Measuring Statistical Dependence via the Mutual Information Dimension

Mahito Sugiyama[1] and **Karsten M. Borgwardt**[1,2]
[1]Machine Learning and Computational Biology Research Group,
Max Planck Institute for Intelligent Systems and
Max Planck Institute for Developmental Biology, Tübingen, Germany
[2]Zentrum für Bioinformatik, Eberhard Karls Universität Tübingen, Germany
{mahito.sugiyama, karsten.borgwardt}@tuebingen.mpg.de

Abstract

We propose to measure statistical dependence between two random variables by the *mutual information dimension* (MID), and present a scalable parameter-free estimation method for this task. Supported by sound dimension theory, our method gives an effective solution to the problem of detecting interesting relationships of variables in massive data, which is nowadays a heavily studied topic in many scientific disciplines. Different from classical Pearson's correlation coefficient, MID is zero if and only if two random variables are statistically independent and is translation and scaling invariant. We experimentally show superior performance of MID in detecting various types of relationships in the presence of noise data. Moreover, we illustrate that MID can be effectively used for feature selection in regression.

1 Introduction

How to measure dependence of variables is a classical yet fundamental problem in statistics. Starting with the Galton's work of Pearson's correlation coefficient [Stigler, 1989] for measuring *linear* dependence, many techniques have been proposed, which are of fundamental importance in scientific fields such as physics, chemistry, biology, and economics.

Machine learning and statistics has defined a number of techniques over the last decade which are designed to measure not only linear but also *nonlinear* dependences [Hastie *et al.*, 2009]. Examples include kernel-based [Bach and Jordan, 2003; Gretton *et al.*, 2005], mutual information-based [Kraskov *et al.*, 2004; Steuer *et al.*, 2002], and distance-based [Székely *et al.*, 2007; Székely and Rizzo, 2009] methods. Their main limitation in practice, however, is the lack of scalability or that one has to specify the type of nonlinear relationship one is interested in beforehand, which requires non-trivial parameter selection.

Recently, in *Science*, a distinct method called *maximal information coefficient* (MIC) has been proposed by Reshef *et al.* [2011] (further analyzed in [Reshef *et al.*, 2013]) that measures any kind of relationships between two continuous variables. They use the mutual information obtained by discretizing data and, intuitively, MIC is the maximum mutual information across a set of discretization levels.

However, it has some significant drawbacks: First, MIC depends on the input parameter $B(n)$, which is a natural number specifying the maximum size of a grid used for discretization of data to obtain the entropy [Reshef *et al.*, 2011, SOM 2.2.1]. This means that MIC becomes too small if we choose small $B(n)$ and too large if we choose large $B(n)$. Second, it has high computational cost, as it is exponential with respect to the number of data points[1], and not suitable for large datasets. Third, as pointed out by Simon and Tibshirani [2012], it does not work well for relationship discovery in the presence of noise.

Here we propose to measure dependence between two random variables by the *mutual information dimension*, or MID, to overcome the above drawbacks of MIC and other machine learning based techniques. First, it contains no parameter in theory and the estimation method proposed in this paper is also parameter-free. Second, its estimation is fast; the average-case time complexity is $O(n \log n)$, where n is the number of data points. Third, MID is experimentally shown to be more robust to uniformly distributed noise data than MIC and other methods.

The definition of MID is simple:

$$\mathrm{MID}(X;Y) := \dim X + \dim Y - \dim XY$$

for two random variables X and Y, where $\dim X$ and $\dim Y$ are the *information dimension* of random variables X and Y, respectively, and $\dim XY$ is that of the joint distribution of X and Y. The information dimension is one of the *fractal dimensions* [Ott, 2002] introduced by Rényi [1959; 1970], and its links to information theory were recently studied [Wu and Verdú, 2010; 2011]. Although MID itself is not a new concept, this is the first study that introduces MID as a measure of statistical dependence between random variables;

[1] Since computing the exact MIC is usually infeasible, they used heuristic dynamic programming for efficient approximation.

to date, MID has only been used for chaotic time series analysis [Buzug *et al.*, 1994; Prichard and Theiler, 1995].

MID has desirable properties as a measure of dependence: For every pair of random variables X and Y, (1) $\mathrm{MID}(X;Y) = \mathrm{MID}(Y;X)$ and $0 \leq \mathrm{MID}(X;Y) \leq 1$; (2) $\mathrm{MID}(X;Y) = 0$ if and only if X and Y are statistically independent (Theorem 1); and (3) MID is invariant with respect to translation and scaling (Theorem 3). Furthermore, MID is related to MIC and can be viewed as an extension of it (see Section 2.4).

