{z}$, the well-known univariate kernel density estimator is

$$\hat{f}(z) = \frac{1}{n}\sum_{j=1}^{n} K_h(z - z_j), \tag{3.6}$$

where Qiu, Gentles and Plevritis (2009) have chosen

$$K_h(z - z_j) = \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{1}{2h^2}(z-z_j)^2},$$

which is a Gaussian kernel with smoothing parameter h. In general, a (scaled) kernel function $K_h(y)$ is a symmetric function such that $\int_{-\infty}^{\infty} K_h(y)dy = 1$. For continuous samples $\mathbf{x}$ and $\mathbf{y}$ of size n, their MI estimator is

$$\hat{I}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n\sum_j K_{h_1}(x_i - x_j)K_{h_2}(y_i - y_j)}{\sum_j K_{h_1}(x_i - x_j)\sum_j K_{h_2}(y_i - y_j)}\right). \tag{3.7}$$

Note that the numerator is a bivariate Gaussian 'product' kernel evaluated at observed sample values. This formulation makes it clear that we are free to choose kernels other than the Gaussian used by Qiu, Gentles and Plevritis (2009).

To further improve performance we use the Epanechnikov kernel, which takes the form

$$K_h(y) = \frac{3}{4h}\left(1 - \left(\frac{y}{h}\right)^2\right) I_{\{|\frac{y}{h}|<1\}}, \tag{3.8}$$

where $I_{\{C\}}$ denotes an indicator function for condition C. For example, $I_{\{|\frac{y}{h}|<1\}} = 1$ when $|\frac{y}{h}| < 1$ and zero otherwise. It can be shown that this

kernel is optimal with respect to asymptotic mean integrated squared error (AMISE, see Wand and Jones (1995)). In short, if $\hat{f}(y)$ is the kernel density estimate of the true density function $f(y)$ the mean integrated squared error (MISE) is

$$E\left[\int(\hat{f}(y)-f(y))^2dy\right].$$

The AMISE is the MISE as $n\to\infty$ such that $h\to 0$ and $nh\to\infty$ (so that $h\to 0$ slower than n^{-1}). The use of the Epanechnikov kernel in (3.7) leads to our preferred estimator for comparisons between continuous variables.

The Epanechnikov kernel provides a number of advantages. Compared to the normal kernel it has a relative asymptotic efficiency of 1.05 (Wand and Jones, 1995), which can be thought of as roughly increasing accuracy to the same extent as a 5% increase in sample size. A particularly important advantage for our intended use in large genomic datasets is the lack of an exponential function in (3.8). This greatly improves computational efficiency.

3.5. *Comparisons between a continuous and a categorical variable.* Comparisons between a continuous and a categorical variable require us to estimate Equation (3.4). Our first estimate for (3.4) is based on an extension of (3.7). Define n_{x_i} as the number of observations with $X=x_i$, and $\sum_{j|X=x_i}$ as a sum over these observations (having n_{x_i} elements). We estimate each integral in (3.4) by taking an average of sample kernel estimates

$$\hat{E}_{Y|X=x_i}\left[\log\left(\frac{f_i(y)}{f(y)}\right)\right]=\frac{1}{n_{x_i}}\sum_{k=1}^{n_{x_i}}\log\left(\frac{\frac{1}{n_{x_i}}\sum_{j|X=x_i}K_h(y_k-y_j)}{\frac{1}{n}\sum_j K_h(y_k-y_j)}\right), \tag{3.9}$$

since, by the law of large numbers (LLN), $\hat{E}_{Y|X=x_i}\to E_{Y|X=x_i}$ as $n\to\infty$. This gives the estimator

$$\hat{I}(X,Y)=\sum_{i=1}^{g}\hat{p}_i\frac{1}{n_{x_i}}\sum_{k=1}^{n_{x_i}}\log\left(\frac{n\sum_{j|X=x_i}K_h(y_k-y_j)}{n_{x_i}\sum_j K_h(y_k-y_j)}\right). \tag{3.10}$$

The LLN is required here as this estimator takes values only at the observed data points, allowing us to avoid a potentially difficult numerical integration and making the estimation straightforward to automate. The $\hat{p}_i$ are simply given by the observed relative frequencies of the x_i.

Substituting the Epanechnikov kernel (3.8) into Equation (3.10) leads to

our preferred estimator for (3.4):

(3.11)
$$\hat{I}(X,Y) = \sum_{i=1}^{g} \hat{p}_i \frac{1}{n_{x_i}} \sum_{k=1}^{n_{x_i}} \log \left(\frac{n \sum_{j|X=x_i} (1 - (\frac{y_k - y_j}{h})^2) I_{\left\{ |\frac{y_k - y_j}{h}| < 1 \right\}}}{n_{x_i} \sum_j (1 - (\frac{y_k - y_j}{h})^2) I_{\left\{ |\frac{y_k - y_j}{h}| < 1 \right\}}} \right).$$

3.6. *Comparisons between two categorical variables.* For comparisons between two categorical variables, we simply replace probabilities in (3.1) with observed relative frequencies:

$$I(X,Y) = \sum_x \sum_y \hat{p}(x,y) \log \left(\frac{\hat{p}(x,y)}{\hat{p}(x)\hat{p}(y)} \right), \tag{3.12}$$

where $\hat{p}(x,y) = \frac{1}{n^2} \sum_x \sum_y I_{\{X=x, Y=y\}}$, $\hat{p}(x) = \frac{1}{n} \sum_x I_{\{X=x\}}$ and $\hat{p}(y) = \frac{1}{n} \sum_y I_{\{Y=y\}}$.

