Characteristics of Local Intrinsic Dimensionality (LID) in Subspaces: Local Neighbourhood Analysis

Tahrima Hashem, Lida Rashidi, James Bailey, Lars Kulik

The University of Melbourne, Australia

Abstract. The local intrinsic dimensionality (LID) model enables assessment of the complexity of the local neighbourhood around a specific query object of interest. In this paper, we study variations in the LID of a query, with respect to different subspaces and local neighbourhoods. We illustrate the surprising phenomenon of how the LID of a query can substantially decrease as further features are included in a dataset. We identify the role of two key feature properties in influencing the LID for feature combinations: correlation and dominance. Our investigation provides new insights into the impact of different feature combinations on local regions of the data.

Keywords. Intrinsic Dimension, Neighbourhood, Subspace.

1 Introduction

Many core operations in data-mining and machine learning are dependent on the choice of similarity measure, as well as the choice of feature space. As the number of features in a dataset increases, the similarity between any pair of data points converges to the distribution mean and the similarity measure loses its discriminability power, i.e., the 'curse of dimensionality'. To overcome this challenge, a range of dimension reduction techniques [1–3] have been developed, to search for a lower dimensional representation that provides a good approximation of the data. A key concept in this context is a dataset's *intrinsic dimensionality* (ID), the minimum number of latent features required to represent the data. This is a natural measure to assess the complexity of a dataset.

In addition to considering the intrinsic dimensionality of an entire dataset, one can also consider intrinsic dimensionality with respect to a particular query object of interest. For this task, one can use local measures of ID [4, 5], which focus on the k-nearest neighbor distances from a specific (query) location in the space. Recently developed *local intrinsic dimensionality* models, i.e., the expansion dimension (ED) [6], the generalised expansion dimension (GED) [7], and local continuous intrinsic dimension (LID) [8, 9], quantify the ID in terms of the growth rate of objects with the expansion in distance from a specific query location. A wide range of applications, e.g., manifold learning, dimension reduction, similarity search [10], local density estimation [11] and anomaly detection [12], have benefited from the use of local ID measures.

In this paper, given a query object, our goal is to analyse how its LID estimates change with respect to different size feature sets. In particular, as more features are used, does the estimated LID of the query increase or decrease? Intuitively, one might expect that as one adds more features, the estimated LID of the query should either increase

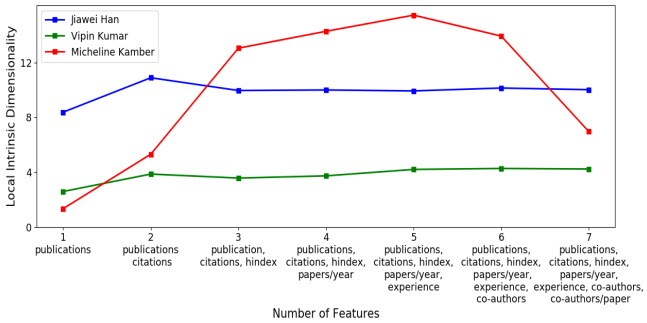


Fig. 1: LID values (computed using the MLE estimator [4]) of three prominent researchers, i.e., *Jiawei Han*, *Micheline Kamber* and *Vipin Kumar* with respect to increasing number of features for the data-mining community of scholars from AMiner.

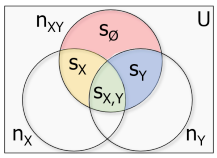


Fig. 2: Neighbourhoods of a query along the individual distance variables, i.e., X and Y as well as their joint distance variable XY.

or remain stable. However, for some situations (in both real and synthetic data), we will demonstrate an unexpected and somewhat counterintuitive phenomenon, that the estimated LID of a query object can actually decrease as more features are used.

We provide a brief example to illustrate the idea. Fig. 1 shows the estimated LID values of three researchers (queries): *Jiawei Han*, *Micheline Kamber* and *Vipin Kumar* from the data-mining community[1] of scholars in the AMiner[2] dataset. We observe that the LID trends are not always smooth as more features are considered. Importantly for researcher *M. Kamber*, there is significant drop in LID when going from 5 to 6 features, and going from 6 to 7 features.

Our purpose is to understand how such a drop in LID is possible and what factors might be responsible. Intuitively, the phenomenon is related to how outlying or inlying the query is within a given subspace, as well as relations between the features themselves, such as their degree of *correlation* and whether a property we call *feature dominance* is present. Developing such an understanding may lead to strategies for more effective feature engineering. Our contributions can be summarised as follows.

1. We identify and illustrate the counter intuitive phenomenon of how the estimated LID of a query object may decrease as more features are considered.
2. We identify the role of two key factors which can influence changes in LID and local neighbourhood for a query: feature dominance and correlation.
3. Given a query object, we study the estimated LID and neighbourhood variations within a feature space and its subspaces, using carefully controlled experiments.

2 Background and Preliminaries

We will first define local intrinsic dimensionality [9] and its estimator [4], and introduce the concept of neighbourhood.

Local Intrinsic Dimensionality: Classical expansion models [6, 8] evaluate the growth rate of the number of data points as the distance to an object of interest increases. E.g., in Euclidean space, when the size of a d-dimensional ball increases by r, it's volume increases by r^d. It is possible to deduce the expansion dimension d from this

[1] https://aminer.org/lab-datasets/soinf/
[2] https://aminer.org/data

growth rate of volume with respect to the size/distance as follows.

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^d \Rightarrow d = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)} \tag{1}$$

The notion of volume is analogous to the probability measure for continuous random variables. The expansion models can be adapted for distance distributions for a given query by replacing the ball set size with the probability of the lower tails of the distribution (Extreme Value Theory), providing a local view of the dimensional structure of the data, as their estimation is restricted to a neighbourhood around the object of interest. Houle et. al. [9] provides the formal definition of LID in light of this theory.