To estimate MID from a dataset, we construct an efficient *parameter-free* method. Although the general strategy is the same as the standard method used for estimation of the *Box-counting dimension* [Falconer, 2003], we aim to remove all parameters from the method using the *sliding window* strategy, where the width of a window is adaptively determined from the number of data points. The average-case and the worst-case complexities of our method are $O(n \log n)$ and $O(n^2)$ with the number n of data points, respectively, which is much faster than the estimation algorithm of MIC and the other state-of-the-art methods such as the Hilbert-Schmidt independence criterion (HSIC) [Gretton *et al.*, 2005] and the distance correlation [Székely *et al.*, 2007] whose time complexities are $O(n^2)$. Hence MID scales up to massive datasets with millions of data points.

This paper is organized as follows: Section 2 introduces MID and analyzes it theoretically. Section 3 describes a practical estimation method of MID. The experimental results are presented in Section 4, followed by conclusion in Section 5.

2 Mutual Information Dimension

In fractal and chaos theory, *dimension* has a crucial role since it represents the complexity of an object based on a "measurement" of it. We employ the information dimension in this paper, which belongs to a larger family of *fractal dimensions* [Falconer, 2003; Ott, 2002].

In the following, let $\mathbb{N}$ be the set of natural numbers including 0, $\mathbb{Z}$ the set of of integers, and $\mathbb{R}$ the set of real numbers. The base of the logarithm is 2 throughout this paper.

We divide the real line $\mathbb{R}$ into intervals of the same width to obtain the entropy of a discretized variable. Formally,

$$G_k(z) := [z, z+1) \cdot 2^{-k} = \left\{ x \in \mathbb{R} \;\middle|\; \frac{z}{2^k} \leq x < \frac{z+1}{2^k} \right\}$$

for an integer $z \in \mathbb{Z}$. We call the resulting system $\mathcal{G}_k = \{\, G_k(z) \mid z \in \mathbb{Z} \,\}$ the *partition* of $\mathbb{R}$ at *level* k. Partition for the two-dimensional space is constructed from $\mathcal{G}_k$ as $\mathcal{G}_k^2 = \{\, G_k(z_1) \times G_k(z_2) \mid z_1, z_2 \in \mathbb{Z} \,\}$.

2.1 Information Dimension

Given a real-valued random variable X, we construct for each level k the discrete random variable X_k over $\mathbb{Z}$, whose probability is given by $\Pr(X_k = z) = \Pr(X \in G_k(z))$ for each $z \in \mathbb{Z}$. We denote the probability mass function for X_k by $p_k(x) = \Pr(X_k = x)$, and that of the joint probability by $p_k(x, y) = \Pr(X_k = x \text{ and } Y_k = y)$.

We introduce the information dimension, which intuitively shows the complexity of a random variable X as the ratio comparing the change of the entropy to the change in scale.

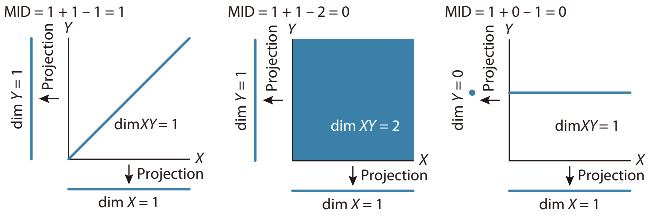

Figure 1: Three intuitive examples. MID is one for linear relationship (left), and MID is zero for independent relationships (center and right).

Definition 1 (Information Dimension [Rényi, 1959]) The *information dimension* of X is defined as

$$\dim X := \lim_{k \to \infty} \frac{H(X_k)}{-\log 2^{-k}} = \lim_{k \to \infty} \frac{H(X_k)}{k},$$

where $H(X_k)$ denotes the *entropy* of X_k, defined by $H(X_k) = -\sum_{x \in \mathbb{Z}} p_k(x) \log p_k(x)$.

The information dimension for a pair of two real-valued variables X and Y is naturally defined as $\dim XY := \lim_{k \to \infty} H(X_k, Y_k)/k$, where $H(X_k, Y_k) = -\sum_{x \in \mathbb{Z}} \sum_{y \in \mathbb{Z}} p_k(x, y) \log p_k(x, y)$, the *joint entropy* of X_k and Y_k. Informally, the information dimension indicates how much a variable fills the space, and this property enables us to measure the statistical dependence. Notice that

$$0 \leq \dim X \leq 1,\ 0 \leq \dim Y \leq 1, \text{ and } 0 \leq \dim XY \leq 2$$

hold since $0 \leq H(X_k) \leq k$ for each k and

$$0 \leq H(X_k, Y_k) \leq H(X_k) + H(Y_k) \leq 2k.$$

In this paper, we always assume that $\dim X$ and $\dim Y$ exist and X and Y are Borel-measurable. Our formulation applies to pairs of continuous random variables X and Y.[2]

2.2 Mutual Information Dimension

Based on the information dimension, the mutual information dimension is defined in an analogous fashion to the mutual information.

Definition 2 (Mutual Information Dimension) For a pair of random variables X and Y, the *mutual information dimension*, or MID, is defined as

$$\mathrm{MID}(X;Y) := \dim X + \dim Y - \dim XY.$$

We can easily check that MID is also defined as

$$\mathrm{MID}(X;Y) = \lim_{k \to \infty} \frac{I(X_k; Y_k)}{k} \tag{1}$$

with the *mutual information* $I(X_k; Y_k)$ of X_k and Y_k defined as $I(X_k; Y_k) = \sum_{x, y \in \mathbb{Z}} p_k(x, y) \log(p_k(x, y)/p_k(x)p_k(y))$.

