The Effect of Random Projection on Local Intrinsic Dimensionality

Michael E. Houle and Ken-ichi Kawarabayashi

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
meh@nii.ac.jp, k_keniti@nii.ac.jp

Abstract. Much attention has been given in the research literature to the study of distance-preserving random projections of discrete data sets, the limitations of which are established by the classical Johnson-Lindenstrauss existence lemma. In this theoretical paper, we analyze the effect of random projection on a natural measure of the local intrinsic dimensionality (LID) of smooth distance distributions in the Euclidean setting. The main contribution of the paper consists of upper and lower bounds on the LID in the vicinity of a reference point after random projection. The bounds depend only on the LID in the original data domain and the target dimension of the projection; as the difference between the target and intrinsic dimensionalities grows, these bounds converge to the LID of the original domain. The paper concludes with a brief discussion of the implications for applications in databases, machine learning and data mining.

1 Introduction

In an attempt to alleviate the effects of high dimensionality, and thereby improve the discriminability of data, simpler representations of the data are often sought by means of a number of supervised or unsupervised learning techniques. One of the earliest and most well-established simplification strategies is dimensional reduction, which seeks a projection to a lower-dimensional subspace that minimizes the distortion of the data. Dimensional reduction has applications throughout machine learning and data mining: these include feature extraction, such as in PCA and its variants [6, 39]; multidimensional scaling [38, 41]; manifold learning [38, 40, 42]; and regression-based similarity learning [43].

Among the various approaches to dimensional reduction, much attention has been given to the study of projections that approximately preserve all pairwise distances within discrete point sets. The limitations of such projections have been established by the classical Johnson-Lindenstrauss (JL) existence lemma [30], which can be stated as follows: given some distortion threshold $0 < \varepsilon < 1$, a set of n points in $\mathbb{R}^m$, and a target dimension $t > (8 \ln n)/\varepsilon^2$, there exists a linear projection $f : \mathbb{R}^m \to \mathbb{R}^t$ such that

$$(1-\varepsilon) \cdot \|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon) \cdot \|u-v\|^2$$

for all pairs of points u and v in the set. This bound on the target dimension has been shown to be asymptotically worst-case optimal for linear projection [33].

Subsequent research has focused on the determination of data transforms that satisfy the JL bounds. Early approaches (such as in [11, 16, 29]) were based on projection to spherically random subspaces; however, the associated transform matrices were dense and expensive to compute. Achlioptas [1] showed that the entries of a projection matrix could be randomly selected from among $\{-1, 0, 1\}$ so as to satisfy the bounds with high probability (after the introduction of a scaling factor). More recent work has been devoted to improving the speed and sparsity of JL transforms [2, 10, 31]. Variants of the JL lemma have also been applied to subspace- and manifold-structured continuous point sets [4, 5].

In general, dimensional reduction requires that an appropriate dimension for the reduced space (or approximating manifold) must be either supplied or learned, ideally so as to minimize the error or loss of information incurred. The dimension of the surface that best approximates the data can be regarded as an indication of the intrinsic dimensionality (ID) of the data set, or of the minimum number of latent variables needed to represent the data. ID thus serves as an important natural measure of the complexity of data.

Over the past decades, many characterizations of the ID of sets have been proposed: classical measures (primarily of theoretical interest), including the Hausdorff dimension, Minkowski-Bouligand or 'box counting' dimension, and packing dimension (for a general reference, see [14, 36]); the correlation dimension [18]; 'fractal' measures of the space-filling capacity or self-similarity of the data [7, 15, 19]; topological estimation of the basis dimension of the tangent space of a data manifold from local samples [6, 39]. Projection-based learning methods such as PCA can produce as a byproduct an estimate of ID.

The aforementioned ID measures can be described as 'global', in that they consider the dimensionality of the set in its entirety. However, when the data set resides on a collection of manifolds, or is distributed according to a mixture of underlying models, global measures may not be indicative of the intrinsic dimensionality in all regions of the set. In order to assess the intrinsic dimensionality in the vicinity of a specified reference point, 'local' ID measures have been proposed that are defined solely in terms of the distances to a set of near neighbors of the reference point. Expansion models, in particular, assess ID in terms of the rate at which the number of encountered objects grows as the considered range of distances expands from the reference location. Such models include the expansion dimension (ED) [32], the generalized expansion dimension (GED) [23], Levina and Bickel's estimator [34], the minimum neighbor distance (MiND) [39], and the local intrinsic dimenension (LID) [3, 20]. The correlation dimension can also be regarded as an expansion model, albeit one that takes into account the growth rates from all points [22, 37]. Local expansion models of ID have also been used in the analysis of a projection-based heuristic for outlier detection [12], and of the complexity of search queries in indexing [8, 24–27, 32].