Definition 1 *Assume a reference object $q \in \mathbb{R}$. Let $X > 0$ be a random variable representing distances from q to other objects.*[3] *If $F(x)$ represents the cumulative distance distribution function of X such that $F(x)$ is continuously differentiable at distance $x \in X$, the local intrinsic dimensionality (LID) of the query q at distance x is defined as:*

$$LID_X(x) = \lim_{\epsilon \to 0} \frac{\ln(F((1+\epsilon)x))/F(x)}{\ln((1+\epsilon)x/x)} = \lim_{\epsilon \to 0} \frac{\ln(F((1+\epsilon)x))/F(x)}{\ln(1+\epsilon)} \tag{2}$$

whenever the limit exists.

Applying L'Hopital's rule to the limits of Eqn. 2, LID can be expressed as follows [9].

Theorem 1 ([9]). *If $F(x)$ represents the cumulative distribution function for a distance variable X and $F(x)$ is continuously differentiable such that $F(x) > 0$ for $x > 0$, then*

$$LID_X(x) = \frac{x \cdot F'(x)}{F(x)} \tag{3}$$

Thus, when $x \in X$ tends to zero, the LID of q can be defined in terms of the limit:

$$LID_X = \lim_{x \to 0} LID_X(x) \tag{4}$$

LID gives a rough indication of the dimensionality of the submanifold containing q that would best fit the distribution of data in the vicinity of q. Comprehensive theory regarding the LID model can be found in [8, 9, 13, 14].

LID Estimation:The k nearest neighbour distances can be considered as extreme events associated with the lower tail of the distance distribution according to the Extreme Value Theory. The tails of the continuous probability distributions converge to the Generalized Pareto Distribution (GPD), under some reasonable assumption [15]. Amsaleg et. al. [4, 5] developed several estimators of LID to heuristically approximate the actual underlying distance distribution by a transformed GPD. The Maximum Likelihood Estimator (MLE) has showed a useful trade-off between efficiency and complexity. For a query object q from a data distribution, the MLE estimator of LID(q) is,

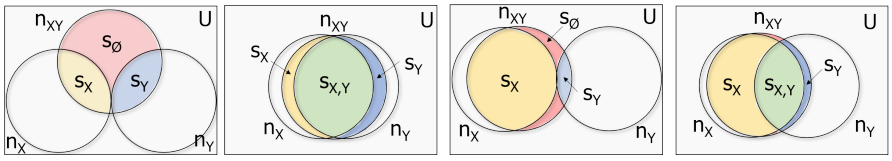
$$\widehat{LID}(q) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log \frac{r_i(q)}{r_k(q)}\right)^{-1} \tag{5}$$

where $r_i(q)$ denotes the distance between q and its i-th nearest neighbour in the sample.

Neighbourhood: Given two features F_X, F_Y[4] and a query object q, we define random variables, X, Y that represent the distance distributions from q to other objects

[3] Suppose $q = 0 \in \mathbb{R}$ and $x_1 = 2 \in X$ are 1 dimensional data values. Then, x_1 directly represents a distance value from q to itself along the X axis.

[4] In fact, our model allows F_X (or F_Y) to be a set of features, rather than a single feature, but for simplicity we will present in the context of being a single feature



(a) Independence (b) Correlation (c) Dominance & Independence (d) Dominance & Correlation

Fig. 3: Four different scenarios of the neighbourhoods for a given query along two distance variables, i.e., X and Y as well as their joint distance variable XY.

using either F_X or F_Y. The joint distribution XY represents the distance distribution from q in the joint space $\{F_X, F_Y\}$. Let LID_X, LID_Y and LID_{XY} be the estimates of the LID for q using X, Y and XY, respectively. The nearest neighbours, n_X, n_Y and n_{XY} that are used to estimate the individual and joint LIDs, are shown as circles in Fig. 2. n_{XY} is a mixture of data objects from n_X, n_Y and U (the whole region).

We use s_X to represent the nearest neighbours within X, that are common with the neighbours in the joint space XY and not with the neighbours in Y (shown in yellow color in Fig. 2). Similarly for s_Y. Thus, $s_X = (n_X \cap n_{XY}) \setminus n_Y$ and $s_Y = (n_Y \cap n_{XY}) \setminus n_X$. Also, $s_{X,Y}$ (the green region) represents the neighbours that are common in both the individual and joint dimensions, $s_{X,Y} = n_X \cap n_Y \cap n_{XY}$. The nearest neighbours in the joint space XY that are not common with any of the neighbours in the individual dimensions are represented as s_ϕ, $s_\phi = n_{XY} \setminus (s_X \bigcup s_Y \bigcup s_{X,Y})$ (the pink region).

For the rest of the paper, we refer the **estimate of the LID value** using Eqn. 5 as the **LID** of a query.

3 Research Questions

The local intrinsic dimensionality (LID) of the query in the joint space XY varies with respect to changes in the local joint neighbourhood (n_{XY}). We next characterise the relationship between the nearest neighbours in the joint space XY and the nearest neighbours in the individual variables, i.e., X and Y, w.r.t. the following two properties:

- **Correlation**: When the two distance variables, X and Y, are positively correlated, one expects that a significant portion of the nearest neighbours in the joint space XY overlap with the nearest neighbours in both X and Y. One also expects that this phenomenon is absent when X and Y are not correlated. i.e. $\mid s^{cor.}_{X,Y} \mid \gg \mid s^{uncor.}_{X,Y} \mid$.
- **Dominance**: A dominant distance variable is one which has a strong influence in determining the nearest neighbours of the query in the joint space XY. If X dominates Y, then a major portion of the nearest neighbours in the joint space XY overlap with the nearest neighbours in X as compared to Y. i.e., $\mid s_X \mid \gg \mid s_Y \mid$.