Informally, the larger $\mathrm{MID}(X;Y)$, the stronger the statistical dependence between X and Y. Figure 1 shows intuitive examples of the information dimension and MID.

[2] Furthermore, it applies to variables with no singular component in terms of the Lebesgue decomposition theorem [Rényi, 1959]. In contrast, Reshef *et al.* [2011] (SOM 6) theoretically analyzed only continuous variables.

2.3 Properties of MID

It is obvious that MID is symmetric $\mathrm{MID}(X;Y) = \mathrm{MID}(Y;X)$ and that $0 \leq \mathrm{MID}(X;Y) \leq 1$ holds for any pair of random variables X and Y, as the mutual information $I(X_k;Y_k)$ in equation (1) is always between 0 and k.

The following is the main theorem in this paper.

Theorem 1 $\mathrm{MID}(X;Y) = 0$ *if and only if* X *and* Y *are statistically independent.*

Proof. ($\Leftarrow$) Assume that X and Y are statistically independent. From equation (1), it directly follows that $\mathrm{MID}(X;Y) = 0$ since $I(X_k;Y_k) = 0$ for all k.

($\Rightarrow$) Assume that X and Y are not statistically independent. We have $\mathrm{MID}(X;Y) = \dim X - \dim X|Y$, where $\dim X|Y = \lim_{k\to\infty} H(X_k|Y_k)/k$ with the conditional entropy $H(Y_k|X_k) = \sum_{x,y\in\mathbb{Z}} p_k(x,y)\log(p_k(x)/p_k(x,y))$. Thus all we have to do is to prove $\dim X \neq \dim X|Y$ to show that $\mathrm{MID}(X;Y) \neq 0$ holds. From the definition of entropy, $\dim X = \lim_{k\to\infty} -\mathbb{E}\log p(X_k)/k$ and $\dim X|Y = \lim_{k\to\infty} -\mathbb{E}\log p(X_k|Y_k)/k$, where $\mathbb{E}$ denotes the expectation. Since $p(X_k)$ and $p(X_k|Y_k)$ go to $p(X)$ and $p(X|Y)$ when $k \to \infty$, $-\mathbb{E}\log p(X_k) > -\mathbb{E}\log p(X_k|Y_k)$ if k is large enough. Thus $\dim X \neq \dim X|Y$ holds. □

Note that $I(X_k;Y_k)$ does not converge to $I(X;Y)$ (and actually goes to infinity) when X and Y are not statistically independent. It is analogue to the entropy $H(X_k) \to \infty$ as $k \to \infty$ and is different from the *differential entropy* of X.

We can also characterize functional relationships with an MID score of one.

Theorem 2 *Let* X *be a random variable with an absolutely continuous distribution. For any function* f *such that* $f(X)$ *also has an absolutely continuous distribution,* $\mathrm{MID}(X;f(X)) = 1$.

Proof. From Theorem 1 in [Rényi, 1959], it follows that $\dim X = \dim f(X) = 1$. Moreover, since $H(X_k) = H(X_k, f(X_k))$, we have $\dim Xf(X) = \dim X = 1$. □

Corollary 1 *Let* X *be a random variable with an absolutely continuous distribution. Given finitely many functions* $f_1, f_2, \ldots, f_m$ *such that each* $f_i(X)$ *has an absolutely continuous distribution. For the multivalued function* F *with* $F(X) = f_i(X)$ $(i \in \{1,2,\ldots,m\})$*,* $\mathrm{MID}(X;F(X)) = 1$.

Ott *et al.* [1984] provide a detailed analysis of the invariance properties of the information dimension. The translation (shift) and scale invariance of MID directly follow.

Theorem 3 *For any* $a, b \in \mathbb{R}$,

$$\mathrm{MID}(X;Y) = \mathrm{MID}(aX;bY) = \mathrm{MID}(X+a;Y+b).$$

2.4 Comparison to MIC

Here we show the relationship between MID and the *maximal information coefficient* (MIC) [Reshef *et al.*, 2011]. Let D be a dataset sampled from a distribution (X,Y), where X and Y are continuous random variables. The dataset D is discretized by a grid G with x rows and y columns, that is, the x-values and the y-values of D are divided into x and y bins, respectively. The probability distribution $D|_G$ is induced by D on the grid G, where the probability mass of each cell is the fraction of data points falling into the cell.