3.7. *Interpretation of MI estimators for mixed comparisons.* Our estimators can measure the degree of association between a continuous variable and a grouping factor (which we term a 'mixed' comparison). This ability allows us to integrate diverse types of omics data by measuring associations between all types of variables observed. For mixed comparisons, we aim for a high MI value when a continuous variable is clearly separated into groups by the categorical variable, and for low values when there is no such relationship. An example from a dataset (Ghazalpour et al., 2006) that we analyze in detail in Section 5 is used in Figure 1. Here we use the approach in Section 3.3. The left figure shows three estimated conditional distributions with little overlap corresponding to a MI value of 0.97. The right figure shows the result of a random permutation of group membership where the substantial overlap of continuous distributions results in a low MI value of 0.05. In this way a high MI value identifies an association such that a discrete variable can distinguish or separate values of a continuous variable. It is also notable that permutation of group membership, while a bijective function, cannot be written in terms of a Jacobian transformation and so does not leave the MI invariant. Thus we can perform a permutation test for the null hypothesis of no association.

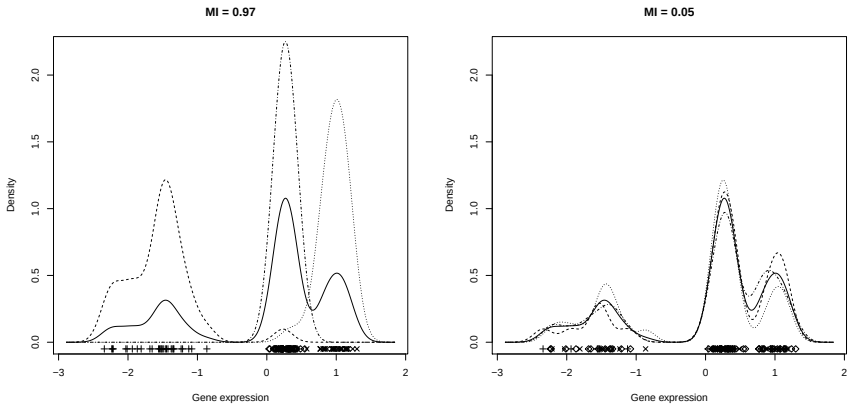

FIG 1. *Interpretation of high and low MI values. This is an example from the Ghazalpour et al. (2006) data showing kernel density estimates of Aqr gene expression according to rs3149884 genotype. The left figure shows the combined ($f(y)$, solid line, see Equation (3.3)) and conditional ($f_i(y)$, dashed lines, see Equation (3.3)) densities highlighting the strong observed association; $MI = 0.97$. The rug plot and densities show a clear separation of the groups (each group has a different symbol in the rug plot). The right figure shows a low MI resulting from a random permutation of the grouping variable; $MI = 0.05$. This demonstrates how greater Kullback–Leibler divergences (the area between each dashed line and the black line) correspond to greater MI.*

3.8. *Bias correction using the jackknife.* Our proposed MI estimator is biased because it uses kernel density estimation, and kernel density estimators are biased (Wand and Jones, 1995). This section describes a bias correction method based on the jackknife.

The jackknife is a nonparametric statistical procedure than can be thought of as a computational simplification of the bootstrap (Efron and Gong, 1983). It can be used for inference but more importantly also gives a correction for estimation bias. As the mean squared error (MSE) of an estimator is defined as $\text{MSE} = \text{bias}^2 + \text{variance}$, a reduction in bias has the potential to greatly improve accuracy under this criterion. The jackknife bias correction is based on a Taylor expansion of the expectation of the estimator (Quenouille, 1956; Efron, 1982). In short, jackknife estimates can be used to subtract the $1/n$ term leaving an estimator with $O(1/n^2)$ bias rather than $O(1/n)$.

Consider a sample of size n, $Y_1, \ldots, Y_n$ from an unknown probability distribution F and a statistic θ that is a function of a realization of this sample $\theta(Y_1, \ldots, Y_n)$. Call $\theta_{(i)}$ the ith jackknife replication of this statistic. $\theta_{(i)} \equiv \theta(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$, that is, the statistic θ calculated on the dataset excluding the ith observation. The jackknife procedure is im-

plemented by calculating all n of the $\theta_{(i)}$ which are used as the basis for inference on θ. If the original estimated value is $\hat{\theta}$ and the estimated jackknife values are $\hat{\theta}_{(i)}$, the jackknife bias corrected estimator is

$$\hat{\theta}_{(.)} = n\hat{\theta} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}.$$

We refer to the application of jackknife bias correction to our MI estimators as 'bias corrected mutual information' (BCMI). Further details are given Section S3 of the supplementary material. A brief simulation study suggested that the asymptotic variance of BCMI goes to zero at a rate of $O(1/n)$, details are given in supplementary Section S4.