We will assess in what circumstances LID_{XY} can be less than the individual estimated LID values, LID_X and LID_Y. We particularly focus on the role of a dominant distance variable and/or the presence of a strong correlation between X and Y. We consider the following four research questions (RQ1-RQ4):

RQ1: *Given a query, when two distance variables are independent (uncorrelated), how can LID_{XY} and n_{XY} be characterised with respect to LIDs and neighbourhoods of the individual dimensions (LID_X, LID_Y, n_X, n_Y)?*

For RQ1, we will analyse a query's characteristics, i.e., inlyingness/outlyingness, in terms of its estimated LID in 2D spaces, when the individual distance variables have no

dependency between them. Fig. 3(a) illustrates this scenario, where we observe $| s_{X,Y} | \cong 0$ and $| s_X | \cong | s_Y |$.

RQ2: *Given a query object, when two distance variables are dependent (correlated), how can LID_{XY} and n_{XY} be characterised with respect to LIDs and neighbourhoods of the individual dimensions (LID_X, LID_Y, n_X, n_Y)?*

Correlation between two distance variables can lead to significant changes in the joint neighbourhood in comparison to the uncorrelated case and we expect the joint LID (estimated) to behave differently from the scenario in RQ1 (see Fig. 3(b)). To demonstrate the impact of correlation on the neighbourhood, consider the top 100 nearest neighbours of a query in XY space, if the correlation between X and Y is 1.0, we can expect that $| s_{X,Y} | \cong 100$.

RQ3: *How are LID_{XY} and n_{XY} influenced when one of the distance variables dominates the other? (X dominates Y or vice versa)*

A dominating distance variable can strongly influence the formation of neighbourhood in the joint space. In Fig. 3(c) we note, a significant part of n_{XY} overlaps with n_X and a small part of it overlaps with n_Y. In this case, we have assumed that the distance variables are independent, i.e., $| s_{X,Y} | \cong 0$. In this case, we expect the query to have neighbourhood characteristics for the joint space that are similar to those for the individual variable X, due to the dominance property of X.

RQ4: *In the presence of both correlation and dominance, how can LID_{XY} and n_{XY} be characterised in terms of (LID_X, LID_Y, n_X, n_Y)?*

Fig. 3(d) illustrates this scenario where we observe a positive correlation between X and Y as $| s_{X,Y} | \gg 0$. We find $| s_X | > | s_Y |$, meaning that X still dominates Y.

4 Experimental Study Using Synthetic Data

We observe the behaviour of LID in multiple univariate (Section 4.1) and bivariate (Sections 4.2-4.4) synthetic datasets that are generated to model the scenarios in the research questions *RQ1-RQ4*. We will later investigate a real dataset in Section 5. For our experiments, we model the ***distance distribution*** instead of the actual data distribution, i.e., the generated data values represent the distances from a query that is located at the origin. Note that the query is not generated by the data generation process. Since we always ensure that the generated values of the synthetic datasets are greater than or equal to 0, the data values along each dimension directly represent the distances from the query to themselves. The Euclidean Norm ($\| \cdot \|_2$) is used to measure distance. We use $k = 100$ neighbours in the MLE estimator of LID (see Eqn 5). Unless otherwise stated, *z-score* normalisation has been applied on both synthetic and real data (i.e. on the raw feature values for F_X or F_Y) before estimating the LIDs of a given query.

4.1 LID in Univariate Synthetic Datasets

To model the distance distributions, we have selected the Weibull distribution as it is lower bounded (given x≥0). Eqn. 6 shows the Weibull probability density and cumulative distribution functions. We generate three uniscaled ($\lambda = 1$) Weibull distributions for different values of shape parameter (κ) in Fig. 4. For shape values, $1 < \kappa < 2.6$, the Weibull pdf is positively skewed (right tail), for $2.6 < \kappa < 3.5$ its coefficient of skewness approaches zero (no tail) and for $\kappa > 3.5$ it is negatively skewed (left tail) [16].

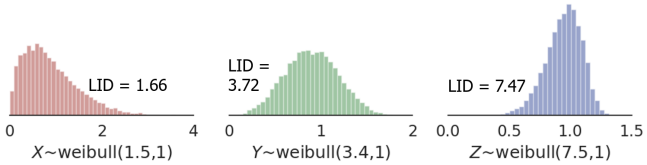


Fig. 4: Histograms of three *Weibull(κ,λ)* distributed distance variables, i.e., X, Y and Z.

$$f_w(x;\kappa,\lambda) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1} \exp^{-(\frac{x}{\lambda})^{\kappa}}$$

$$F_w(x;\kappa,\lambda) = 1 - \exp^{-(\frac{x}{\lambda})^{\kappa}} \qquad F_w^{-1}(x;\kappa,\lambda) = \lambda[-\ln(1-x)]^{\frac{1}{\kappa}} \tag{6}$$

$$LID_X = \lim_{x\to 0} \frac{x \cdot f_w(x;\kappa,\lambda)}{F_w(x;\kappa,\lambda)} = \lim_{x\to 0} \frac{\frac{d}{dx}(x \cdot f_w(x;\kappa,\lambda))}{\frac{d}{dx}(F_w(x;\kappa,\lambda))} = \kappa \tag{7}$$

The theoretical LID of a Weibull distributed distance variable is derived in Eqn. 7 and is equal to the shape value. Also, experimentally the query (at origin) obtains LID values 1.66, 3.72 and 7.47, corresponding to κ values, i.e., 1.5, 3.4, 7.5, respectively. Thus, the larger the shape value of the Weibull distribution, the higher the LID and the more outlying the query is relative to other objects.

4.2 Bivariate Synthetic Datasets Generation

We generate six bivariate synthetic datasets. Each dataset consists of 10,000 data points. Four datasets, i.e., D1:*ND-Independent*, D2:*ND-Correlated*, D3:*D-Independent*, and D4: *D-Correlated*, are generated using the lower bounded *Weibull* distribution. Thus, the generated random variables, i.e., X and Y, can be treated as continuous distance variables for a query at origin $(0,0)$. To achieve control over the *mean* (μ), *variance* (σ^2), *skewness* (β) (i.e., measure of symmetry), and *kurtosis* (γ) (i.e., measure of whether the data is heavy-tailed or not in relation to the normal distribution) of the distance distributions, we generate two further datasets, D5:*ED-Independent* and D6:*ED-Correlated*, using the *Pearson* distribution family, which is effective in modelling skewed observations [17]. In this case, we ensure that all data values are greater or equal to 0, so that the generated data values correspond to distances from the query.