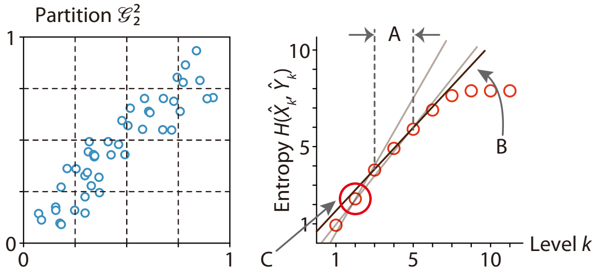

Figure 2: Example of a dataset (left) and illustration of estimation process of $\dim XY$ (right). **A**: The width w used for linear regression is adaptively determined from n. **B**: The best-fitting line. Its slope is the estimator $d(X,Y)$ of $\dim XY$. **C**: $H(\hat{X}_2, \hat{Y}_2)$ from partition on the left.

The *characteristic matrix* $M(D)$ is defined as $M(D)_{x,y} = I^*(D,x,y)/\log\min\{x,y\}$, where $I^*(D,x,y)$ is the maximum of the mutual information $I(D|_G)$ over all grids G of x columns and y rows. Then, MIC is defined as $\mathrm{MIC}(D) = \max_{xy<B(n)} M(D)_{x,y}$. Here, when n goes to infinity, we can easily check that $\mathrm{MID}(X;Y) = \lim_{n,\,x,\,y\to\infty} M(D)_{x,\,y}$ almost surely since $D|_G \to (X,Y)$ as $n,x,y \to \infty$ almost surely. Thus we can say that MIC is the maximum value under the constraint $xy < B(n)$ specified by the user, while MID is the limit of $M(D)$ when $x, y \to \infty$.

This difference provides more power of detection to MID, especially for multivalued functions which are referred to as *non-functional* in [Reshef *et al.*, 2011]. MID is one for such relationships (Corollary 1) as MID focuses on the *ratio* of the changes in the limit, while MIC does not.

3 Estimation of MID

We construct an estimation method for the information dimension and MID on finite samples. While the information dimension itself has been studied from a theoretical perspective [Ott, 2002; Ott *et al.*, 1984; Wu and Verdú, 2010], actual estimation of its value from a finite dataset remains a challenge. For this reason, we design a novel estimation method for the information dimension and MID.

In the following, we always assume that data are in the unit interval $\mathcal{I} = [0,1] \times [0,1]$, which can be achieved by normalization. Since MID is translation and scaling invariant, this transformation has no effect on MID.

3.1 Preliminaries

We modify the definition of $G_k(z)$ as follows: the domain $\mathrm{dom}(G_k) = \{0,1,\ldots,2^k-1\}$, $G_k(z) := [z, z+1) \cdot 2^{-k}$ if $z \in \{0,1,\ldots,2^k-2\}$, and $G_k(z) := [z, z+1] \cdot 2^{-k}$ if $z = 2^k - 1$. Moreover, we define $G_k^2(z_1,z_2) := G_k(z_1) \times G_k(z_2)$ for a pair of integers z_1, z_2. We write $\mathcal{G}_k = \{\, G_k(z) \mid z \in \mathrm{dom}(G_k) \,\}$ and $\mathcal{G}_k^2 = \{\, G_k^2(z_1,z_2) \mid z_1,\, z_2 \in \mathrm{dom}(G_k) \,\}$.

Given a dataset $D = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ sampled from a distribution (X,Y). We use the natural es-

Algorithm 1 Estimation of MID

```
Input: Dataset $D$
Output: Estimator of MID
for each $Z \in \{X, Y, (X, Y)\}$ do
    Compute $H(\hat{Z}_1), H(\hat{Z}_2), \dots$ until each point is isolated;
    Find the best fitted regression line for $k$ v.s. $H(\hat{Z}_k)$ for each window of width $w$ (equation (2)) started from $s \in S_Z(w)$;
    $d(Z) \leftarrow$ the gradient of the obtained line;
end for
Output $d(X) + d(Y) - d(X, Y)$;
```

timator $\hat{X}_k$ of X_k, where the probability is defined as

$$\Pr(\hat{X}_k = z) := \frac{\#\{\, (x, y) \in D \mid x \in G_k(z) \,\}}{n}$$

($\#A$ denotes the number of elements of the set A) for all $z \in \mathrm{dom}(G_k)$, and similarly defined for $\Pr(\hat{Y}_k = z)$ and the joint probability $\Pr(\hat{X}_k = z_1, \hat{Y}_k = z_2)$. This is the fraction of points in D falling into each cell. Trivially, $\hat{X}_k$ (resp. $\hat{Y}_k$) converges to X_k (resp. Y_k) almost surely when $n \to \infty$.

3.2 Estimation via Sliding Windows

Informally, the information dimension $\dim X$ (resp. $\dim Y$, $\dim XY$) can be estimated as the *gradient* of the regression line over k v.s. the entropy $H(\hat{X}_k)$ (resp. $H(\hat{Y}_k)$, $H(\hat{X}_k, \hat{Y}_k)$) since $H(\hat{X}_k) \simeq \dim X \cdot k + c$ such that the difference of the two sides goes to zero as $k \to \infty$. This approach is the same as the standard one for the *box-counting dimension* [Falconer, 2003, Chapter 3]. However, since n is finite, only a finite range of k should be considered to obtain the regression line. In particular, $H(\hat{X}_k)$ is monotonically nondecreasing as k increases and finally it converges to some constant value, where each cell contains at most one data point.