4. Simulation studies. Simulation is used to assess the performance of the jackknife bias correction and the effect of various choices for the kernel density estimation parameters. Our approach uses kernel density estimation and thus has the smoothing bandwidth as a free parameter. The choice for this parameter can be considerably narrowed by the use of plug-in bandwidth estimation (Sheather and Jones, 1991), which reduces the choice to a discrete number of integers corresponding to the number of 'levels' of bandwidth estimation used (often between 1 and 5; Wand and Jones (1995) state that 2 is a common choice). We perform a simulation study to evaluate the accuracy of our estimators for each choice of plug-in level with and without jackknife bias correction.

4.1. *Mixed comparisons.* This section concerns comparisons between a categorical and a continuous variable. We consider a 3 group categorical variable denoted $\{A, H, B\}$ corresponding to the 3 genotypes for SNP data, where H is the heterozygote.

Firstly, conditional on each category we simulate normally distributed continuous data with varying degrees of separation between groups. The means for the three simulated genotype categories $\{A, H, B\}$ are $\{-5, 0, 5\}$ for 'large separation', $\{-0.5, 0, 0.5\}$ for 'small separation' and $\{0, 0, 0\}$ for 'no separation'. All normal distributions have unit variance. There are equal numbers of observations for each of the 3 genotypes, with a total sample size of either $n = 150$ or $n = 30$. In each case we perform 10,000 simulation runs. The target MI values are determined by evaluating (3.4) using numerical integration, with estimation accuracy measured by MSE. Numerical integration was performed using the R `integrate()` function which reported error bounds less than 10^{-4}. The smoothness of the density functions makes it unlikely that the numerical integration introduces any systematic bias.

Due to the small observed MSE values we present results in terms of $-\log_{10}(\text{MSE})$, higher values indicate greater accuracy. Results are given in Table 2. For the large separation scenario with $n = 150$ the bias correction reduces MSE for plug-in level 1 and 2 but increases MSE for higher levels. For large separation with $n = 30$ the bias correction continues to reduce MSE for levels 1 through 4. For the small separation scenario the bias correction greatly improves MSE for all bandwidth levels and both sample sizes. This is also the case for the no separation scenario. Boxplots of all simulation results are given in Section S5 of the supplementary material.

	Level									
Separation	1	**1b**	2	**2b**	3	**3b**	4	**4b**	5	**5b**
$n = 150$										
Large	3.5	**3.9**	5.2	**5.4**	5.3	**4.5**	4.8	**4.3**	4.7	**4.3**
Small	3.1	**5.5**	3.1	**5.4**	3.0	**5.3**	3.0	**5.2**	2.9	**5.1**
None	2.9	**5.0**	2.9	**5.0**	2.9	**5.0**	2.8	**4.9**	2.8	**4.8**
$n = 30$										
Large	1.5	**1.7**	1.9	**2.2**	2.6	**3.3**	3.4	**4.1**	4.4	**3.3**
Small	2.1	**3.9**	2.0	**3.8**	2.0	**3.7**	1.9	**3.6**	1.8	**3.4**
None	1.9	**3.7**	1.9	**3.6**	1.8	**3.5**	1.8	**3.4**	1.7	**3.2**

TABLE 2

Simulation results for normally distributed data. This table shows $-\log_{10}(MSE)$ for two sample sizes ($n = 150$ and $n = 30$) with various degrees of separation between groups as described in Section 4.1. The number of levels chosen for plug-in bandwidth selection was varied, and for each scenario MI and BCMI were calculated; the use of the bias correction is indicated by bold type and 'b' next to the number of levels.

4.2. *U-shaped associations.* A major advantage of MI-based methods is the ability to identify non-monotonic associations. It is common practice to give a numerical coding to a SNP under the assumption of a linear allelic effect (for example, see Ghazalpour et al. (2006)). Here we code the genotype categories as $\{A \to 0, H \to 1, B \to 2\}$. If the data are such that a central category gives a baseline gene expression level with both other groups showing an increase (or both a decrease) these correlation estimates will be near zero whereas MI will assign a high score. Furthermore, MI estimates will not change if the labels (not the data) of the categorical variables are permuted. If a correlation measure is to be used, a better approach is given by the 'heterozygote advantage/disadvantage model' (Laird and Lange, 2010). For the heterozygote advantage model we code the genotypes $\{A \to 0, H \to 1, B \to 0\}$. Figure 2 shows boxplots of simulation results where the continuous distributions are normal with means $\{5, 0, 5\}$ and unit variance. We use 10,000 simulated values for each measure, with $n = 150$.

The target MI value of 0.62 was determined by numerical integration of (3.4). For these data the linear coding targets a value of zero, whereas the MI-based approach and the heterozygote advantage coding both identify a strong association. We choose a single plug-in level of 3 for this comparison as the difference in MSE due to the choice of plug-in levels is small compared to the difference in MSE between the MI measures and the correlation measures. MI is able to identify this association without having to explicitly choose a model that is designed for this purpose.