We consider two scenarios, *Scenario 1* and *Scenario 2*, where we model *dominance* and *non-dominance* between two distance variables. For each scenario we generate two different types of datasets, i.e., uncorrelated and correlated, using a Gaussian Copula (described below). Datasets D1 and D2, model the scenarios stated in $RQ1$ and $RQ2$, respectively, whereas D3 and D4, model the scenarios in $RQ3$ and $RQ4$, respectively. Datasets D5 and D6 illustrate the LID behaviour for the same phenomena as D3 and D4, using extremely skewed and heavy-tailed distance distributions.

Scenarios	Title	Name of datasets	Distribution Type	Description of the parameters	Rank Correlation
Scenario 1	D1.	ND-Independent	Weibull	$\kappa_X = 4, \lambda_X = 1, \kappa_Y = 6, \lambda_Y = 1$	$\alpha_s = 0$
	D2.	ND-Correlated	same as D1	same as D1	$\alpha_s = 0.89$
Scenario 2	D3.	D-Independent	same as D1	$\kappa_X = 4, \lambda_X = 8, \kappa_Y = 6, \lambda_Y = 1$	$\alpha_s = 0$
	D4.	D-Correlated	same as D1	same as D3	$\alpha_s = 0.89$
	D5.	ED-Independent	Pearson	μ_X=7, σ_X^2=0.5, β_X=−1.75, γ_X=9 μ_Y=8, σ_Y^2=1, β_Y=0, γ_Y=3	$\alpha_s = 0$
	D6.	ED-Correlated	same as D5	same as D5	$\alpha_s = 0.89$

Table 1: Description of distribution and correlation parameters of synthetic bivariate datasets.

When generating the bivariate distance distributions, our goal is to illustrate the circumstances where the query has different LID values in individual dimensions. We investigate the properties of the 2D local neighbourhood around the query, where LID_{XY} may show an expected increase or unexpected decrease with respect to the individual LIDs LID_X and LID_Y. To ensure that we have different LID values in X and Y dimensions, we use smaller values for the shape parameter in X than Y, making $LID_X < LID_Y$, leveraging our observations in Section 4.1.

We use a copula [18, 19] to generate both the correlated and uncorrelated datasets. Copulas (C) provide a way to model correlated multivariate data. According to Sklar's Theorem [18], any multivariate cumulative distribution function can be expressed in terms of the marginal cumulative distribution functions of the random variables, together with a copula describing their dependence structure (α) (see Eqn. 8).

$$F(x, y) = C(F^1(x), F^2(y); \alpha) \tag{8}$$

The *Gaussian Copula* (C_g) generates correlated uniformly distributed values from a multivariate normal distribution with a given linear correlation (α_p). Thus, a correlated multivariate distribution with the same or different marginal distributions can be obtained by applying the desired inverse cumulative distribution functions (*ICDF*) to the corresponding uniform variables. We follow this technique to generate the four bivariate datasets, i.e., D1, D2, D3 and D4, using the following steps.

- **Step 1**:We use a Gaussian copula C_g with selected linear correlation parameter to sample bivariate uniformly distributed values $U = [U_1, U_2]$ for $U \in [0, 1]$.
- **Step 2**: We apply the *ICDF* of the Weibull distribution (F_w^{-1}) to U_1 and U_2, with the given parameters for each dimension i.e., κ_X, κ_Y, λ_X, and λ_Y, and obtain the desired marginal (Weibull) distributions for X and Y; the process is known as inverse transform sampling [20].

Though we need to provide the linear correlation (α_p) as an input to C_g, this linear correlation is not preserved during the inverse sampling because F_w^{-1} is a non-linear function (see Eqn. 6). However, $F_w^{-1}(u)$ is monotonically increasing for $u \in U$ and $\kappa, \lambda > 0$, and under any monotonic transformation, rank correlation, e.g., *Spearman*'s correlation coefficient (α_s), is preserved [18]. There remains a one-to-one mapping between α_p and α_s for normally distributed data [21] (see Eqn. 9). Hence, the value of α_s between the Weibull distributed variables is almost identical to the initial value of α_p specified in C_g, since C_g is constructed from normally distributed data.

$$\alpha_s = (6/\pi) * sin^{-1}(\alpha_p/2) \tag{9}$$

For generating the uncorrelated datasets, i.e., D1 and D3, we use α_p=0 (α_s=0) for C_g. We use the same scale (λ=1) and different shapes (κ), i.e., 4 and 6, for X and Y in the D1 dataset. In D3, we use a larger value of scale for X (λ=8) than Y (λ=1), so that X can be treated as a dominating distance variable. In fact, we intend to observe how X with its heavily-tailed neighbours, dominates Y in selecting the neighbours in the 2D space (X, Y). On the other hand, we use α_p=0.9 (α_s=0.89) to C_g for the generation of correlated datasets in both non-dominance and dominance cases, i.e., D2 and D4. We use the same Weibull parameters as D1 and D2, for D3 and D4, respectively.

We model extreme scenarios of dominance in both the absence and presence of correlation using D5 and D6 datasets, respectively. The data values are sampled from a *Pearson* distribution family [17]. In D5, X follows a negatively skewed (β_X=−1.75)

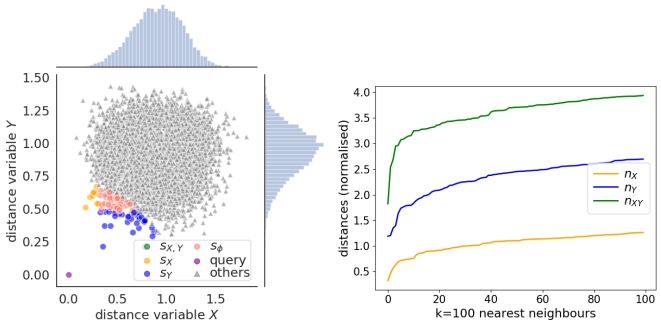
heavy-tailed (γ_X=9) distribution whereas Y models symmetric (β_Y=0) light-tailed (γ_Y=0) distribution (see Table 1). Since both of them are independently sampled, they are uncorrelated. Due to the very skewed distribution along X, we are able to see the drop of joint LID even after applying the *z-score normalisation* in these datasets. Since it is not straightforward to obtain the *ICDF* for the X dimension with the given Pearson parameters, we generate the correlated Pearson numbers in the following step[5].