To effectively determine the range of k, we use the *sliding window* strategy to find the best fitted regression line. We set the width w of each window to the maximum value satisfying

$$|\mathcal{G}_w| = 2^w \leq n \text{ and } |\mathcal{G}_w^2| = 4^w \leq n \qquad (2)$$

for estimation of $\dim X$ or $\dim Y$ and $\dim XY$, respectively. If $|\mathcal{G}_w| = 2^w > n$ holds, then there exists a dataset D such that $H(\hat{X}_w) - H(\hat{X}_{w-1}) = 0$, and hence we should not take the point $(w, H(\hat{X}_w))$ into account in the regression for estimating $\dim X$. Hence w in equation (2) gives the *upper bound* of the width of each window that can be effectively applied to *any* dataset in estimation of the dimension.

3.3 MID Estimator

Here we give an estimator of MID using the sliding window strategy. For simplicity, we use the symbol Z to represent the random variables X, Y, or the pair (X, Y).

Let w be the maximum satisfying equation (2) and define

$$S_Z(w) := \left\{ s \in \mathbb{N} \;\middle|\; \begin{array}{l} H(\hat{Z}_{k+1}) - H(\hat{Z}_k) \neq 0 \text{ for any} \\ k \in \{s, s+1, \dots, s+w-1\} \end{array} \right\}.$$

This is the set of starting points of the windows. Note that this set is always finite since $H(\hat{Z}_k)$ converges. We denote the resulting coefficient by $\beta(s)$ and the coefficient of determination by $R^2(s)$ when we apply linear regression to the set of w points $\{\, (s, H(\hat{Z}_s)), (s+1, H(\hat{Z}_{s+1})), \dots, (s+w-1, H(\hat{Z}_{s+w-1})) \,\}$.

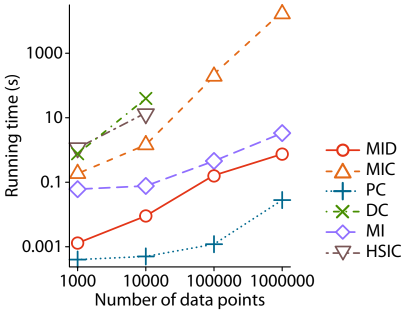

Figure 3: Running time. Note that both axes have logarithmic scales. Points represent averages of 10 trials.

Definition 3 (Estimator of information dimension) Given a dataset D. Define the estimator $d(Z)$ ($Z = X$, Y, or the pair (X, Y)) of the information dimension $\dim Z$ as

$$d(Z) := \beta\left(\arg\max_{s \in S_Z(w)} R^2(s)\right)$$

and define the estimator of MID as

$$\mathrm{MID}(D) := d(X) + d(Y) - d(X, Y).$$

The estimator of MID is obtained by Algorithm 1 and Figure 2 shows an example of estimation. The average-case and the worst-case time complexities are $O(n \log n)$ and $O(n^2)$, respectively, since computation of $H(\hat{Z}_k)$ takes $O(n)$ for each k and it should be repeated $O(\log n)$ and $O(n)$ times in the average and the worst case.

4 Experiments

We evaluate MID experimentally to check its efficiency and effectiveness in detecting various types of relationships, and compare it to other methods including MIC. We use both synthetic data and gene expression data. Moreover, we apply MID to feature selection in nonlinear regression for real data to demonstrate its potential in machine learning applications.

Environment: We used Mac OS X version 10.7.4 with 2 × 3 GHz Quad-Core Intel Xeon CPU and 16 GB of memory. MID was implemented in C[3] and compiled with `gcc` 4.2.1. All experiments were performed in the R environment, version 2.15.1 [R Core Team, 2012].

Comparison partners: We used Pearson's correlation coefficient (PC), the distance correlation (DC) [Székely *et al.*, 2007], mutual information (MI) [Kraskov *et al.*, 2004], the Hilbert-Schmidt independence criterion (HSIC) [Gretton *et al.*, 2005], and MIC (heuristic approximation) [Reshef *et al.*, 2011]. DC was calculated by the R `energy` package; MI and MIC by the official source code[4]; HSIC was implemented in

[3] The source code for MID is available at http://webdav.tuebingen.mpg.de/u/karsten/Forschung/MID

[4] MIC: http://www.exploredata.net, MI: http://www.klab.caltech.edu/~kraskov/MILCA

R (computationally expensive processes were done by efficient packages written in C or Fortran). All parameters in MIC were set to the default values. Gaussian kernels with the width set to the median of the pairwise distances between the samples were used in HSIC, which is a popular heuristics [Gretton *et al.*, 2005; 2008].

4.1 Scalability

We checked the scalability of each method. We used the linear relationship without any noise, and varied the number n of data points from $1,000$ to $1,000,000$.