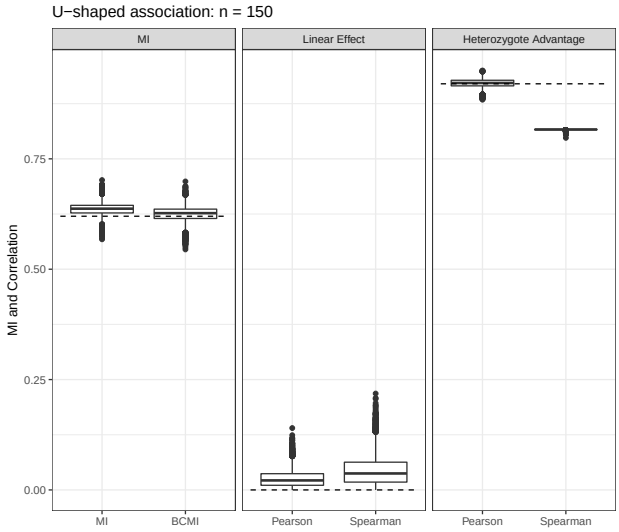

FIG 2. *Boxplots of simulation results for a U-shaped relationship with categories; see Section 4.2 for details. MI and correlation are both plotted on the same numerical scale. The two boxplots to the left show MI and BCMI estimates with plug-in level 3. The two boxplots in the middle show Pearson and Spearman correlation estimates for the (incorrect) linear allelic effect coding and the two boxplots on the right show Pearson and Spearman correlation estimates for the heterozygote advantage coding. The target values are given by dashed horizontal lines, namely 0.62 for MI and BCMI (see text), 0 for Pearson and Spearman correlations with the linear effect coding, and 0.92 for Pearson correlation with the heterozygote advantage coding.*

4.3. *Skewed data.* Skewed or heteroscedastic data or data with outlying groups are challenging to analyze. To investigate such data we sample from gamma distributions under two scenarios. The distributions are chosen to have different amounts of skew and variability and are shown in Figure 3. The distributions in the left plot are chosen to give little separation between groups while being skewed. Each group is given a different variance. The distributions in the right plot are chosen to give a higher MI, with two skewed

distributions (with the same variance, but different skew) and a symmetrical group with less variance but a much larger mean. In both cases we use $n = 150$ (50 in each group) with 10,000 simulations runs per estimator. We parameterize the gamma distributions by their mean and variance to ease comparisons with our previous results. Results are shown in Table 3. The bias correction increases MSE for plug-in level 1 in the second scenario, but reduces MSE in all other levels and reduces MSE in all levels in the first scenario. Boxplots of these results are given in Section S5.1 of the supplementary material.

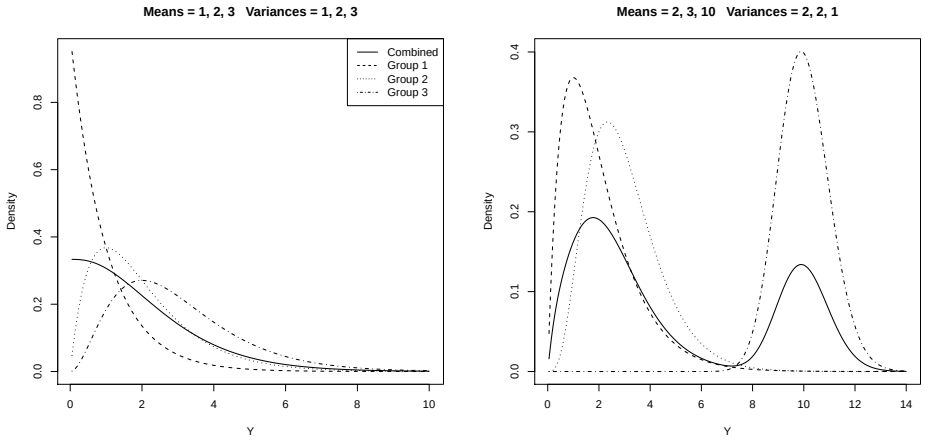

FIG 3. *Distributions for skewed data simulations. The left figure shows the density functions used to create skewed heteroskedastic data, the right figure shows densities used for skewed data with an outlying symmetrical group. The solid line is $f(y)$ (Equation 3.3) with the dashed lines representing each of the $f_i(y)$, labelled as Groups 1, 2 and 3. In the left plot the densities of Groups 1, 2 and 3 have means 1, 2 and 3 with variances 1, 2 and 3, respectively. In the right plot the densities of Groups 1, 2 and 3 have means 2, 3 and 10 with variances 2, 2 and 1, respectively.*

4.4. *Further simulation results.* Simulation results for comparisons between pairs of continuous variables are presented in Section S6 of the supplementary material. Simulation results for comparisons between pairs of categorical variables are given in Section S7. A summary of all simulation results is given in Section S8.

