ISSN 1346-5597

NII National Institute of Informatics

NII Technical Report

Measuring Dependency via Intrinsic Dimensionality

Simone Romano, Oussama Chelly, Vinh Nguyen, James Bailey, Michael E. Houle

NII-2016-004E
June 2016

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Measuring Dependency via Intrinsic Dimensionality

Simone Romano*†, Oussama Chelly*, Vinh Nguyen†, James Bailey†, Michael E. Houle*
*National Institute of Informatics, Tokyo, Japan †CIS Department, The University of Melbourne, Australia
Emails: {simone.romano,vinh.nguyen,baileyj}@unimelb.edu.au, {chelly,meh}@nii.ac.jp

***Abstract*—Measuring the amount of dependency among multiple variables is an important task in pattern recognition. In the last few years, many new dependency measures have been developed for the exploration of functional relationships. In this paper, we develop a dependency measure between variables based on an extreme-value theoretic treatment of intrinsic dimensionality. Our measure identifies variables with low intrinsic dimension — that is, those that support embeddings of the data within low-dimensional manifolds. To build a dependency measure on strong foundations, we theoretically prove a connection between information theory and intrinsic dimensionality theory. This allows us also to propose novel estimators of intrinsic dimensionality. Finally, we show that our dependency measure enables to find patterns that cannot be found by other state-of-the-art measures on real and synthetic data.**

I. Introduction

Exploring patterns of dependency among variables is one of the first steps in gaining insights into a new data set. For classic linear dependency between pairs of variables, measures such as the Pearson correlation coefficient are very widely used. For the more general case where the two variables share a non-linear functional relationship, the dependency can be identified according to the recently proposed Maximal Information Coefficient (MIC) [1], where a MIC value of 1 indicates a noiseless functional relationship. A noiseless functional dependency between two variables characterizes their strong dependency. Some examples are: transcript levels of a particular gene that functionally oscillate during the cell cycle determines whether the expression of the gene is dependent upon the cell cycle [1]; socio-economic factors for different countries that are functionally related are clearly also strongly dependent [2].

Over the last few years, several measures of the dependency among multiple variables have been proposed in the literature. Among these, a notable information-theoretic measure of dependency is the Multivariate mAximal Correlation (MAC) score [3]. Another recent measure is the Universal Dependency Score (UDS) [4], which was introduced as a more computationally efficient alternative to MAC. Nonetheless, such measures target only relationships among variables that are strictly functional: much in the same way as MIC does for pairs of variables, MAC and UDS report a score of 1 whenever one of the variables can be expressed as a strict function of the remaining variables.

Despite their importance, functional relationships are not the most general of the interesting relationships that can be determined through dependency analysis — *it is possible for data to be embedded in low-dimensional manifolds without exhibiting a functional relationship*. As an example, in Figure 1 we consider the dependency between two variables X and Y derived from a publicly-available data set consisting of traffic sensor measurements within the city of Melbourne in Australia [5]. Every sensor counts the number of vehicles passing by fixed locations within the road network of Melbourne, binned in intervals of 15 minutes. The two sensors X and Y in Figure 1 are clearly strongly dependent: the increase in traffic at sensor X is associated with an increase in traffic at sensor Y, with the rate of increase depending on the day of the week the measurement was taken. Given that they are well-described by two 1-dimensional manifolds, X and Y must be considered to be strongly dependent; however, the relationship is not functional, and cannot be determined by the current state-of-the-art dependency measures.

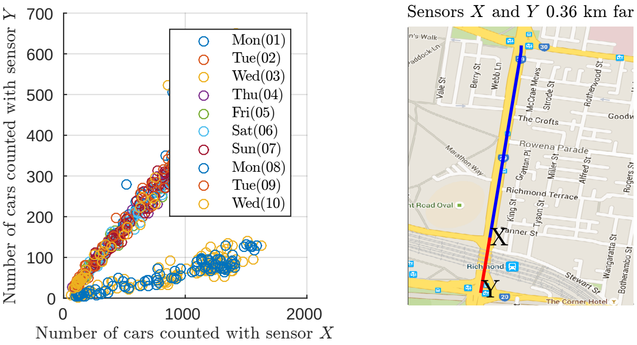

Figure 1. Number of vehicles counted in 15-minute intervals by sensor X, plotted against the number of vehicles counted by sensor Y. The data plotted was collected in Melbourne, Australia during the first ten days of January 2007. Despite the existence of two distinct patterns according to the day on which measurements were made, there is clearly a strong dependency between the two sensors. However, state-of-the-art dependency measures struggle to identify non-functional dependencies of this kind.

For data distributions that can be modelled in terms of low-dimensional manifolds, dependency measures that capture the dimensional characteristics of the data may succeed even when the variables do not admit a functional relationship. The recently-proposed Mutual Information Dimension (MID) [6] was recently proposed to identify dependency relationships between pair of variables. MID is efficient to compute, and it does not require any additional parameters for its computation. Nevertheless, the authors note that the performance of MID may suffer when the data is subjected to additive noise [6]. Moreover, the MID formulation does not naturally extend to the case of multiple variables.

In this paper, we introduce the Intrinsic Dimensional Dependency (IDD) measure for D continuous variables $X = (X_1 \dots X_D)$. IDD has the following features: (i) $0 \leq \mathrm{IDD}(X) \leq 1$; (ii) $\mathrm{IDD}(X) = 0$ iff all variables are independent; (iii) $\mathrm{IDD}(X) = 1$ if the underlying distribution of X is restricted to a constant number of 1-dimensional manifolds. IDD is based on the local intrinsic dimensionality theory proposed in [7], which allows us to overcome the challenges preventing the extension of MID to handle multiple variables. Our contributions include the following:

- identifacation of the connection between information theory and local intrinsic dimensionality theory;
- novel global estimators of dimensionality, and an explanation of their relationship to existing estimators;
- the Intrinsic Dimensional Dependency (IDD) measure for finding variables that embed low dimensional manifolds;
- an experimental demonstration of the ability of IDD to identify novel relationships on real and synthetic data.