Results of running time are shown in Figure 3 (we could not calculate DC and HSIC for more than $n = 100,000$ due to their high space complexity $O(n^2)$). These results clearly show that on large datasets, MID is much faster than other methods including MIC which can detect nonlinear dependence. Notice that time complexities of MIC, DC, and HSIC are $O(n^2)$. PC is faster than MID, but unlike MID, it cannot detect nonlinear dependence.

4.2 Effectivity in Measuring Dependence

First, we performed ROC curve analysis using synthetic data to check both precision and recall in detection of various relationship types in the presence of uniformly distributed noise data. We prepared twelve types of relationships shown in Figure 4a, same datasets were used in [Reshef *et al.*, 2011]. For generation of noisy datasets, we fixed the ratio of uniformly distributed noise in a dataset in each experiment, and varied from 0 to 1. Figure 4c shows examples of linear relationships with noise ($n = 300$). For instance, if the noise ratio is 0.5, 150 points come from the linear relationship, and the remaining 150 points are uniformly distributed noise.

We set the number of data points $n = 300$. In each experiment, we generated 1000 datasets, where 500 are sampled from the relationship type with noise (*positive* datasets), and the remaining 500 sets are just noise, that is, composed of statistically independent variables (*negative* datasets). Then, we computed the AUC score from the ranking of such 1000 datasets. Thus both precision and recall are taken into account, while Reshef *et al.* [2011] evaluated only recall.

Figure 4b shows results of ROC curve analysis for each relationship type. We omit results for noise ratio from 0 to 0.4 since all methods except for PC have the maximum AUC in most cases. In more than half of the cases, MID showed the best performance. Specifically, the performance of MID was superior to MIC in all cases except for two sinusoidal types. Moreover, although MI showed better performance than MID in four cases, its performance was much worse than MID and MIC in sinusoidal types. In addition, MI is not normalized, hence it is difficult to illustrate the strength of dependence in the real world situations, as mentioned by Reshef *et al.* [2011]. In contrast to MI, MID showed reasonable performance in all relationship types. Since MID was shown to be much faster than MIC, DC, and MI, these results indicate that MID is the most appropriate among these methods for measuring dependence in massive data. Note that MIC's performance is often worst except for sinusoidal types, which confirms the report by Simon and Tibshirani [2012].

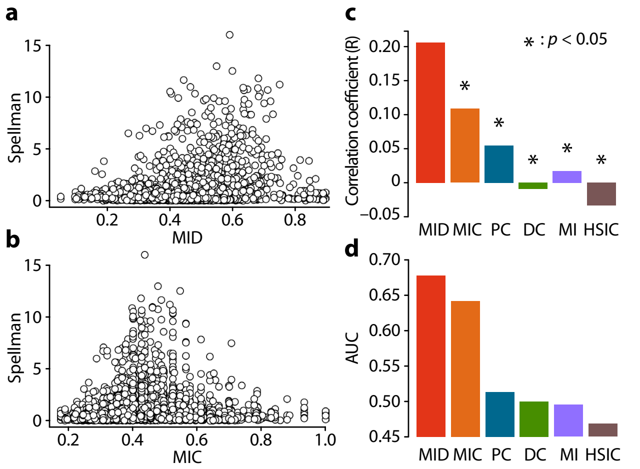

Figure 5: Associations of gene expression data. (**a**, **b**) MID and MIC versus Spellman's score. (**c**) Correlation coefficient (R). (**d**) AUC scores.

Figure 4d is the average of the means of AUC over all relationship types (noise ratio from 0.5 to 0.9 were taken into account). The score of MID was significantly higher than other methods (the Wilcoxon signed-rank test, $\alpha = 0.05$).

Next, we evaluated MID using the cdc15 yeast (*Saccharomyces cerevisiae*) gene expression dataset from [Spellman *et al.*, 1998], which was used in [Reshef *et al.*, 2011] and contains 4381 genes[5]. This dataset was originally analyzed to identify genes whose expression levels oscillate during the cell cycle. We evaluated dependence between each time series against time in the same way as [Reshef *et al.*, 2011, SOM 4.7]. The number $n = 23$ for each gene.

Figure 5a, b illustrate scatter plots of MID and MIC versus Spellman's score. We observe that genes with high Spellman's scores tend to have high MID scores: the correlation coefficient (R) between Spellman's scores and MID was significantly higher than others ($\alpha = 0.05$, Figure 5c).

To check both precision and recall in real data, we again performed ROC curve analysis. That is, we first ranked genes by measuring dependence, followed by computing the AUC score. Genes whose Spellman's scores are more than 1.314 were treated as positives as suggested by Spellman *et al.* [1998]. Figure 5d shows the resulting AUC scores. It confirms that MID is the most effective compared to other methods in measuring dependence. The AUC score of PC is low since it cannot detect nonlinear dependence. DC, MI, and HSIC also show low AUC scores. The reason might be the lack of power for detection due to the small n.