In this paper, we will be concerned with the LID model of intrinsic dimensionality, which can be regarded as an extension of the (generalized) expansion

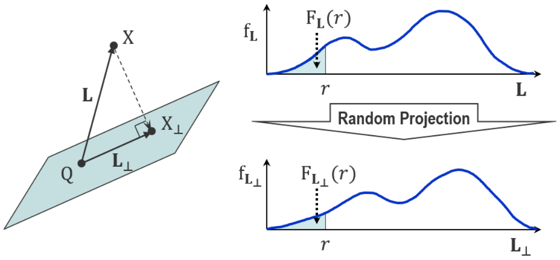

Fig. 1: Random projection of the distribution associated with the smooth random distance variable $\mathbf{L}$, defined as the Euclidean distance from reference point Q to a sample point X drawn from some domain in $\mathbb{R}^m$. The projection induces a new random distance variable $\mathbf{L}_\perp$, defined as the distance from Q to the projection of X in a randomly-oriented t-dimensional subspace.

dimension to the statistical setting of smooth distributions over the non-negative reals [3, 20–22]. Instead of regarding intrinsic dimensionality as a characteristic of a collection of data points (as evidenced by their distances from a supplied reference location), the LID is a direct characterization of the complexity of the underlying distribution itself. With this latter perspective, an original data set drawn from a metric space defines a sample of distances from this underlying distribution, from which one can seek the intrinsic dimensionality of the distribution of distances to some fixed reference location. Note that the model does not require that the sample data be constrained to lie on a manifold.

The LID formulation can be shown to be equivalent to a formulation of the indiscriminability of the underlying smooth distance distribution as evidenced by its cumulative distribution function F. The indiscriminability is modeled as a function $\mathrm{ID}_F(r)$ of the distance $r \in [0, \infty)$, which tends to the local intrinsic dimension value $\mathrm{ID}^*_F \triangleq \lim_{r \to 0^+} \mathrm{ID}_F(r)$ as the radius r vanishes. ID^*_F has been shown to be equivalent to the notion of the 'degree' or 'index' in the statistical theory of extreme values (EVT); indeed, the EVT index has been interpreted as a form of dimension within statistical contexts [9]. Practical methods that have been developed within EVT for the estimation of the index, including the well-known Hill estimator and its variants obtained through maximum likelihood estimation [3, 28, 34], can all be applied to LID (for a survey, see [17]).

In this theoretical paper, we will be concerned with the effect on LID when the distance distribution is subjected to a random projection (as illustrated in Figure 1). As the main contribution of our paper, we prove that under reasonable assumptions, for the LID formulation for Euclidean distance distributions derived from a reference location within a global data distribution, a randomly-oriented linear projection produces a distance distribution (relative to the projective subspace) whose LID value at the reference location, $\mathrm{ID}^*_{F_\perp}$, satisfies

$$\frac{t \cdot \mathrm{ID}^*_F}{t + \mathrm{ID}^*_F} \le \mathrm{ID}^*_{F_\perp} \le \mathrm{ID}^*_F \ .$$

The result indicates that LID is stable under random projection whenever the projection dimension t significantly exceeds ID^*_F, and that stability is lost as t approaches ID^*_F.

Whereas the JL lemma determines a lower limit above which a projection can always be found so as to (approximately) preserve all pairwise distances, our result considers the effect of projection on the distribution of distances from an individual reference location. Bounds on the ID after projection are also known for the Hausdorff dimension: here, Mattila [35] proved that for almost all projections of an analytic set E from $\mathbb{R}^m$ to the set $E_\perp$ in $\mathbb{R}^t$, the Hausdorff dimension $\mathrm{HD}(E_\perp)$ of the projection equals $\min\{\mathrm{HD}(E), t\}$. However, the Hausdorff dimension is a global measure of ID defined on analytic sets — it does not give any insight into the problem considered in this paper: the effect of projection on the local ID of distance distributions, and the discriminability of distance measures.