The proofs of all theorems and propositions presented in this paper are available in the supplement at https://sites.google.com/site/iddpaper.

II. Related Work

Table I summarizes the features of the dependency measures presented in the introduction. The recently proposed dependency measures capable of scoring functional relationships are based on information theory. MIC [1] targets pairs of variables, whereas MAC [3] and UDS [4] target functional relationships among multiple variables. All these measures are computed using the Shannon entropy obtained by discretizing the continuous variables under analysis. In order to assure that the maximum is achieved at 1 for the case when the relationship is functional, a suitable upper bound is used as a normalization factor. MID is also based on discretization and Shannon entropy estimation for pairs of variables. However, its foundations rely on the dimensionality theory of Rényi [8], who in 1959 was the first to identify the connection between information theory and dimensionality. This connection allows MID to go one step further than functional relationships, in targeting relationships defined by 1-dimensional manifolds over two variables. Although there exist many applications of dimensionality theory to the analysis of chaotic time series [9], to the best of our knowledge, MID is the first application to dependency analysis in pattern recognition.

To date there exist many different measures of dimensionality — see [10] for a recent survey. One of the most popular measures, the correlation dimension of Grassberger and Procaccia [11], has been shown to be related to an entropic measure, the α-Rényi entropy [12], within the research literature on chaotic time series [9], [13]. In this paper, we show that the correlation dimension and Rényi entropy can be expressed in terms of the local measure of intrinsic dimensionality introduced in [7]. This intriguing link allows us to build a dependency measure among multiple variables based on information theory and dimensionality theory.

Table I

Features of dependency measures: handling functional relationships, manifold relationships, and multiple variables.

	Dependency Measure	Func.	Manifold	Multiple
MIC	Maximal Information Coefficient	✓		
MAC	Multivariate mAximal Correlation	✓		✓
UDS	Universal Dependency Score	✓		✓
MID	Mutual Information Dimension	✓	✓	
IDD	Intrinsic Dimensional Dependency	✓	✓	✓

III. Intrinsic Dimensionality Theory

Intrinsic dimensionality theory studies the expressibility of high-dimensional data in terms of a small number of latent variables, such as those needed to describe a manifold of low dimension [7], [8], [10]. Let $X = (X_1 \dots X_D)$ be D continuous variables that represent a data set $\{x_i\}_{i=1\dots n}$ of n data points. We begin our discussion with the α-Rényi dimension, defined as follows:

$$\dim_\alpha(X) \triangleq \lim_{\delta \to 0^+} \frac{H_\alpha(X,\delta)}{\log 1/\delta}, \tag{1}$$

where $H_\alpha(X,\delta) \triangleq \frac{1}{1-\alpha} \log \left(\sum_{\delta-\text{boxes}} p(x,\delta)^\alpha \right)$ is the α-Rényi entropy [12]. Here, $p(x,\delta)$ is the estimated probability mass function of the discretized variable X using boxes of size δ. Intuitively, the α-Rényi entropy quantifies the space-filling capacity of the data, and $\dim_\alpha(X)$ measures its growth rate: the smaller the growth rate of the entropy, the smaller the dimensionality of the manifold that embeds X. Interestingly, if α is allowed to tend to 1, the α-Rényi entropy tends to the Shannon entropy $H(X,\delta) = -\sum p(x,\delta) \log p(x,\delta)$, in which case $\dim_\alpha(X)$ tends to the *information dimension*:

$$\dim(X) \triangleq \lim_{\alpha \to 1} \dim_\alpha(X) = \lim_{\delta \to 0^+} \frac{H(X,\delta)}{\log 1/\delta}. \tag{2}$$

For the case when $\alpha = 2$, the α-Rényi dimension reduces to another well-known measure of dimensionality, the *correlation dimension*. For this case, it is convenient to use an alternative definition of $\dim_\alpha(X)$ in terms of the generalized correlation integral [13]:

$$\dim_\alpha(X) \triangleq \lim_{r \to 0^+} \frac{\log C_\alpha(X,r)}{\log r}, \tag{3}$$

where

$$C_\alpha(X,r) \triangleq \left(\int \left(\int f(y) \mathbb{1}(x,y,r) \, \mathrm{d}y \right)^{\alpha-1} f(x) \, \mathrm{d}x \right)^{\frac{1}{\alpha-1}}. \tag{4}$$

Here, $f(x)$ is the p.d.f. of X, and $\mathbb{1}(x,y,r) = \mathbb{1}(\|x-y\| < r)$ is the indicator function activated when the Euclidean distance between x and y is smaller than the specified radius r. When $\alpha = 2$, the correlation integral $C_2(X,r) = \iint f(y)f(x)\mathbb{1}(x,y,r) \, \mathrm{d}y \, \mathrm{d}x$ [11] has an intuitive interpretation: $C_2(X,r)$ is the probability of finding two points at distance less than r in the support of X.

We have just seen that a single generalized measure the α-Rényi dimension, is capable of generalizing two independently-proposed measures of dimensionality: the information dimension and the correlation dimension. In the next section, we will show that there exist also connections between the α-Rényi dimension and other measures of dimensionality recently proposed in the pattern recognition literature.