- **Step 1:** We generate independent (uncorrelated) Pearson values, P_1 and P_2 using the same parameters as D5 and sort them in ascending order.
- **Step 2:** We generate the correlated uniform values, i.e., U_1 and U_2, with $\alpha_p = 0.9$ (equivalent to α_s=0.89) from the Gaussian copula C_g.
- **Step 3:** After sorting the uniform values in ascending order, we obtain two indices, in_1 and in_2, describing the rearranged order of U_1 and U_2, respectively.
- **Step 4:** We position the sorted values of P_1 and P_2 in the same order as the indices, in_1 and in_2, to obtain the final Pearson variables, P_1^c and P_2^c for dimensions X and Y, respectively, in D6 dataset.

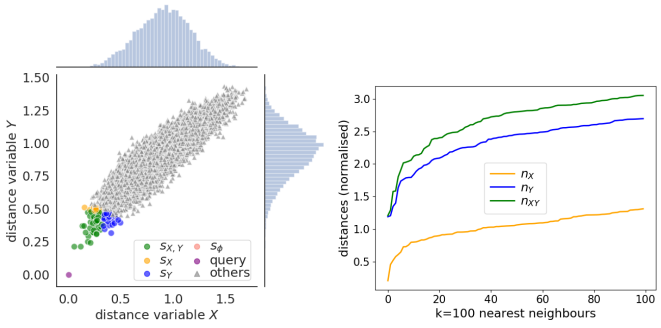
4.3 Scenario 1 (Non-Dominance)

D1 and D2 are generated in a setting where there is no dominant feature (Table 1). Fig. 5(a) and Fig. 5(b) provide the scatter plots and nearest neighbour distance graphs for D1 and D2, respectively. It is clearly notable from the scatter plots that the data values are correlated in D2 (elliptical shape) whereas in D1 they are not (circular shape).

[5] https://au.mathworks.com/help/stats/generate-correlated-data-using-rank-correlation.html



(a) D1:ND-Independent



(b) D2:ND-Correlated

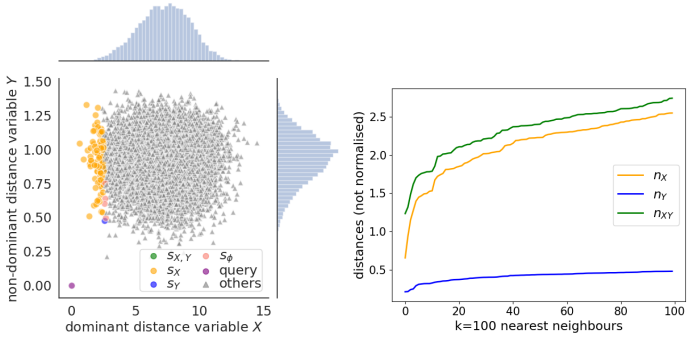
Fig. 5: Scatter plots and **normalised** distance graphs of datasets D1 and D2 modelling the *uncorrelated* and *correlated* variables, respectively, in the *non-dominance* setting.

In Fig. 5(a), we observe that the distances of the local neighbours with respect to X variable are relatively smaller than that of Y. This happens because Y has a larger shape than X. Hence the distribution of Y is skewed more to the left in comparison to X. As a result, LID_Y is larger than LID_X. We further note that there is no common neighbour between the joint space and the individual dimensions, i.e., $s_{X,Y}$=0 (no green dots). The distances of the neighbours along the joint variable XY are far in comparison to the individual variables. Hence, the query obtains a very high LID value, LID_{XY} = 10.18, that is approximately the summation of the individual LIDs, LID_X = 4.87 and LID_Y = 6.35, which matches with results mentioned in [8].

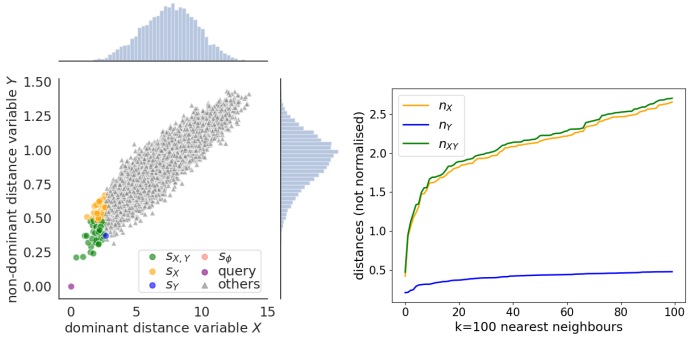
Fig. 5(b) corresponds to the correlated variables X and Y of dataset D2. We observe that 48% of the neighbours, n_{XY} are overlapped with both n_X and n_Y, i.e., $s_{X,Y}$=48 (green dots). We found LID_X=3.78, LID_Y=6.35 and LID_{XY}=6.46. Note that the joint LID in the correlated case is 6.46 which is smaller than the joint LID(=10.18) of the uncorrelated case with the same parameter settings. Thus, if the continuous distance variables are positively correlated, the query finds its 2D neighbours to be more common with the neighbours of the individual dimensions and thus obtains a smaller LID in comparison with the independent case. **This observation answers $RQ1$ and $RQ2$ for the independent and correlated distance variables in the non-dominance setting**.