4.3 Effectivity in Feature Selection

Finally, we examined MID as a scoring method for feature selection. To illustrate the quality of feature selection, we performed kNN regression using the selected features to estimate the nonlinear target function, where the average of the k nearest neighbors is calculated as an estimator for each data

[5] http://www.exploredata.net/Downloads/Gene-Expression-Data-Set

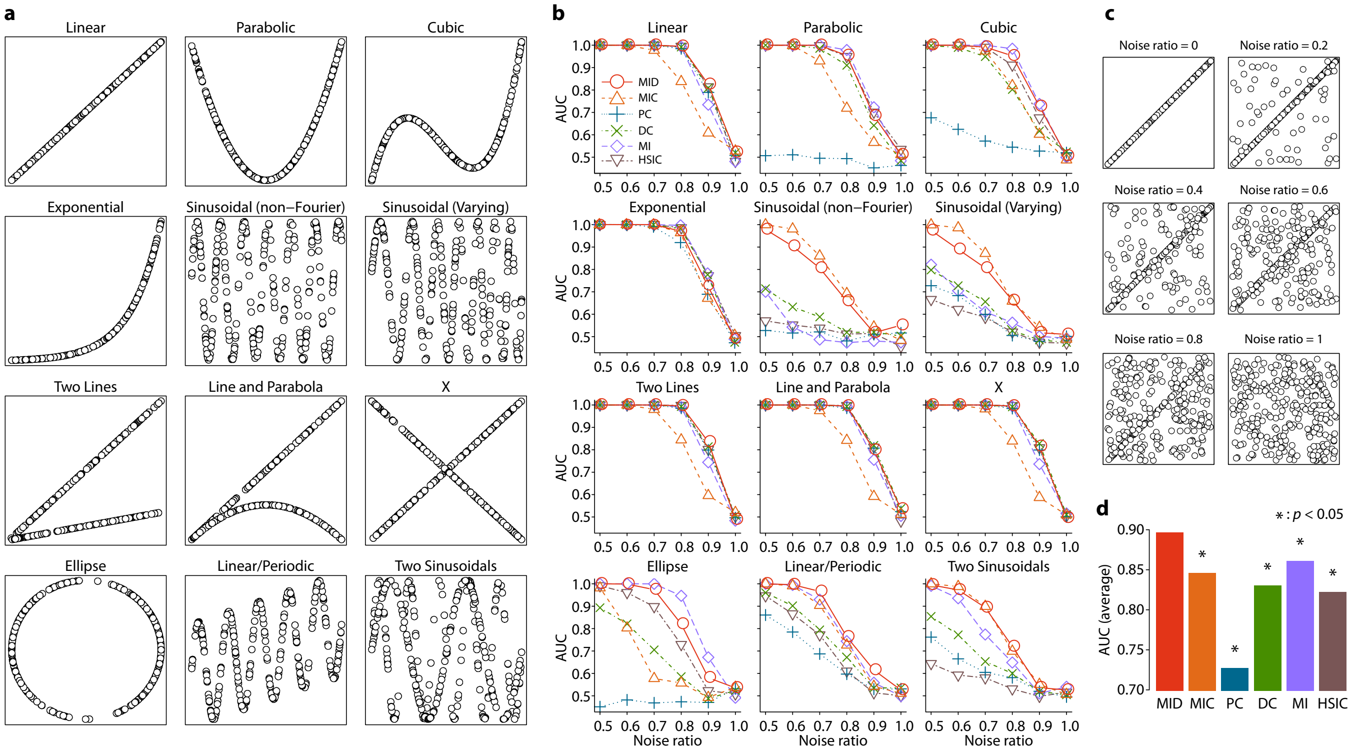

Figure 4: ROC curve analysis for synthetic data. (**a**) Relationship types ($n = 300$ each). (**b**) AUC scores. (**c**) Examples of noisy data for linear relationship ($n = 300$ each). (**d**) The average of the means of AUC over all relationship types.

point (k was set to 3). We measured the quality of prediction by the *mean squared error* (MSE). For each dataset, we generated training and test sets by 10-fold cross validation and performed feature selection using only the training set. Eight real datasets were collected from UCI repository [Frank and Asuncion, 2010] and StatLib[6].

We adopted the standard *filter* method, that is, we measured dependence between each feature and the target variable, and produced the ranking of the features from the training set. We then applied kNN regression (based on the training set) to the test set repeatedly, varying the number of selected top ranked features. Results for each dataset are shown in Figure 6a, and the average R^2, which is the correlation between predicted and target values over all datasets, is summarized in Figure 6b. MID performed significantly better than any other methods (the Wilcoxon signed-rank test, $\alpha = 0.05$).

5 Conclusion

We have developed a new method based on dimension theory for measuring dependence between real-valued random variables in large datasets, the mutual information dimension (MID). MID has desirable properties as a measure of dependence, that is, it is always between zero and one, it is zero if and only if variables are statistically independent, and it is translation and scaling invariant. Moreover, we have constructed an efficient parameter-free estimation method for MID, whose average-case time complexity is $O(n \log n)$. This method has been experimentally shown to be scalable and effective for various types of relationships, and for nonlinear feature selection in regression.