A. Local Intrinsic Dimensionality and the α-Rényi Dimension

The Local Intrinsic Dimensionality (Local ID, or LID) measure [7], original proposed for applications in the data mining community, has recently been shown to have deep connections to the statistical theory of extreme values [14],

[15]. Given a data point x, the local ID at x at distance r is defined as:

$$\mathrm{ID}(x,r) \triangleq \lim_{\epsilon\to 0^+} \frac{\log F_R\big(x,(1+\epsilon)r\big) - \log F_R(x,r)}{\log(1+\epsilon)} \tag{5}$$
$$= \frac{r f_R(x,r)}{F_R(x,r)},$$

where $R \geq 0$ is a continuous random variable denoting the distance from x to other data points, $F_R(x,r)$ is its c.d.f. and $f_R(x,r)$ its p.d.f. The *local intrinsic dimension* at x is in turn defined as the limit as the radius tends to zero:

$$\mathrm{ID}(x) = \lim_{r\to 0^+} \mathrm{ID}(x,r).$$

$\mathrm{ID}(x)$ measures the dimensionality of the manifold embedded in X at the locality x. Here we prove that this local measure of dimensionality contributes to the global measure of dimensionality $\dim_\alpha(X)$ according to the following relationship:

Theorem 1. *Let X be a set of D continuous variables, $f(x)$ the p.d.f. of the distribution from which X is drawn, and $\mathrm{ID}(x)$ the local intrinsic dimension at the locality x. The α-Rényi dimension can be expressed as*

$$\dim_\alpha(X) = \frac{\int f^\alpha(x)\,\mathrm{ID}(x)\,\mathrm{d}x}{\int f^\alpha(x)\,\mathrm{d}x}.$$

The result above leads to interesting characterizations of the information dimension and correlation dimension:

- The *information dimension*, defined in Equation (2) in terms of the Shannon entropy, can also be computed as the expectation of the local ID. That is, $\dim_\alpha(X)$ for $\alpha = 1$ is equal to:
$$\dim(X) = \int f(x)\,\mathrm{ID}(x)\,\mathrm{d}x;$$
- The *correlation dimension* $\dim_2(X)$ is the expected local ID with respect to a distribution whose p.d.f. is the normalized square of the original p.d.f.:
$$\dim_2(X) = \frac{\int f(x)^2 \mathrm{ID}(x)\,\mathrm{d}x}{\int f(x)^2\,\mathrm{d}x}.$$

In addition to linking dimensionality theory, information theory, and local intrinsic dimensionality theory, the equations above allows us to propose novel estimators of dimensionality.

B. Estimation of Dimensionality using Local Estimators

In [14], several estimators of local ID have been proposed; among these, the Maximum Likelihood Estimator (MLE) exhibited a useful trade-off between statistical efficiency and complexity. Given a locality x, the MLE estimator of ID makes use of the distances between the first k-th nearest neighbors and x: :

$$\widehat{\mathrm{ID}}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log\frac{d_i(x)}{d_k(x)}\right)^{-1}, \tag{6}$$

where $d_i(x)$ denotes the Euclidean distance between x and its i-th nearest neighbor (NN). The MLE estimator of ID can be derived from the statistical theory of extreme values, wherein the smallest kNN distances can be seen as extreme events of the underline distribution of distances. The MLE estimator is indeed equivalent to the established Hill estimator for power-law distributions [16].

Using the MLE estimator of local ID, it is possible to obtain estimates for $\dim_\alpha(X)$ as defined in Theorem 1.

Theorem 2. *The kNN estimator of $\dim_\alpha(X)$ is:*

$$\widehat{\dim_\alpha}(X) = \frac{\sum_{i=1}^{n} \widehat{\mathrm{ID}}(x_i)(d_k(x_i)^{-D})^{\alpha-1}}{\sum_{i=1}^{n}(d_k(x_i)^{-D})^{\alpha-1}}.$$

The formula allows the estimation of the information dimension $\dim(X)$ and the correlation dimension $\dim_2(X)$ when $\alpha = 1$ and $\alpha = 2$, respectively. Interestingly, the information dimension can simply be computed as the average local ID estimated over each of the n data samples:

$$\widehat{\dim}(X) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathrm{ID}}(x_i) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k}\sum_{i=1}^{k}\ln\frac{d_i(x)}{d_k(x)}\right)^{-1}. \tag{7}$$

It is intriguing to note that the formula above has already been used as measure of global dimensionality [14], [17]. However, in this paper we have demonstrated for the first time that the average of local ID estimates across data samples is a theoretically sound estimator of the information dimension.

IV. The Intrinsic Dimensional Dependency

In this section, we define a dependency measure for the variables X which is maximized when the intrinsic dimensionality $\dim(X)$ is small. Intuitively, our measure can be interpreted as a divergence between the data representational dimension $\sum_{i=1}^{D}\dim(X_i)$ and its intrinsic dimensionality $\dim(X)$. Moreover, for the sake of interpretability, the measure is normalized using a suitable upper bound. in order to produce values in the range $[0,1]$.

Definition 1. *The Intrinsic Dimensional Dependency for X:*

$$\mathrm{IDD}(X) \triangleq \frac{\sum_{i=1}^{D}\dim(X_i) - \dim(X)}{\sum_{i=1}^{D}\dim(X_i) - \max_i \dim(X_i)}.$$

IDD has the following properties:

Proposition 1. *Let X be a set of D continuous variables:*

1) $0 \leq \mathrm{IDD}(X) \leq 1$;
2) $\mathrm{IDD}(X) = 0$ *iff all X_i are independent;*
3) $\mathrm{IDD}(X) = 1$ *if there exist one or more manifolds of dimension 1 whose union embeds X;*
4) $\mathrm{IDD}(X) = 1$ *if there exists $1 \leq i \leq D$ such that for all $j \neq i$, X_j is a a function or multivalued function of X_i.*

We choose to base IDD on the information dimension $\dim(X)$, so as to exploit the useful properties of the Shannon entropy which do not hold true for the α-Rényi entropy [18]. IDD can also be seen as the normalized extension to multiple variables of $\mathrm{MID}(X,Y) = \dim(X) + \dim(Y) - \dim(X,Y)$.