4.4 Scenario 2 (Dominance)

For the dominance scenario, we consider D3:*D-Independent* and D4:*D-Correlated* datasets. In order to observe the dominance property of a distance variable, we do not standardise



(a) D3:D-Independent



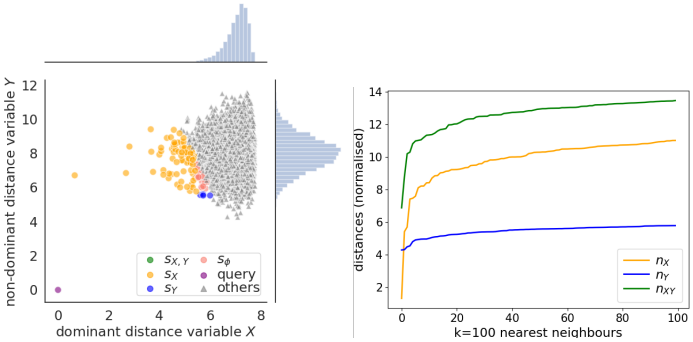
(b) D4:D-Correlated

Fig. 6: Scatter plots and nearest neighbours distance graphs of D3 and D4 datasets modelling the *uncorrelated* and *correlated* variables, respectively, in the *dominance* setting.

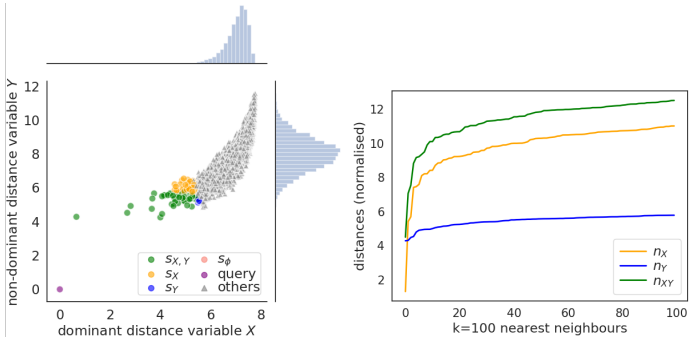
these datasets. We note from the scatter plots that the data shows greater variance in X compared to Y and the distances of the nearest neighbours for Y remain almost constant, whereas there is a steady increase in the distances for X as the number of nearest neighbours grows (see Fig. 6). For the query at the origin, LID_Y is 6.35 for both D3 and D4 datasets while LID_X is 4.87 and 3.80 in D3 and D4, respectively.

For D3 dataset, we observe a drop in LID value with respect to the joint distance variable XY compared to Y, i.e., LID_{XY} = 5.70 while LID_Y=6.35. We note that a major portion of the neighbours in XY overlap with the neighbours from X, i.e., s_X = 95 (the orange dots). There is no overlapping between the neighbours of XY and the individual dimensions X and Y, i.e., $s_{X,Y}$ = 0. As a result, the distances along XY are following the similar trend of along X (Fig. 6(a)). Thus in scenarios where one of the features X is dominant and has a lower LID value in comparison to the non-dominant feature Y, the LID value in the joint space LID_{XY} becomes smaller than that of LID_Y. However, if the dominant variable does not have such property (low LID), we do not observe this reduction of LID value in the joint space (**answering question** $RQ3$).

We demonstrate the correlated scenario of D4 in Fig. 6(b). We find 99% (s_X+$s_{X,Y}$= 51%+48%) of n_{XY} are overlapped with n_X and 48% of them are common with n_Y (green dots). The uniformity of the distances of n_X has a significant influence on the distances of the neighbours in the joint space (XY), causing significant reduction in the LID of



(a) D5:ED-Independent



(b) D6:ED-Correlated

Fig. 7: Scatter plots and nearest neighbours distance graphs of D5 and D6 datasets modelling the *uncorrelated* and *correlated* variables, respectively, in the *dominance* setting.

Scenarios	Title	Name of the datasets	Neighbourhood in (XY)				LID estimates		
			s_X	s_Y	$s_{X,Y}$	s_ϕ	LID_X	LID_Y	LID_{XY}
Scenario 1	D1.	ND-Independent	9	35	0	56	4.87	6.35	10.18
	D2.	ND-Correlated	13	38	48	01	3.78	6.35	6.46
Scenario 2	D3.	D-Independent	95	01	0	04	4.87	6.35	5.70
	D4.	D-Correlated	51	01	48	0	3.80	6.35	3.93
	D5.	ED-Independent	75	04	00	21	7.63	16.59	13.74
	D6.	ED-Correlated	44	05	51	0	7.63	16.59	9.80

Table 2: Changes in local neighbourhood and LID estimation w.r.t. the distance variables X, Y and XY for all six synthetic datasets in absence(/presence) of dominance and correlation.

XY, i.e., LID_{XY}(=3.93) $\ll$ LID_Y(=6.35). n.b. the joint LID in D4 (3.93) is smaller than the joint LID of the uncorrelated case in D3 (5.70) (**answering question** $RQ4$).

Extreme Distributions: D5:*ED-Independent* and D6:*ED-Correlated* illustrate the extreme case of dominance for uncorrelated and correlated Pearson random variables, respectively (see Fig. 7). Here, X has a negative long tailed asymmetric distribution whereas Y follows a short tailed symmetric distribution. The query is an outlier in both dimensions, but it obtains smaller LID in X (LID_X = 7.63) than Y (LID_Y = 16.59) for both datasets. This phenomenon occurs since the query is surrounded by a group of outliers in X, whereas all the nearest neighbours are quite far away from the query in Y.

In D5, the uncorrelated dataset, the neighbours along XY mostly intersect with the neighbours along X as we find s_X=75. Since they are uncorrelated there is no overlap among the neighbours in the joint space XY and the individual dimensions X and Y. Here, X has a bigger influence on LID_{XY} since the nearest neighbours of the query in X are having the dominant distances. We find LID_{XY} = 13.74 which is smaller than LID_Y(=16.59). We obtain a similar LID behaviour in D6. However, due to the correlation between X and Y, 51% of n_{XY} overlaps with both n_X and n_Y in D6. A significant portion of n_{XY} is coming only from n_X, i.e., s_X = 44%, which causes a drop in LID_{XY} = 9.80 as compared to LID_Y(=16.59). Note that for the correlated case the reduction of LID value in the joint space is much greater than the uncorrelated case.