MID overcomes the drawbacks of MIC: (1) MID contains no parameter; (2) MID is fast and scales up to massive datasets; and (3) MID shows superior performance in detecting relationships in the presence of uniformly distributed noise. In [Reshef *et al.*, 2011], an *additive* noise model was considered; MID might not work well for this type of noise model. Since MID is based on dimension theory, it tends to judge that the true distribution is a two-dimensional plane, with no dependence. We will further explore the exciting connection between dimension theory and statistical dependence estimation to address this challenge.

References

[Bach and Jordan, 2003] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003.

[Buzug *et al.*, 1994] Th. Buzug, K. Pawelzik, J. von Stamm, and G. Pfister. Mutual information and global strange attractors in Taylor-Couette flow. *Physica D: Nonlinear Phenomena*, 72(4):343–350, 1994.

[Falconer, 2003] K. Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 2003.

[6] http://lib.stat.cmu.edu/datasets/

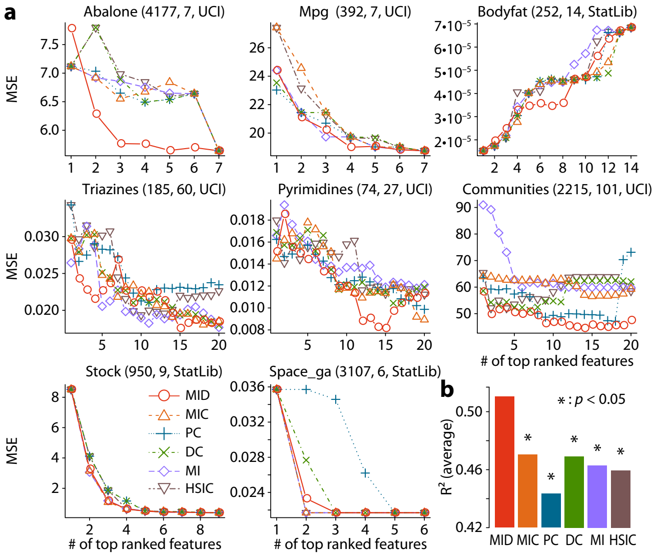

Figure 6: Feature selection for real data. (**a**) MSE (mean squared error) against top ranked features (n, d, and data source are denoted in each parenthesis). For *Triazines*, *Pyrimidines*, and *Communities*, we only illustrate results from the first 20 top ranked features. Points represent averages of 10 trials. (**b**) The average of the means of R^2 over every datasets.

[Frank and Asuncion, 2010] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[Gretton *et al.*, 2005] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 63–77. Springer, 2005.

[Gretton *et al.*, 2008] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, pages 585–592, 2008.

[Hastie *et al.*, 2009] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009.

[Kraskov *et al.*, 2004] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138–1–16, 2004.

[Ott *et al.*, 1984] E. Ott, W. D. Withers, and J. A. Yorke. Is the dimension of chaotic attractors invariant under coordinate changes? *Journal of Statistical Physics*, 36(5):687–697, 1984.

[Ott, 2002] E. Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 2 edition, 2002.

[Prichard and Theiler, 1995] D. Prichard and J. Theiler. Generalized redundancies for time series analysis. *Physica D: Nonlinear Phenomena*, 84(3–4):476–493, 1995.

[R Core Team, 2012] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2012.

[Rényi, 1959] A Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Hungarica*, 10(1–2):193–215, 1959.

[Rényi, 1970] A Rényi. *Probability Theory*. North-Holland Publishing Company and Akadémiai Kiadó, Publishing House of the Hungarian Academy of Sciences, 1970. Republished from Dover in 2007.

[Reshef *et al.*, 2011] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[Reshef *et al.*, 2013] D. N. Reshef, Y. A. Reshef, M. Mitzenmacher, and P. C. Sabeti. Equitability analysis of the maximal information coefficient, with comparisons. *arXiv:1301.6314*, 2013.

[Simon and Tibshirani, 2012] N. Simon and R. Tibshirani. Comment on "Detecting novel associations in large data sets" by Reshef *et al.*, *Science* 2011, 2012. http://www-stat.stanford.edu/~tibs/reshef/comment.pdf.

[Spellman *et al.*, 1998] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[Steuer *et al.*, 2002] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18(supple 2):S231–S240, 2002.

[Stigler, 1989] S. M. Stigler. Francis Galton's account of the invention of correlation. *Statistical Science*, 4(2):73–79, 1989.

[Székely and Rizzo, 2009] G. J. Székely and M. L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236–1265, 2009.

[Székely *et al.*, 2007] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

[Wu and Verdú, 2010] Y. Wu and S. Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Transactions on Information Theory*, 56(8):3721–3748, 2010.

[Wu and Verdú, 2011] Y. Wu and S. Verdú. MMSE dimension. *IEEE Transactions on Information Theory*, 57(8):4857–4879, 2011.