In order to obtain an estimator for IDD, we make use the Equation (7) for the estimation of $\dim(X)$. However, in order to make IDD *invariant to the distribution of the marginals in X* we carry out a copula transformation [19] in which for each variable, the raw value of a data point is substituted with its rank. In practice, to avoid numerical instability due to very small distance values, we add a very small amount of noise ε to X [20]. Furthermore, to decrease the computational time when the number of variables is small, we build a KD-tree for faster distance computation. These steps are summarized in Algorithm 1. If D is small (for example, $D < 10$), the average computational complexity of IDD is in $O(Dn \log n + nk \log n)$. On the other hand, as D increases the computational cost tends to the worst-case complexity, $O(Dn \log n + nkn)$. For the applications of dependency that are of typical interest, D is usually small — indeed, only for a small number of variables can the relationships be explained and interpreted by the user [21].

Algorithm 1 Estimation of $\mathrm{IDD}(X)$ based on kNN sets.

```
IDD(X, k)
Copula transform $X$
$X = X + \varepsilon$, where $\varepsilon = 10^{-6}$ Gaussian noise
Build KD-trees for $X$ and $X_i$
Compute $\widehat{\dim}(X)$ and $\widehat{\dim}(X_i)$ according to Equation (7).
return $\mathrm{IDD}(X)$ as per Definition 1
```

V. Experimental Evaluation

In this section we experimentally evaluate our estimators, as well as the performance of IDD as a dependency measure. In Section V-A we discuss the characteristics of our estimators of dimensionality, and in Section V-B we evaluate the performance of IDD on synthetic and real data sets. All code will be made publicly available at https://google.sites.com/site/iddpaper.

A. *Experiments about Estimation of Dimensionality*

In this section, we compare the estimator of dimensionality $\widehat{\dim}_\alpha$ against other previously proposed measures of dimensionality. In particular, we compare it against the Grassberger-Procaccia (GP) [11], Hein, and Takens estimators of dimensionality [22]. Our target is to gain insights on the different characteristics of $\widehat{\dim}_\alpha$ for different choices of the parameter α. Indeed, in-depth studies of the performance of the estimator of information dimension $\widehat{\dim}_1$ can be found in [14], [17].

1) Synthetic Data Sets: Measures of dimensionality are usually tested on synthetic data sets for which the true dimension of the manifold embedded in X is known. Here we use the same data sets proposed in [22]. We test our kNN estimator for $\widehat{\dim}_\alpha(X)$, which is solely based on distance computation, and hence can be computed efficiently using a KD-Tree. This estimator must satisfy the usual extreme-value-theoretic assumptions, in that the neighborhood size k should be small relative to the total number of points n [14]; for these sets, we therefore fix $k = 100$.

Table II shows the average value of the different estimators of dimensionality for a collection of synthetic data sets of two different sizes, $n = 1000$ and $n = 10000$. The performance of $\widehat{\dim}_\alpha$ is comparable to those of other previously proposed measures, and is particularly good when the data size is larger. We also note a slight tendency for $\widehat{\dim}_\alpha$ to decrease as α increases; however, the cause is not completely evident from this list of synthetically crafted data sets. We next identify a scenario for which α can significantly impact the estimation.

Table II
Synthetic data sets with known true dimensionality and estimators. As an extreme-value-theoretic estimator, $\widehat{\dim}_\alpha$ is seen to perform better on larger data sets, although there is a slight tendancy for it to decrease as α increases.

n	Data Set	D	Dim.	Hein	GP	Tak.	$\widehat{\dim}_1$	$\widehat{\dim}_{1.5}$	$\widehat{\dim}_2$
1000	Sphere	11	10	9	8.83	9.19	8.18	8.18	8.18
	Dense	6	4	4	3.61	3.63	3.6	3.54	3.49
	Swiss roll	3	2	2	1.95	1.94	2.49	2.36	2.24
	Moebius	3	2	2	1.98	1.99	2.52	2.46	2.42
10000	Sphere	11	10	9.9	9.53	9.59	9.12	9.12	9.12
	Dense	6	4	4	3.72	3.77	3.89	3.86	3.79
	Swiss roll	3	2	2	2.01	2.01	1.98	1.99	1.99
	Moebius	3	2	2	2	1.99	2.01	2.03	2.04

2) Bigger α Decreases the Contribution of Noisy Points: For real data, even when X is well described by a low-dimensional manifold, there typically exist a substantial proportion of noise points lying far from the manifold. If their neighborhood includes many points in the vicinity of the manifold, noise points can be expected to have very high estimated ID values. The information dimension $\dim(X)$ estimated as per Equation 7 weights every local $\mathrm{ID}(x)$ contribution equally. On the other hand, when we estimate $\widehat{\dim}_\alpha$, the $\mathrm{ID}(x)$ contribution is penalized by a factor proportional to $\frac{1}{d_k(x)^{D(\alpha-1)}}$; in this case, the larger the value α, the further the k-th nearest neighbor is to x, and the more the contribution of $\mathrm{ID}(x)$ is penalized.

Figure 2 illustrates how $\widehat{\dim}_\alpha$ can decrease as α increases. The noise points (those far from the elliptical manifold) have high estimated ID, and their contributions are greatly penalized when α is large. In this example, the estimated dimensionality of the ideal manifold is 1.

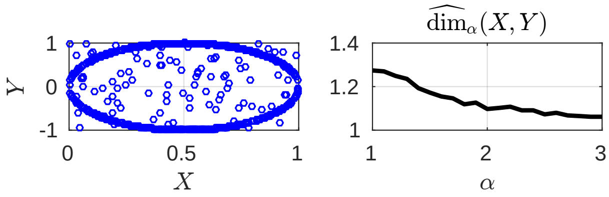

Figure 2. Estimates of $\widehat{\dim}_\alpha$ as α is varied: the larger the value of α, the greater the penalty on the contributions of noise points.