Our results for synthetic datasets are summarised in Table 2.

5 Experiments with Real Data

The AMiner dataset is a large academic social network comprising 1.7M authors, 2.1M papers and 4.3M coauthor relationships. We consider 7 numerical features: publications (*pub*), citations (*ct*), h-index (*hi*), papers/year (*ppy*), co-authors, co-authors/paper

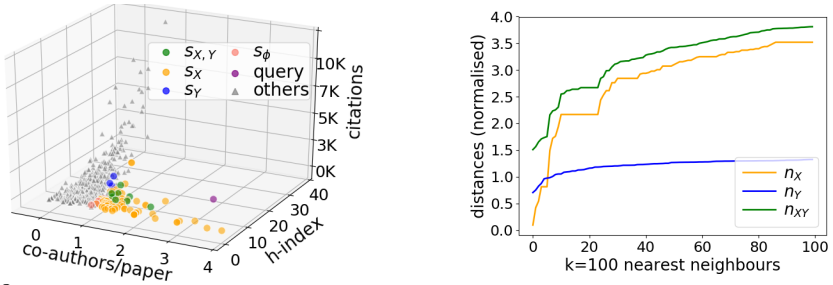


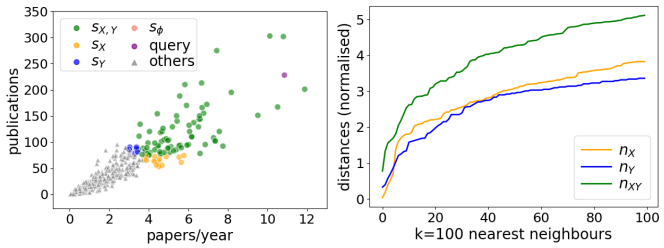
Fig. 8: The left figure is a scatter plot of the data-mining community of researchers from AMiner for the query *M. Kamber* and the three features *ct*, *hi* and *avgco*. Nearest neighbours distance graph is shown on the right. *M. Kamber* obtains LID(*ct*, *hi*) = 10.85 < LID(*ct*, *hi*, *avgco*) = 5.2.

(*avgco*), and research experience. We analysed the LID behaviour in this dataset by considering different authors as the query and estimating the LID value for various combinations of features. We consider two prominent researchers, i.e., *Micheline Kamber* and *Jeffrey Xu Hu* from the data-mining community consisting of 641 researchers, as queries to model different phenomena described in Sections 4.3-4.4. We use k=100 in the MLE estimator, but obtained similar results (not reported) for k = 30, 60.

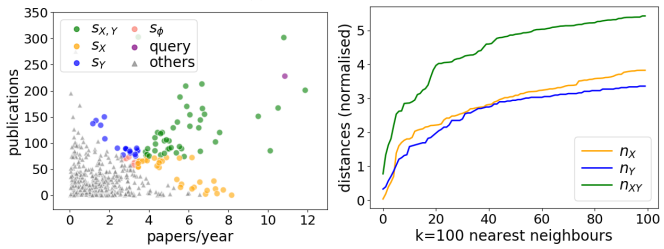
Case Study 1- Dominance: Given *M. Kamber* as the query, and the three features, *ct*, *hi* and *avgco*, we illustrate how dominance influences the LID value. In this scenario, the dominant variable X corresponds to the distances from the query to other authors on the feature *avgco*. While the non-dominant variable Y corresponds to the distances between query to others with respect to the two features *citations* and *hindex*. We find that LID_X = LID(*avgco*) = 3.39 and LID_Y = LID(*ct*, *hi*) = 10.85.

Fig. 8 provides the 3D scatter plot and the normalised distance graph of 100 nearest neighbours that are used to estimate the LID along the distance variables, X, Y and XY. It is evident that the query is an outlier in both X and Y. We find that 81% of the neighbours (the orange dots) in XY are coming only from X, i.e., s_X=81 (see Table 3), which is the reason for obtaining a smaller LID value of 5.2 for the 3D feature-set (*ct, hi, avgco*), i.e., LID_{XY} = 5.2, after adding *avgco* to the 2D feature-set (*ct, hi*).

Case Study 2- Correlation: We observe the LID behaviour on AMiner dataset in terms of the correlation of the features. In our experiments, we also explored the effect of decorrelation on the LID value. Consider the query: *Jeffrey* and the two features papers/year and publications. Here, X represents the distances from the query to others on *ppy* while Y on *pub*. *Jeffrey* obtains LID values of 2.5 and 3.1 with respect to *ppy* and *pub*, respectively (see Table 3).



(a) Before Decorrelation



(b) After Decorrelation

Fig. 9: Scatter plots of the data-mining community of researchers, and distance graphs of the query *Jeffrey Xu Hu* before and after decorrelation of the features papers/year and publications. *Jeffrey* obtains LID values of 3.4 and 4.2 before and after decorrelation, respectively.

Query	Features	Scenarios	Correlation Coefficients	n_{XY}				LID_X	LID_Y	LID_{XY}
				s_X	s_Y	$s_{X,Y}$	s_ϕ			
M. Kamber	*X*:*avgco*=3.6, *Y*:*ct*=1546,*hi*=4	Dominance	–	81	3	8	8	3.39	10.85	5.2
Jeffrey Xu Hu	*X*:*ppy* = 10.9, *Y*: *pub* = 228	Correlation	α_s = 0.96	16	7	77	0	2.5	3.1	3.4
		Decorrelation	α_s = 0.30	31	18	44	7	2.5	3.1	4.2

Table 3: Joint neighbourhood and LID values for the distance variables X, Y and XY in the dominance, correlation and decorrelation scenarios. The second column describes the query values for the features.