The remainder of the paper is organized as follows. In the next section, we give an overview of LID and its properties. In Section 3, we give a proof of our main result. For the initial part, we borrow the projection framework and Chernoff bound employed by [11] in their proof of the JL lemma. In Section 4, we conclude with a discussion of the implications of the LID projection bounds.

2 Local Intrinsic Dimensionality

In this section, we present an overview of the measure of local ID for distance distributions as formulated in [21].

2.1 Intrinsic Dimensionality and Indiscriminability

The LID model as first proposed in [20] takes a distributional view of data — instead of inferring dimensional characteristics from a sample of points, dimensionality is modeled in terms of a distribution of non-negative scalar values, as one would expect to see from the distances calculated from a reference location to points generated according to some hidden process.

As a motivating example from m-dimensional Euclidean space, consider the situation which the volumes V_1 and V_2 are known for two balls of differing radii r_1 and r_2, respectively, centered at a common reference point. The dimension m can be deduced from the ratios of the volumes and the distances to the reference point, as follows:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \implies m = \frac{\ln(V_2/V_1)}{\ln(r_2/r_1)} .$$

For finite data sets, GED formulations are obtained by estimating the volume of balls by the numbers of points they enclose [23]. In contrast, for continuous real-valued random distance variables, the notion of volume is naturally analogous to that of probability measure. ID can then be modeled as a function of distance $\mathbf{X} = r$, by letting the radii of the two balls be $r_1 = r$ and $r_2 = (1+\epsilon)r$, and letting $\epsilon \to 0$. The following definition generalizes this notion even further, to any real-valued function (not necessarily a cumulative distribution function) that non-zero in the vicinity of $r \neq 0$.

Definition 1. *Let F be a real-valued function that is non-zero over some open interval containing $r \in \mathbb{R}$, $r \neq 0$. The* intrinsic dimensionality of F at r *is defined as*

$$\mathrm{IntrDim}_F(r) \;\triangleq\; \lim_{\epsilon \to 0} \frac{\ln\left(F((1+\epsilon)r)/F(r)\right)}{\ln(1+\epsilon)}\,,$$

whenever the limit exists.

The intrinsic dimensionality of the cumulative distribution function F of a distance distribution has also been shown in [20, 21] to be equivalent to a measure of its indiscriminability. The discriminability of a random distance variable $\mathbf{X}$ is assessed in terms of the relative rate at which probability measure increases as the distance increases.

Definition 2. *Let F be a real-valued function that is non-zero over some open interval containing $r \in \mathbb{R}$, $r \neq 0$. The* indiscriminability of F at r *is defined as*

$$\mathrm{InDiscr}_F(r) \;\triangleq\; \lim_{\epsilon \to 0} \frac{F((1+\epsilon)r) - F(r)}{\epsilon \cdot F(r)}\,,$$

whenever the limit exists.

The following fundamental theorem adapted from [21] shows that for distance distributions with continuously differentiable cumulative distribution functions, the notions of indiscriminability and intrinsic dimensionality are in fact one and the same. The proof follows by applying l'Hôpital's rule to the limits in Definitions 1 and 2.

Theorem 1 ([21]). *Let F be a real-valued function that is non-zero over some open interval containing $r \in \mathbb{R}$, $r \neq 0$. If F is continuously differentiable at r, then*

$$\mathrm{ID}_F(r) \triangleq \frac{r \cdot F'(r)}{F(r)} = \mathrm{IntrDim}_F(r) = \mathrm{InDiscr}_F(r).$$

When considering the local intrinsic dimensionality of a distance distribution, the question arises as to how the choice of r should be made. Asymptotically, as the number of data samples rise, for any fixed positive integer k the k-nearest neighbor radius can be seen to tend to zero. For this reason, we are especially interested in the case where $r \to 0$. Accordingly, we define the local intrinsic dimensionality (LID) to be the limit of the indiscriminability as $r \to 0$, whenever the limit exists:

$$\mathrm{ID}^*_F \triangleq \lim_{r \to 0^+} \mathrm{ID}_F(r)\,.$$

2.2 Two Properties of Local ID

We now state (without proof) two technical results from [20] needed for the proof of the main theorem of this paper.