A larger choice of α therefore yields less sensitivity to noise. When designing a dependency measure between variables, a noiseless relationship should not achieve the same score as a noisy relationship. For this reason, IDD is proposed in terms of the information dimension $\widehat{\dim}_1$.

B. *Experiments about Dependency between Variables*

In this section we carry out experiments on IDD computed as per Algorithm 1, on synthetic and real data sets.

1) Choosing k for IDD: Here we discuss the sensitivity of IDD with regards to its parameter k. In Figure 3 we show the average value of IDD at the variation of k for one noiseless (red) and one noisy (blue) relationship among $n = 1000$ points. For the noiseless relationship, the value of IDD is close to 1 for a wide range of values k. Nonetheless, IDD for the noisy relationship is close to 0 if we choose k to be small. In general, if the chosen locality size k is too small, the global relationship between variables may not be detectable by IDD; k should therefore be chosen with care. Even though the best choice of k varies from data set to data set, in this work we chose to fix $k \approx n/4$, so as to avoid focusing on localities that are too small. MID, the other dependency measure based on dimensionality discussed in this paper, has the same issue in that it focuses on very small localities. Nonetheless, this issue cannot be solved straightforwardly because MID does not allow for tuning of a neighborhood size parameter.

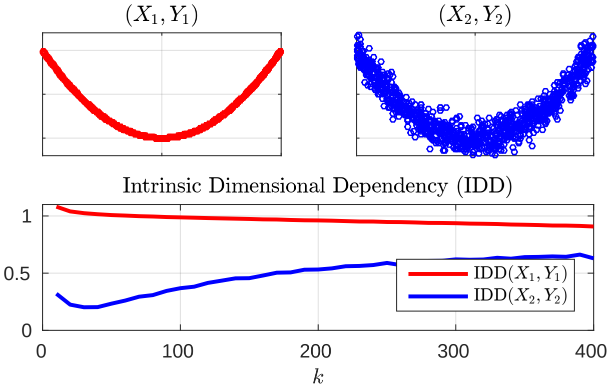

Figure 3. The performance of IDD as k is varied, for one noiseless (red) and one noisy (blue) relationship among $n = 1000$ points. If k is chosen too small, no relationship will be identified even if there exists only a small amount of noise. In this work we chose $k \approx n/4$.

2) Synthetic Relationships: In this section we crafted different relationships among the D variables X, in order to experimentally demonstrate the properties of IDD in Proposition 1. We tested IDD as well as the two other measures capable of handling multiple variables, MAC and UDS — for these two competitors, all parameters were set to their default values [3], [4]. Figure 4 shows the average value of the analyzed measures on the following relationships induced on $n = 1000$ points:

Rel. A: All variables are identical, and hence functionally related. X_1 is a uniform variable in $[0, 1]$ and $X_j = X_1$ for all $j = 2 \ldots D$. IDD and MAC show the desired behavior: their score is always 1. On the other hand, the UDS value is significantly below 1 and increases only slightly as the number of variables D increases.

Rel. B: The relationship is multivalued functional. X_1 is uniform, $X_2 = \{X_1, 20X_1\}$, and $X_j = X_2$ for all $j = 3 \ldots D$. Such scenarios occur whenever there is a latent categorical variable that determines the different trends (as in the example in Figure 1). IDD consistently scores this relationship with the value 1. MAC scores this dependency with 1 only when D is sufficiently large, whereas UDS shows quite different values for this type of relationship as D increases.

Rel. C: There is a functional relationship between one variable and the remaining variables: $X_1 = \left(\frac{1}{D-1}\sum_{j=2}^{D} X_j\right)^2$ where X_j are uniform variables. We observe that as the total number of variables D increases, the strength of the overall dependency in X should decrease. In addition, when $D = 2$, the relationship $X_1 = (X_2)^2$ is functional, and should be scored with value 1. This indeed is the behavior of IDD. On the other hand, MAC does not always decrease when D increases, although it is equal to 1 when $D = 2$. The performance of UDS is even worse, as it is equal to 0 when $D > 2$.

Rel. D: All variables are independent. In this case all measures should be equal to 0. IDD in this case shows a slightly increasing baseline value that can be explained by the estimation bias in $\widehat{\dim}(X)$. Nonetheless, its increasing trend is much weaker than that of MAC trend. Furthermore, IDD could be possibly adjusted for better performance using techniques recently proposed in the literature [2]. UDS is the best in this scenario: it is identically equal to 1 for any choice of D.

3) Real Data Sets: Here we explore the relationships between different traffic sensors (variables) in the city of Melbourne using data from [5]: each of the 1084 sensors disseminated in the city counts the number of vehicles passing by in 15 minute intervals. We focus on the first 10 days of the year 2007, which yields a total of $n = 960$ data points for each sensor. Moreover given that our focus is on continuous variables, we consider only sensors for which at least half the produced values were unique. IDD allows us to identify multivalued functional relationships where the trend depends on the particular day of the week considered. An example of such a relationship is given in Figure 1. In order to compare dependency measures for pairs of variables, such as MIC and MID, we perform the following experiment: we identify the top 100 dependent pairs of sensors according to the given dependency measure. *Due to the nature of traffic flow, the top 100 pairs identified by the dependency measure should consist of sensors that are geographically close.* Table III shows the average distance in kilometers among the top 100 sensor pairs, according to each of the dependency measures tested. As it is able to identify multivalued functional relationships, the distance averaged over the top 100 sensors according to IDD is smaller than that computed according to other measures.

Table III
DISTANCE IN KM FOR THE TOP 100 PAIRS OF DEPENDENT SENSORS: SENSORS CLOSE TO EACH OTHERS ARE DEPENDENT AND IDD IDENTIFIES MORE OF SUCH SENSORS.