The features *ppy* and *pub* are highly correlated, i.e., $\alpha_s(X, Y)$= 0.96, where *Jeffrey* obtains LID^{corr}_{XY}=3.4. After removing the correlation [22], i.e., by random permutation of objects in X and Y, yielding $\alpha_s(X, Y)$=0.3, *Jeffrey* obtains LID^{decor}_{XY}=4.2 which is larger than LID^{corr}_{XY}. Fig. 9(a) and Fig. 9(b) display the scatter and distance plots before and after the decorrelation, respectively. We note in Fig. 9(b), that the no. of common neighbours between XY and the individuals, i.e., X and Y, ($s_{X,Y}$) decreases (green dots) while the neighbours in s_X and s_Y increases (orange and blue dots) as compared to Fig. 9(a). The distance plot in Fig. 9(a) shows a more uniform distance distribution, compared to Fig. 9(b) which shows an abrupt increase at multiple locations of the plot.

6 Discussion

As a default, one might expect that the local intrinsic dimensionality of a query should increase as more features are used. However, our studies using both real and synthetic data indicate that under certain conditions such as dominance of a feature or presence of correlation between features, the estimated LID of a query can instead decrease. During the expansion of an existing feature space, significant changes might occur in the neighborhood local to the query. Our studies found, when a query's local neighborhoods are dissimilar with respect to different features, this phenomenon could occur. Some general observations are:

- **Independence**: When the features are independent, the LID in merged space is approximately the summation of the LIDs of the individual features. It matches with the theoretical observation of LID in joint space as stated in [8, 13].
- **Dominance**: When a dominant feature with low LID (LID_X), is combined with a feature with high LID (LID_Y), the LID in the joint space will be lie between LID_X and LID_Y ($LID_X < LID_{XY} \leq LID_Y$).
- **Correlation**: In the presence of a positive correlation, when a feature with low LID (LID_X) is combined with another feature with high LID (LID_Y), the joint LID is much smaller than the summation of the LIDs of the individual dimensions ($LID_{XY} \ll (LID_X + LID_Y)$). The stronger the correlation, the larger the reduction in LID_{XY}.
- **Dominance and Correlation**: In the presence of positive correlation, when a dominant feature with low LID (LID_X) is combined with another feature with high LID (LID_Y), the joint LID is between LID_X and LID_Y ($LID_X < LID_{XY} \ll LID_Y$). The stronger the dominance and correlation, the larger the reduction in LID_{XY}.

7 Conclusions

We have analysed the behaviour of local intrinsic dimensionality (LID) for changes in the feature-space as well as the neighbourhood of a query. We considered two key factors, correlation and dominance, that can cause the LID to decrease when more features are considered. Thus, increasing the number of features may not always result in an increase in the (local) complexity of the data around a query object. Our observations may provide insights into the feature selection and enumeration process, as well as ob-

ject inlyingness/outlyingness across subspaces. For the future, it will be interesting to develop further theory to understand these findings.

References

1. C. Bouveyron, G. Celeux, and S. Girard, "Intrinsic dimension estimation by maximum likelihood in probabilistic pca," *Pattern Recognition Letters*, vol. 32, pp. 1706–1713, 2011.
2. J. B. Tenenbaum, V. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.
3. S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *ICCV*, vol. 290, pp. 2323–2326, 2000.
4. L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, and M. Nett, "Extreme-value-theoretic estimation of local intrinsic dimensionality," *DMKD*, vol. 32, no. 6, pp. 1768–1805, 2018.
5. ——, "Estimating local intrinsic dimensionality," in *SIGKDD*, 2015, pp. 29–38.
6. D. R. Karger and M. Ruhl, "Finding nearest neighbors in growth-restricted metrics," in *Proceedings of the Thiry-fourth Annual ACM STOC*, 2002, pp. 741–750.
7. M. E. Houle, H. Kashima, and M. Nett, "Generalized expansion dimension," in *ICDMW*, 2012, pp. 587–594.
8. M. E. Houle, "Dimensionality, discriminability, density and distance distributions," in *ICDMW*, 2013, pp. 468–473.
9. ——, "Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications," in *SISAP*, 2017, pp. 64–79.
10. M. E. Houle, X. Ma, M. Nett, and V. Oria, "Dimensional testing for multi-step similarity search," in *ICDM*, 2012, pp. 299–308.
11. J. Von Brünken, M. Houle, and A. Zimek, "Intrinsic dimensional outlier detection in high-dimensional data," *NII Technical Reports*, pp. 1–12, 01 2015.
12. M. E. Houle, E. Schubert, and A. Zimek, "On the correlation between local intrinsic dimensionality and outlierness," in *SISAP*, 2018, pp. 177–191.
13. M. E. Houle, "Inlierness, outlierness, hubness and discriminability: An extreme-value-theoretic foundation," *NII Technical Reports*, pp. 1–32, 03 2015.
14. ——, "Local intrinsic dimensionality II: multivariate analysis and distributional support," in *SISAP*, 2017, pp. 80–95.
15. S. G. Coles, *An introduction to statistical modeling of extreme values*. Springer, 2001, vol. 208.
16. D. N. Rousu, "Weibull skewness and kurtosis as a function of the shape parameter," *Technometrics*, vol. 15, no. 4, pp. 927–930, 1973.
17. K. Pearson, "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material," *Philosophical Transactions of the Royal Society of London Series A*, vol. 186, pp. 343–414, 1895.
18. R. B. Nelsen, *An Introduction to Copulas*. Springer Science Business Media Inc., 2006.
19. T. Takeuchi, "Constructing a bivariate distribution function with given marginals and correlation: Application to the galaxy luminosity function," *Monthly Notices of the Royal Astronomical Society*, vol. 406, 04 2010.
20. M. G. Kendall, A. Stuart, and J. K. Ord, Eds., *Kendall's Advanced Theory of Statistics*. Oxford University Press, Inc., 1987.
21. M. G. Kendall, "Rank and product-moment correlation," *Biometrika*, vol. 36, no. 1/2, pp. 177–193, 1949.
22. A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas, "Assessing data mining results via swap randomization," *ACM TKDD*, vol. 1, no. 3, 2007.