In the context of distance distributions with smooth cumulative distribution functions, the indiscriminability of a cumulative distribution function after transformation can be decomposed into two factors: the indiscriminability of the cumulative distribution function before transformation, and the indiscriminability of the transform itself.

Theorem 2 ([20]). *Let g be a real-valued function that is non-zero and continuously differentiable over some open interval containing $r \in \mathbb{R}$, except perhaps at r itself. Let f be a real-valued function that is non-zero and continuously differentiable over some open interval containing $g(r) \in \mathbb{R}$, except perhaps at $g(r)$ itself. Then*

$$\mathrm{ID}_{f \circ g}(r) = \mathrm{ID}_g(r) \cdot \mathrm{ID}_f(g(r))$$

whenever $\mathrm{ID}_g(r)$ and $\mathrm{ID}_f(g(r))$ are defined. If $r = f(r) = g(r) = 0$, then

$$\mathrm{ID}^*_{f \circ g} = \mathrm{ID}^*_f \cdot \mathrm{ID}^*_g$$

whenever ID^*_f and ID^*_g are defined.

The second technical result needed establishes upper and lower bounds on the expansion of probability measure over a fixed range of distances, in terms of upper and lower bounds on LID values over the range.

Theorem 3 ([20]). *Let F be a real-valued function that is non-zero and continuously differentiable over some open interval containing $[a, b] \subset \mathbb{R}$, where $0 < a \leq b$. Let $\overline{\mathrm{ID}}_F(a, b)$ and $\underline{\mathrm{ID}}_F(a, b)$ be the supremum and infimum of $\mathrm{ID}_F(r)$ taken over the range $r \in [a, b]$. Then*

$$\left(\frac{b}{a}\right)^{\underline{\mathrm{ID}}_F(a,b)} \leq \frac{F(b)}{F(a)} \leq \left(\frac{b}{a}\right)^{\overline{\mathrm{ID}}_F(a,b)} .$$

3 Intrinsic Dimensionality after Projection

In this section, as the main contribution of this paper, we examine the effect of random projection on the distribution of distances to a reference point induced by a data distribution in Euclidean space. In particular, we prove the following upper and lower bounds on the LID of the reference point after projection of the data distribution to a t-dimensional subspace (which we will refer to as $\mathrm{ID}^*_{F_\perp}$), in terms of both t and the original LID value (which we will refer to as ID^*_F).

Theorem 4. *Let $\mathbf{L}$ be a random variable representing the Euclidean distance from some fixed reference point Q to a randomly-generated point $X \in \mathbb{R}^m$. Also, let $\mathbf{L}_\perp$ be the random variable representing the Euclidean distance between the images of these points under a uniform random projection $\psi : \mathbb{R}^m \to \mathbb{R}^t$ to an* *arbitrarily-oriented subspace of target dimension* $t < m$*. Let* F *and* $F_\perp$ *be the respective cumulative distribution functions of* $\mathbf{L}$ *and* $\mathbf{L}_\perp$*. If there exists some* $\epsilon > 0$ *such that both* F *and* $F_\perp$ *are continuously differentiable over the interval* $(0, \epsilon)$*, and if the limits* ID^*_F *and* $\mathrm{ID}^*_{F_\perp}$ *both exist, then*

$$\frac{t \cdot \mathrm{ID}^*_F}{t + \mathrm{ID}^*_F} \;\leq\; \mathrm{ID}^*_{F_\perp} \leq \mathrm{ID}^*_F \;.$$

It should be emphasized that the theorem is a statement concerning the distribution of $\mathbf{L}_\perp$, and not the random projection of a particular fixed data set. The random variable $\mathbf{L}_\perp$ follows the distribution of distances to Q obtained when generating a data point, and then subjecting the point to a random projection before measuring its distance to Q (see Figure 1). We do not reason in terms of collections of data samples, but rather on the effect of projection on the distance to Q of a single data sample.

3.1 Random Projection

For the initial part of our proof, we borrow the projection framework and Chernoff error bound formula employed by Dasgupta and Gupta [11] in their proof of the Johnson-Lindenstrauss Theorem (an excellent treatment of which can also be found in [13]). However, instead of using their framework to analyze the probability of obtaining a low-distortion embedding of a fixed data set, we will use it to bound the growth rates within neighborhoods of the reference point before and after projection.