IDD	MID	MIC	MAC	UDS
6.6 ± 5.5	7.1 ± 5.0	7.1 ± 5.5	7.4 ± 5.6	7.5 ± 5.4

We also collected data on the daily energy consumption for 12 different buildings at the University of Melbourne over the year 2013. We paired each daily value with the maximum and minimum temperatures registered in Melbourne, as retrieved from http://www.bom.gov.au/. Some buildings show a higher energy consumption when the outdoor temperature is high. Our target is to identify the building which is most dependent on the outdoor temperature. Figure 5 shows the top depen-

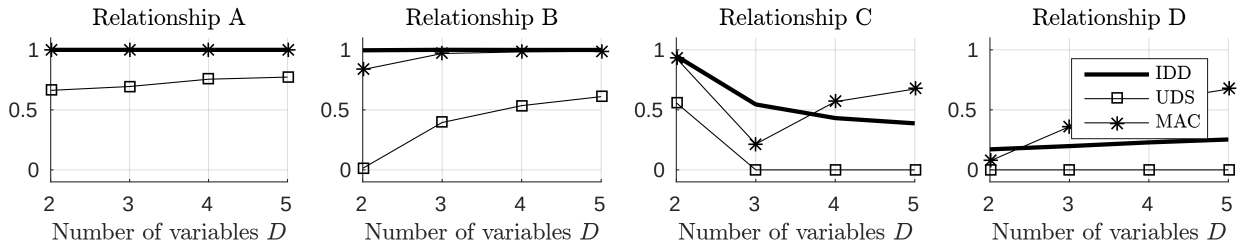

Figure 4. Synthetic relationships. IDD is always in $[0, 1]$; moreover, it is equal to 1 for functional and multivalued functional relationships (Rel. A and B).

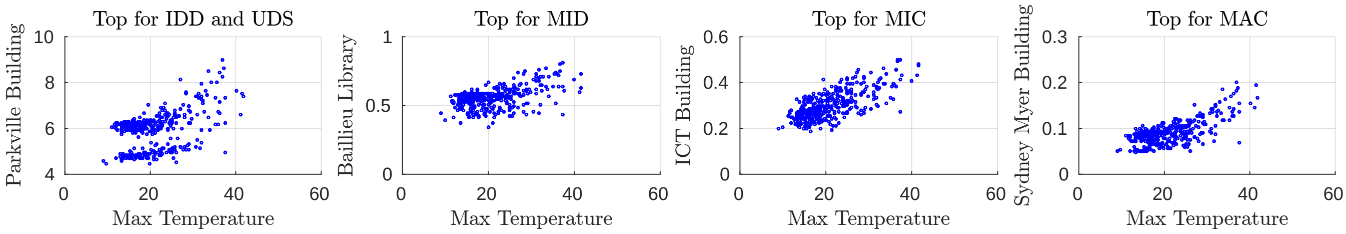

Figure 5. Energy consumption in MWh for the buildings at the University of Melbourne that are most dependent on outdoor temperature (measured in C°). The top building is the one whose consumption ranked highest in terms of its dependency on temperature, according to the stated dependency measure. IDD allows the identification of the double trend of the 'Parkville Building' due to the two different regimes of the cooling system. The other measures see this double trend as noise.

dent building according to different dependency measures. IDD scores the 'Parkville Building' as the most dependent on temperature, even though the relationship is multivalued functional. There indeed exist two sharp increasing trends with regards to temperature, corresponding to two different functioning regimes of the cooling system. The other measures do not identify this situation as a multivalued trend — they simply reject it as noise. UDS shares the same top scoring relationship with IDD; however this behavior is not to be expected in general, given that UDS targets only relationships that are strictly functional.

VI. Conclusion

In this paper, we presented a new dependency measure between multiple continuous variables based on dimensionality theory, the Intrinsic Dimensional Dependency (IDD). IDD is computed on top of novel estimators of dimensionality that we presented and experimentally validated in this paper. IDD is capable of identifying variables that are well-described by low-dimensional manifolds. This type of dependency between variables is not targeted by previously existing methods. IDD thus has the potential to become a useful tool for analysts to explore patterns of relationships in data.

References

[1] 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*, 2011.

[2] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoor, "A framework to adjust dependency measure estimates for chance," *arXiv preprint arXiv:1510.07786*, 2015.

[3] H. V. Nguyen, E. Müller, J. Vreeken, P. Efros, and K. Böhm, "Multivariate maximal correlation analysis," in *ICML*, 2014, pp. 775–783.

[4] H.-V. Nguyen, P. Mandros, and J. Vreeken, "Universal dependency analysis," in *Proc. SIAM International Conference on Data Mining (SDM)*, 2016, p. *to appear*.

[5] F. Schimbinschi, X. V. Nguyen, J. Bailey, C. Leckie, H. Vu, and R. Kotagiri, "Traffic forecasting in complex urban networks: Leveraging big data and machine learning," in *IEEE Big Data*, 2015.

[6] M. Sugiyama and K. M. Borgwardt, "Measuring statistical dependence via the mutual information dimension," in *IJCAI*, 2013.

[7] M. E. Houle, "Dimensionality, discriminability, density and distance distributions," in *Data Mining Workshops (ICDMW)*, 2013.

[8] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 1-2, pp. 193–215, 1959.

[9] D. Prichard and J. Theiler, "Generalized redundancies for time series analysis," *Physica D: Nonlinear Phenomena*, vol. 84, no. 3, pp. 476–493, 1995.

[10] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," *Information Sciences*, vol. 328, pp. 26–41, 2016.

[11] P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Physical review letters*, vol. 50, no. 5, p. 346, 1983.

[12] A. Renyi, "On measures of entropy and information," 1961.

[13] C. Diks and S. Manzan, "Tests for serial independence and linearity based on correlation integrals," *Studies in Nonlinear Dynamics & Econometrics*, vol. 6, no. 2, 2002.