5. Analysis of liver tissue data. This section demonstrates the usefulness of BCMI using the liver tissue dataset introduced in Section 2. The publicly available data have been normalized and filtered as described in

	Level									
	1	**1b**	2	**2b**	3	**3b**	4	**4b**	5	**5b**
n = 150 Means = 1, 2, 3; Vars = 1, 2, 3	3.4	**4.2**	3.1	**4.8**	2.9	**5.7**	2.8	**6.4**	2.7	**5.3**
n = 150 Means = 2, 3, 10; Vars = 2, 2, 1	4.5	**3.8**	4.4	**4.9**	3.9	**6.7**	3.7	**5.9**	3.6	**5.4**

TABLE 3

Simulation results for skewed data. Data are generated following gamma distributions as shown in Figure 3 with means and variances as given. This table shows $-\log_{10}(MSE)$ *for various choices of the level parameter for plug-in bandwidth estimation with bias correction indicated by bold type and the letter 'b'.*

Ghazalpour et al. (2006) such that the 3421 available genes are among the most variable and most highly connected (in the sense that the row sums of a correlation matrix between them has large values). This makes our task easier by removing uninformative variables and lowering the overall number of comparisons. Note that this filtering biases the dataset in favor of containing variables with linear associations.

5.1. *Associations between continuous and categorical variables.* In this section we focus on associations between gene expression measures and SNPs as this is the most novel aspect of our development of MI. We find that the ability to identify non-monotonic associations as discussed in Section 4.2 is borne out in our analysis of real data. For example, a U-shaped association was observed between the Car9 gene on chromosome 4 and the rs3702474 SNP on chromosome 16; see Figure 4. These box and violin plots (Hintze and Nelson, 1998) were produced with the R package `vioplot` using the default 1.5 times interquartile range threshold to determine outliers. For clarity large outlying gene expression values less than -2 have been removed from the plot; two are in the A group (-7.29 and -4.56), one is in the H group (-2.94). These plots show boxplots (black rectangles) with superimposed kernel density estimates. The kernel densities are truncated at the minimum and maximum observed values within each group. The medians (white circles) show the general U-shaped nature of this relationship; the kernel densities indicate the bulk of the data. This degree of overlap results in the moderate BCMI of 0.23 (0.2 without outliers). Note that our BCMI estimation procedure is able to handle the bimodality within genotype A.

A common approach for the analysis of SNP data is to assume an additive effect for each copy of an allele, for example using the numerical coding $\{A \to 0, H \to 1, B \to 2\}$ (as done in Ghazalpour et al. (2006)). We refer to

the use of this coding as the linear allele effect linear model, $y_i = \beta_0 + x_i\beta_1$, where $x_i \in \{0, 1, 2\}$. The $\{0, 1, 2\}$ SNP coding is also used for the calculation of Pearson and Spearman correlation values.

Table 4 shows association values and p-values for the relationship between Car9 and rs3702474 with and without the outliers. Association is measured using Pearson correlation, Spearman correlation and BCMI. P-values are given for the linear allele effect linear model ('Linear'), analysis of variance ('ANOVA') where the SNP values are used as a grouping variable, and a BCMI permutation test ('BCMI perm'). The permutation test p-value is the result of 3 million random permutations of the SNP variable and is defined as the proportion of resulting BCMI values greater than or equal to the raw observed value. The Pearson correlation is not robust to outliers, changing from positive to negative due to the influence of just the outliers. The Spearman correlation is more robust, but misses the non-monotonic relationship and instead detects only a small negative association (removal of the outliers increases the magnitude of the Spearman correlation). BCMI is much more robust to the effect of the outliers. The linear model is unable to detect any statistically significant association with or without outliers. ANOVA shows no evidence of association at all when outliers are left in the data, but indicates very strong evidence of a difference between groups when the outliers are removed. The permutation test gives very small p-values both with and without outliers included, although the p-value is increased when the outliers are removed from the analysis.

A common use for association measures is to scan through a dataset to identify the strongest associations. Section S9 of the supplementary material contains tables of some of the strongest associations identified. Interestingly, there were several very strong associations between genes and SNPs lying on different chromosomes (termed 'trans' associations). At least 2 U-shpaed associations are discovered: the above-mentioned association between Car9 and rs3702474, and another between Olfr599 and rs3674895. Both of these are obscured by the shape of their associations and the presence of outliers. Some of the strongest trans associations have non-significant ANOVA and linear model p-values, and would thus have not been identified without using BCMI. Conversely, the overall highest BCMI values also corresponded to results with strong statistical significance of ANOVA or linear models, indicating that we did not miss any obvious strong signals within the data. Section S9 also details how we dealt with linkage disequilibrium (LD) resulting from the experimental design.