Without loss of generality, we may assume that our reference point Q coincides with the origin of our original Euclidean space $\mathbb{R}^m$. Under this assumption, all the distances of interest coincide with the length of randomly-generated vectors from either the original data distribution, or the data distribution obtained after random projection.

The proof framework of [11] considers the effect of random projection on the length of a fixed vector, by first considering the effect on the associated normalized (unit length) vector. The authors note that the distribution of lengths of projection of this fixed unit vector onto a randomly-selected space is the same as the distribution of the lengths of the projection of a randomly-selected unit vector onto a fixed space (see Figure 2). Within this setting, the expected length of projection of a random m-dimensional unit vector to a t-dimensional subspace, as well as Chernoff-style bounds on the probability of the length varying from this expected length, are established by the following two lemmas.

Lemma 1 ([11]). *Let* $Y = (Y_1, \ldots, Y_m)$ *be a vector selected uniformly at random from the unit sphere in* $\mathbb{R}^m$*. Let* $Y_\perp \in \mathbb{R}^t$ *be the projection of* Y *onto any* t *of its coordinates, where* $0 < t < m$*. Then*

$$\mathbf{E}[\|Y_\perp\|^2] \;=\; \frac{t}{m}.$$

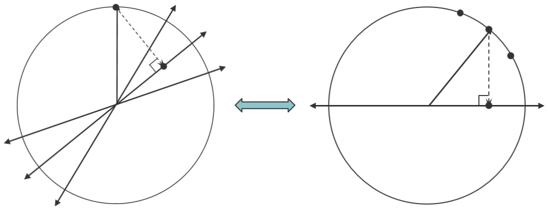

Fig. 2: The distribution of lengths of projection of a fixed unit vector onto a randomly-selected space containing the origin (as shown on the left) is the same as the distribution of the lengths of the projection of a randomly-selected unit vector onto a fixed space (as shown on the right).

Proof: Since the Y_i are identically distributed, we may without loss of generality assume that the projection spans the first t coordinates of $\mathbb{R}^m$. Thus,

$$\mathbf{E}[\|Y_\perp\|^2] = \mathbf{E}\left[\sum_{i=1}^{t} Y_i^2\right] = \sum_{i=1}^{t} \mathbf{E}[Y_i^2] = t \cdot \mathbf{E}[Y_j^2]$$

for any choice of $j \in \{1, \ldots, m\}$. Since Y is a unit vector, we also have

$$1 = \mathbf{E}[\|Y\|^2] = \mathbf{E}\left[\sum_{i=1}^{m} Y_i^2\right] = m \cdot \mathbf{E}[Y_j^2]\,.$$

Combining these two expressions, the result follows. □

Lemma 2 ([11]). *Let $Y = (Y_1, \ldots, Y_m)$ be a vector selected uniformly at random from the unit sphere in $\mathbb{R}^m$. Let $Y_\perp \in \mathbb{R}^t$ be the projection of Y onto any t of its coordinates, where $0 < t < m$. If $\beta < 1$, then*

$$\Pr\left[\|Y_\perp\|^2 \leq \beta\frac{t}{m}\right] \;\leq\; \beta^{\frac{t}{2}}\left(1 + \frac{t(1-\beta)}{m-t}\right)^{\frac{m-t}{2}},$$

and if $\beta > 1$, then

$$\Pr\left[\|Y_\perp\|^2 \geq \beta\frac{t}{m}\right] \;\leq\; \beta^{\frac{t}{2}}\left(1 + \frac{t(1-\beta)}{m-t}\right)^{\frac{m-t}{2}}.$$

For details of this latter proof, we refer the reader to [11]. Here, we note only that the proof of the Chernoff-style bound of Lemma 2 relies heavily on the expected squared vector length stated in Lemma 1.

3.2 Proof of Theorem 4

Although the proof of our main result makes use of Lemma 2 from [11], the proof strategy thereafter is quite different, and considerably more complex. Whereas their proof of the Johnson-Lindenstrauss lemma made direct use of the Chernoff-style bound of Lemma 2, our result instead uses Theorem 3 to establish a Chernoff-style bound in terms of the local IDs of the cumulative distribution functions associated with $\mathbf{L}$ and $\mathbf{L}_\perp$. It then relies on a careful choice of the parameter β so as to guide the convergence of a double-limit process towards the desired bounds relating $\mathrm{ID}^*_{F_\perp}$ to ID^*_F.