[14] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K.-i. Kawarabayashi, and M. Nett, "Estimating local intrinsic dimensionality," in *SIGKDD*. ACM, 2015, pp. 29–38.

[15] M. E. Houle, "Inlierness, outlierness, hubness and discriminability: an extreme-value-theoretic foundation," NII, Technical Report NII-2015-002E, Mar 2015.

[16] B. M. Hill *et al.*, "A simple general approach to inference about the tail of a distribution," *The annals of statistics*, vol. 3, pp. 1163–1174, 1975.

[17] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *NIPS*, 2004, pp. 777–784.

[18] A. Teixeira, A. Matos, and L. Antunes, "Conditional rényi entropies," *Information Theory, IEEE Transactions on*, vol. 58, 2012.

[19] B. Póczos, Z. Ghahramani, and J. Schneider, "Copula-based kernel dependency measures," *arXiv preprint arXiv:1206.4682*, 2012.

[20] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[21] N. X. Vinh, J. Chan, S. Romano, J. Bailey, C. Leckie, K. Ramamohanarao, and J. Pei, "Discovering outlying aspects in large datasets," *Data Mining and Knowledge Discovery*, pp. 1–36, 2016.

[22] M. Hein and J.-Y. Audibert, "Intrinsic dimensionality estimation of submanifolds in r d," in *ICML*. ACM, 2005, pp. 289–296.

Measuring Dependency via Intrinsic Dimensionality

SUPPLEMENTARY MATERIAL

PROOFS OF PROPOSITIONS AND THEOREMS

Theorem 1. *Let X be a set of D continuous variables, $f(x)$ the p.d.f. of the distribution from which X is drawn, and $\mathrm{ID}(x)$ the local intrinsic dimension at the locality x. The α-Rényi dimension can be expressed as*

$$\dim_\alpha(X) = \frac{\int f^\alpha(x)\,\mathrm{ID}(x)\,\mathrm{d}x}{\int f^\alpha(x)\,\mathrm{d}x}.$$

Proof. We first note that the following holds true for the generalized correlation integral in Equation (4):

$$\begin{aligned} C_\alpha(X,r) &= \left(\int \left(\int f(y)\mathbb{1}(x,y,r)\,\mathrm{d}y\right)^{\alpha-1} f(x)\,\mathrm{d}x\right)^{\frac{1}{\alpha-1}} \\ &= \left(\int F_R^{\alpha-1}(x,r) f(x)\,\mathrm{d}x\right)^{\frac{1}{\alpha-1}}, \end{aligned}$$

where $F_R(x,r) = \int f(y)\mathbb{1}(x,y,r)\,\mathrm{d}y$ is the number of points at distance smaller than r from x. We then use l'Hôpital's rule on the definition of $\dim_\alpha(X)$ in Equation (3):

$$\begin{aligned} \dim_\alpha(X) &= \lim_{r\to 0^+} \frac{\log\left(\int F_R^{\alpha-1}(x,r)f(x)\,\mathrm{d}x\right)}{(\alpha-1)\log r} \\ &\overset{\text{H.}}{=} \lim_{r\to 0^+} \frac{r\int(\alpha-1)F_R^{\alpha-2}(x,r)f_R(x,r)f(x)\,\mathrm{d}x}{(\alpha-1)\int F_R^{\alpha-1}(x,r)f(x)\,\mathrm{d}x} \\ &= \lim_{r\to 0^+} \frac{\int F_R^{\alpha-1}(x,r)\frac{rf_R(x,r)}{F_R(x,r)}f(x)\,\mathrm{d}x}{\int F_R^{\alpha-1}(x,r)f(x)\,\mathrm{d}x} \\ &= \lim_{r\to 0^+} \frac{\int F_R^{\alpha-1}(x,r)\mathrm{ID}(x,r)f(x)\,\mathrm{d}x}{\int F_R^{\alpha-1}(x,r)f(x)\,\mathrm{d}x}. \end{aligned}$$

As r tends to 0^+, $F_R(x,r)$ tends to $f(x)$. Therefore:

$$\dim_\alpha(X) = \frac{\int f^\alpha(x)\mathrm{ID}(x)\,\mathrm{d}x}{\int f^\alpha(x)\,\mathrm{d}x}$$

□

Theorem 2. *The kNN estimator of* $\dim_\alpha(X)$ *is:*

$$\widehat{\dim}_\alpha(X) = \frac{\sum_{i=1}^n \widehat{\mathrm{ID}}(x_i)(d_k(x_i)^{-D})^{\alpha-1}}{\sum_{i=1}^n (d_k(x_i)^{-D})^{\alpha-1}}.$$

Proof. We first prove a more general result: if $K(\cdot)$ is a kernel function with width h, then for $\alpha \geq 1$,

$$\widehat{\dim}_\alpha(X) = \frac{\sum_{i=1}^n \widehat{\mathrm{ID}}(x_i)\Big(\sum_{j=1}^n K(\|x_i - x_j\|, h)\Big)^{\alpha-1}}{\sum_{i=1}^n \Big(\sum_{j=1}^n K(\|x_i - x_j\|, h)\Big)^{\alpha-1}}.$$

To prove this, note that for $\alpha \geq 1$, $\dim_\alpha(X) = \frac{\int f(x)f(x)^{\alpha-1}\mathrm{ID}(x)\,\mathrm{d}x}{\int f(x)f(x)^{\alpha-1}\,\mathrm{d}x}$. The p.d.f. $f(x)$ of X can be estimated with kernel functions $K(\cdot)$ via summation over all data points x_i: $\hat{f}_X(x) = \frac{1}{n}\sum_{j=1}^n \frac{1}{h}K(\|x - x_j\|, h)$. If we have a reliable data set of n i.i.d data points, the expected value $\int f(x)g(x)\,\mathrm{d}x$ of any function $g(x)$ over the p.d.f. $f(x)$ can be estimated with $\frac{1}{n}\sum_{i=1}^n g(x_i)$. Therefore the denominator of $\dim_\alpha(X)$ can be estimated with $\frac{1}{n}\sum_{i=1}^n \hat{f}_X(x_i)^{\alpha-1} = \frac{1}{n}\sum_{i=1}^n(\frac{1}{n}\sum_{j=1}^n \frac{1}{h}K(\|x_i - x_j\|, h))^{\alpha-1}$. The numerator is instead equal to $\frac{1}{n}\sum_{i=1}^n \widehat{\mathrm{ID}}(x_i)(\frac{1}{n}\sum_{j=1}^n \frac{1}{h}K(\|x_i - x_j\|, h))^{\alpha-1}$.