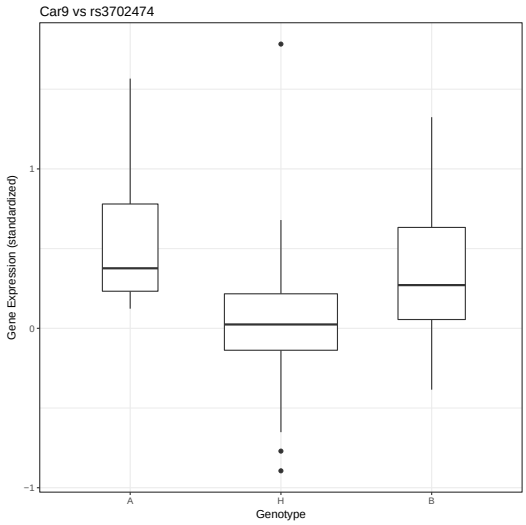

FIG 4. *Box and violin plots for gene expression of Car9 grouped according to rs3702474 genotype, with large negative-valued outliers removed (see Section 5.1). The width of each boxplot is proportional to the square root of the sample size in each group.*

Association measures

	Pearson	Spearman	BCMI
With outliers	0.14	-0.04	0.23
Without outliers	-0.14	-0.10	0.2

P-values

	Linear	ANOVA	BCMI perm
With outliers	0.169	0.389	7.3×10^{-6}
Without outliers	0.165	1.1×10^{-4}	1.8×10^{-5}

TABLE 4
Association measures and p-values for the U-shaped relationship between the Car9 gene and the rs3702474 SNP; see Section 5.1 for details.

5.2. *An exploratory visualization of results.* Association measures are routinely used to infer biochemical networks and functionally related sub-networks referred to as modules (for example Ghazalpour et al. (2006)). The BCMI values calculated by our approach are easily imported into network analysis software for this purpose (for example the actively-developed open source software Cytoscape (Shannon et al., 2003); we use version 2.8.3 released in May 2012).

5.3. *A permutation criterion for network visualization.* We choose to take a highly conservative approach and visualize the dependency structure as a network where nodes are connected based on Bonferroni-adjusted statistical significance. To remain consistent we would prefer to use a single procedure to assess statistical significance for all types of comparison. In keeping with the nonparametric focus of this work we define our visualization criteria based upon the results of a permutation test. Permutation tests are reliable in the sense that they maintain their nominal level, but are computationally intensive. The number of resamples required can be very large in multiple testing scenarios with large numbers of comparisons. The approximation described below works well to overcome the computational expense. Note that a practical, alternative approach would be to simply choose an arbitrary cutoff to select only the very strongest identified associations.

5.4. *An approximate permutation test.* To reduce the computational burden of calculating a separate permutation null hypothesis distribution for each pairwise comparison (that is, applying the approach used in the single gene/SNP example above separately for each comparison) we follow Efron (2010, section 6.5). This approach combines all distinct pairwise comparisons to obtain a single approximate null hypothesis distribution, which is then used for all comparisons.

Full details of the approach are given in Section S10.1 of the supplementary material. Essentially, combining the comparisons allows us to greatly increase the number of draws from the permutation null hypothesis distribution that can be obtained from a single permutation. The resulting computational advantage is balanced by the fact that the empirical distribution of the resamples will be only an approximation to the null hypothesis distribution for any given comparison.

5.5. *Approximate permutation tests for the liver tissue data.* As described fully in Section S10.2, we apply the approximate permutation test separately for the three different types of comparisons generated by the liver tissue data: discrete/discrete, continuous/continuous and discrete/continuous. Based on the total number of comparisons of all types, a Bonferroni-adjusted threshold of 4.93×10^{-9} for the p-values was chosen in order to control the familywise error rate at $\alpha = 0.05$. To achieve this level, the numbers of resamples required were 40, 60 and 400 for the continuous/continuous, discrete/continuous and discrete/discrete comparisons, respectively.

An example of one of the networks thus discovered is given in Figure 5. In Figure 5 we can see some of the LD between the SNPs on chromosome 8 with some subsets of SNPs in higher LD than others (roughly five such

groups are evident, as seen by the darker lines indicating high BCMI, for example the group of SNPs at the bottom of the plot). No association between genes reached the level required to add an edge between them in the plot (BCMI > 0.45). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway information for the genes in Figure 5 is given in Table 5.

In addition Section S11 of the supplementary material contains network plots of all variables in the liver tissue data. Interestingly we note that the SNPs clustered into groups according to chromosome, as expected from the high degree LD present in the F2 intercross. Also, strong mixed-type associations caused groups of SNPs to 'attach' to a main group of associated gene expression variables. We also make a comparison to the original analysis of these data (Ghazalpour et al., 2006) and show that the genes within a group they identify as being associated with mouse weight tend to be near weight in an association network.

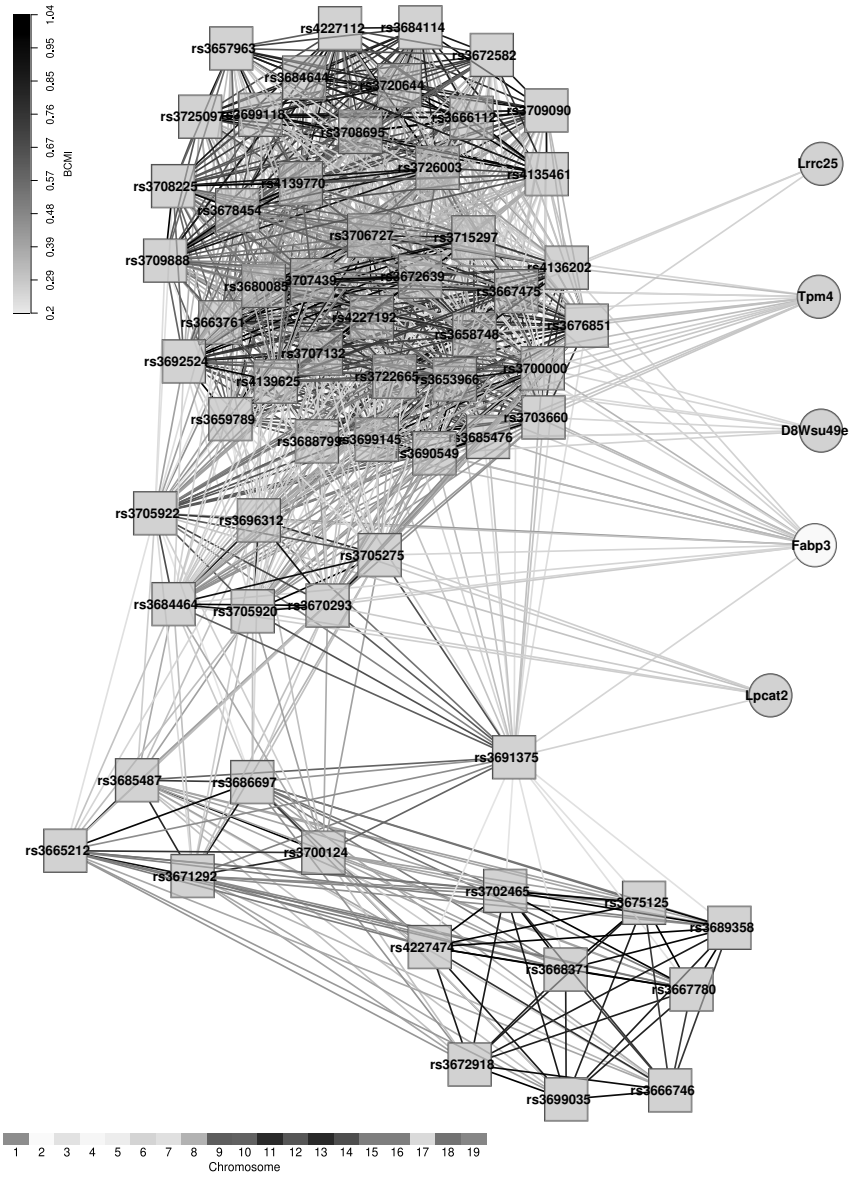

FIG 5. *Network plot for chromosome 8 determined by the permutation-based criterion applied to all types of association; for details see Sections 5.3 to 5.5. KEGG annotations for the genes in this plot are given in Table 5. Each gene expression level or SNP corresponds to a node in the network and is connected by an edge if the corresponding BCMI value is greater than the appropriate cutoff from Table S3. Genes are shown with circular nodes, SNPs are shown with square nodes. Genes and SNPs are color-coded according to chromosome; in this figure green nodes correspond to chromosome 8 and yellow nodes correspond to chromosome 4. The color of each edge reflects the corresponding BCMI value, ranging continuously from light-blue (low BCMI) to black (high BCMI). The electronic version of this plot uses vector graphics and can be enlarged as needed.*

Gene	Chromsome	KEGG pathways
Tpm4	8	Cardiovascular Diseases; Cardiac muscle contraction; Circulatory System; Cardiomyopathy
Fabp3	4	Fatty acid-binding protein 3; Muscle and heart; Endocrine System; PPAR signaling pathway
Lpcat2	8	Glycerophospholipid metabolism; Ether lipid metabolism; Metabolic pathways
Lrrc25	8	Unknown
D8Wsu49e	8	Unknown

TABLE 5
KEGG pathway annotations for the genes in Figure 5.

6. Discussion. We have developed an approach for exploring the dependency structure of large and complex datasets. Using BCMI as a consistent framework for quantifying dependency, we can search for strong associations for any type of variable, continuous or discrete. The generality of this measure makes it possible to detect a wide class of associations, which can be used to combine datasets containing different types of variables. Once calculated, BCMI can then be used to either inform or be directly fed into subsequent analysis. Although this work was motivated by the need to explore large genomic datasets, other efforts have been made to detect wide classes of novel associations in other settings (for example Reshef et al. (2011)) and BCMI is also suitable for such settings.

A potential application in genomics is the automated inference of gene ontologies, which have recently been inferred based on networks built from correlation measures (Dutkowski et al., 2013). Similarly, the network modules identified by Ghazalpour et al. (2006) are also ultimately based on pairwise Pearson correlation. An obvious first step to improving such procedures is to replace correlation with a more general measure, for which BCMI is highly suitable. The ability to include categorical variables allows the possibility of integrating diverse sources of genomic data, such as copy number variation or methylation status.

We use kernel density estimation to estimate mutual information values. Other methods, such as those based on sample spacings (Kraskov, Stogbauer and Grassberger, 2004), were evaluated but results were found to be highly variable. Density estimation by a mixture of spline functions (Kauermann and Schellhase, 2009) is accurate, but extremely computationally expensive.