With regards to the kNN estimator, it is possible to prove that $K(\|x_i - x_j\|) = \frac{\mathbb{1}(\|x_i - x_j\| \leq r)}{V_D(r)}$ is a proper kernel, where r is a given radius and $V_D(r) = \frac{\pi^{D/2}}{\Gamma(D/2+1)}r^D$ is the volume of a D-dimensional sphere with radius r. A valid choice for the radius r is the distance $d_k(x_i)$ from x_i to its kth nearest neighbor. Given that the number of data points at distance less than or equal to $d_k(x_i)$ from x_i is exactly k, we have $\frac{1}{n}\sum_{i=1}^n \frac{I(\|x_i - x_j\| \leq d_k(x_j))}{V_D(d_k(x_j))} = \frac{1}{n}\frac{k}{V_D(d_k(x_j))} = \frac{1}{n}\frac{k\Gamma(D/2+1)d_k(x_i)^{-D}}{\pi^{D/2}}$. The result follows from algebraic manipulations. □

Proposition 1. *Let X be a set of D continuous variables:*

1) $0 \leq \mathrm{IDD}(X) \leq 1$*;*
2) $\mathrm{IDD}(X) = 0$ *iff all X_i are independent;*
3) $\mathrm{IDD}(X) = 1$ *if there exist one or more manifolds of dimension 1 whose union embeds X;*
4) $\mathrm{IDD}(X) = 1$ *if there exists $1 \leq i \leq D$ such that for all $j \neq i$, X_j is a a function or multivalued function of X_i.*

Proof.
Point 1: By definition, $\dim(X) = \lim_{\delta\to 0^+}\frac{H(X,\delta)}{\log 1/\delta}$. Then regarding the lower bound of IDD, $\sum_{i=1}^D \dim(X_i) - \dim(X)$ is equal to:

$$\begin{aligned} &= \sum_{i=1}^D \lim_{\delta\to 0^+}\frac{H(X_i,\delta)}{\log 1/\delta} - \lim_{\delta\to 0^+}\frac{H(X,\delta)}{\log 1/\delta} \\ &= \lim_{\delta\to 0^+}\frac{1}{\log 1/\delta}\Big(\sum_{i=1}^D H(X_i,\delta) - H(X,\delta)\Big) \\ &= \lim_{\delta\to 0^+}\frac{1}{\log 1/\delta}\mathrm{KL}\Big(p_X(x,\delta)\|p_{X_1}(x_1,\delta)\cdots p_{X_D}(x_D,\delta)\Big), \end{aligned}$$

where KL is the Kullback-Leibler divergence, which is greater or equal to 0 for any $\delta > 0$. Regarding the upper bound of IDD, we use the known fact that the Shannon entropy satisfies $H(X) \geq \max_i H(X_i)$ to prove the following inequalities for

$\sum_{i=1}^{D} \dim(X_i) - \dim(X)$:

$$\begin{aligned}
&= \sum_{i=1}^{D} \lim_{\delta\to 0^+} \frac{H(X_i,\delta)}{\log 1/\delta} - \lim_{\delta\to 0^+} \frac{H(X,\delta)}{\log 1/\delta} \\
&\leq \sum_{i=1}^{D} \lim_{\delta\to 0^+} \frac{H(X_i,\delta)}{\log 1/\delta} - \lim_{\delta\to 0^+} \frac{\max_i H(X_i,\delta)}{\log 1/\delta} \\
&= \sum_{i=1}^{D} \lim_{\delta\to 0^+} \frac{H(X_i,\delta)}{\log 1/\delta} - \max_i \lim_{\delta\to 0^+} \frac{H(X_i,\delta)}{\log 1/\delta} \\
&= \sum_{i=1}^{D} \lim_{\delta\to 0^+} \frac{H(X_i,\delta)}{\log 1/\delta} - \max_i \dim(X_i).
\end{aligned}$$

Since the Shannon entropy is a continuous function, and since X is continuous, it is possible to interchange the limit and max operations.

Point 2: As shown for Point 1 above, $\sum_{i=1}^{D} \dim(X_i) - \mathrm{IDD}(X)$ is equal to

$$\lim_{\delta\to 0^+} \frac{1}{\log 1/\delta} \mathrm{KL}\Big(p_X(x,\delta) \| p_{X_1}(x_1,\delta) \cdots p_{X_D}(x_D,\delta)\Big).$$

The result follows from the fact that for any $\delta > 0$, the KL divergence is equal to 0 iff all variables X are independent.

Point 3: If there exist a manifold or multiple manifolds of dimension 1 embedded in X, then $\mathrm{ID}(x) = 1$ for any locality x. With $\dim(X) = 1$ being the expected ID over the p.d.f. of X, we have that $\mathrm{IDD}(X) = 1$. According to Theorem 1 in [8] if X_i is a continuous random variable, $\dim(X_i) = 1$. Given that we are considering continuous random variables X_i, $\max_i \dim(X_i) = 1$, and therefore $\mathrm{IDD}(X) = 1$.

Point 4: follows immediately from Point 3. □