Our approach is helped by the fact that in order to calculate information measures, we are essentially estimating $E(f(y))$ rather than $f(y)$ itself. This quantity is referred to as the *information potential* in Principe (2010). The use of the LLN for estimating information potentials is an interesting aspect

of our approach. Recall that we estimate $E(f(y))$ by calculating kernel-based estimates of $\hat{f}$ for the observed y values and then taking a sample average of these. Thus if our $\hat{f}$ values are inaccurate we add error only once into the procedure rather than twice as we would if we were to use numerical integration to evaluate $E(f(y))$.

Several alternative association measures have been proposed for identifying very general classes of dependency, such as the maximal information coefficient (MIC) (Reshef et al., 2011), the Brownian distance covariance (DCOR) (Szekely and Rizzo, 2009) and generalized correlation (GCOR) (Hall and Miller, 2009, 2011). When comparing pairs of continuous variables in the liver tissue data we found that BCMI gave high values to relationships showing a mutually exclusive or 'L-shaped' pattern (which is also readily identified by MIC, as discussed in Reshef et al. (2011)), whereas DCOR and GCOR performed better for linear associations with performance comparable to Pearson correlation. We did not make a detailed comparison of these measures with BCMI as they are infrequently used, require discrete variables to be given a numerical coding, and are all more computationally expensive than BCMI.

Computational tractability in particular is is an important feature of any method that is intended to be applied to large and high-dimensional data. Our approach is successful in balancing a desire to identify a wide class of possible associations with the ability to calculate association measures for large numbers of pairwise comparisons in reasonable time.

We have described a procedure for a bias corrected estimate of MI for all types of comparisons that can arise from data consisting of discrete and continuous variables. In ongoing research we are extending this to other types of data such as ordinal or censored (for example, survival) data. Kernel density estimates for censored data exist and can potentially be used with very little modification of our estimators (see for example Padgett and McNichols (1984) and Marron and Padgett (1987)). The application of MI to ordinal data is much less developed in the literature and is likely to be a difficult problem.

BCMI is a useful tool that can identify some nonmonotonic associations that can be missed by other correlation measures. It is robust to the presence of outliers and is fast to compute. Most importantly, it can be applied to all kinds of comparisons arising from a collection of continuous and categorical variables. BCMI is therefore highly suitable for an initial exploration of the dependency structure of high dimensional genomic data.

An R package `mpmi` is available at `http://r-forge.r-project.org/projects/mpmi/`.

SUPPLEMENTARY MATERIAL

Supplement A: Additional material
(supplement.pdf). Supplementary material is available and includes the proof of Proposition 3.1, alternative information measures, simulations and more results for the genomic and clinical data used in the paper.

References.

Chu, J., Weiss, S. T., Carey, V. J. and Raby, B. A. (2009). A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC systems biology* **3** 55.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley.

Dawy, Z., Goebel, B., Hagenauer, J., Andreoli, C., Meitinger, T. and Mueller, J. C. (2006). Gene mapping and marker clustering using Shannon's mutual information. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3** 47–56.

Dutkowski, J., Kramer, M., Surma, M. A., Balakrishnan, R., Cherry, J. M., Krogan, N. J. and Ideker, T. (2013). A gene ontology inferred from molecular networks. *Nature biotechnology* **31** 38–45.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. SIAM.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37** 36–48.

Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J. and Horvath, S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight. *Mammalian Genome* **18** 463–472.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusis, A. J. et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics* **2** e130.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics* **18** 533–550.

Hall, P. and Miller, H. (2011). Determining and Depicting Relationships Among Components in High-Dimensional Variable Selection. *Journal of Computational and Graphical Statistics* **20** 988–1006.

Hintze, J. L. and Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician* **52** 181–184.

Kauermann, G. and Schellhase, C. (2009). Density estimation with a penalized mixture approach Technical Report, Centre for Statistics, Bielefeld University.

Kraskov, A., Stogbauer, H. and Grassberger, P. (2004). Estimating mutual information. *Physical Review E* **69** 066138.

Laird, N. M. and Lange, C. (2010). *The fundamentals of modern statistical genetics*. Springer.

Marron, J. and Padgett, W. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics* **15** 1520–1535.

Padgett, W. and McNichols, D. T. (1984). Nonparametric density estimation from censored data. *Communications in Statistics-Theory and Methods* **13** 1581–1611.

Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Verlag.

Qiu, P., Gentles, A. J. and Plevritis, S. K. (2009). Fast calculation of pairwise mutual information for gene regulatory network reconstruction. *Computer Methods and Programs in Biomedicine* **94** 177–180.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science* **334** 1518–1524.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13** 2498–2504.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **53** 683–690.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18** S231–S240.

Szekely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics* **3** 1236–1265.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Chapman & Hall/CRC.

Wang, S., Yehya, N., Schadt, E. E., Wang, H., Drake, T. A. and Lusis, A. J. (2006). Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS genetics* **2** e15.

Christopher Pardy
School of Mathematics and Statistics
Faculty of Science
UNSW Sydney, NSW 2052 Australia
E-mail: cpardy@unsw.edu.au

Susan R Wilson
School of Mathematics and Statistics
Faculty of Science
UNSW Sydney, NSW 2052 Australia

Mathematical Sciences Institute
Australian National University
Canberra, ACT, Australia
E-mail: Sue.Wilson@anu.edu.au