For any vector X, let $Y = X/\|X\|$ be the unit vector obtained by the normalization of X, and let $Y_\perp$ and $X_\perp$ be the projections of X and Y under ψ.

To prove the lower bound, we first consider the cumulative probability of the squared distance distribution after projection by ψ. Since ψ is not known *a priori*, for any squared distance threshold $r > 0$, this probability can be expressed as

$$F_\perp(\sqrt{r_\perp}) \;=\; \mathbf{Pr}[\mathbf{L}_\perp^2 \le r_\perp] \;=\; \mathbf{Pr}[\|X_\perp\|^2 \le r_\perp] \;=\; \mathbf{Pr}\left[\|Y_\perp\|^2 \le \frac{r_\perp}{\|X\|^2}\right].$$

For any choice of $0 < \beta < 1$, this probability can be bounded by

$$\begin{aligned}\mathbf{Pr}[\mathbf{L}_\perp^2 \le r_\perp] &= \mathbf{Pr}\left[\|Y_\perp\|^2 \le \frac{r_\perp}{\|X\|^2} \bigwedge \|Y_\perp\|^2 \ge \beta\frac{t}{m}\right] \\ &\qquad\qquad + \mathbf{Pr}\left[\|Y_\perp\|^2 \le \frac{r_\perp}{\|X\|^2} \bigwedge \|Y_\perp\|^2 < \beta\frac{t}{m}\right] \\ &\le \mathbf{Pr}\left[\beta\frac{t}{m} \le \frac{r_\perp}{\|X\|^2}\right] + \mathbf{Pr}\left[\|Y_\perp\|^2 \le \beta\frac{t}{m}\right] \\ &\le \mathbf{Pr}\left[\mathbf{L}^2 \le \frac{m}{\beta t}r_\perp\right] + \mathbf{Pr}\left[\|Y_\perp\|^2 \le \beta\frac{t}{m}\right] \\ &\le \mathbf{Pr}\left[\mathbf{L}^2 \le \frac{1}{\beta}r\right] + \mathbf{Pr}\left[\|Y_\perp\|^2 \le \beta\frac{t}{m}\right],\end{aligned}$$

where $r_\perp = \frac{t}{m}r$, the expected squared length of the projection under ψ of a vector of squared length r.

Recall that the length of a fixed unit vector after uniform random projection to a t-dimensional space follows the same distribution as the length of a uniform random unit vector after a fixed projection to $\mathbb{R}^t$. Lemma 2 can therefore be applied to yield

$$\begin{aligned}\mathbf{Pr}[\mathbf{L}_\perp^2 \le r_\perp] &\le \mathbf{Pr}\left[\mathbf{L}^2 \le \frac{1}{\beta}r\right] + \beta^{\frac{t}{2}}\left(1 + \frac{t(1-\beta)}{m-t}\right)^{\frac{m-t}{2}} \\ &\le \mathbf{Pr}[\mathbf{L}^2 \le r/\beta] + \beta^{\frac{t}{2}}\left(e^{\frac{(1-\beta)t}{m-t}}\right)^{\frac{m-t}{2}} \\ &\le \mathbf{Pr}[\mathbf{L}^2 \le r/\beta] + \beta^{\frac{t}{2}}e^{\frac{t}{2}(1-\beta)} \;\le\; \mathbf{Pr}[\mathbf{L}^2 \le r/\beta] + \beta^{\frac{t}{2}}e^{\frac{t}{2}}, \qquad (1)\end{aligned}$$

since $0 < \beta < 1$.

Since F and $F_\perp$ are assumed to be continuously differentiable over the range of distances $(0, \epsilon)$, the cumulative distribution functions of $\mathbf{L}^2$ and $\mathbf{L}_\perp^2$ must also be continuously differentiable over $(0, \epsilon^2)$. Let ID□ and ID□$_\perp$ denote the LID values of the cumulative distribution functions of $\mathbf{L}^2$ and $\mathbf{L}_\perp^2$, respectively. We can therefore apply Theorem 3 to obtain

$$\mathbf{Pr}[\mathbf{L}^2 \le r/\beta] \le \mathbf{Pr}[\mathbf{L}^2 \le \delta] \cdot \left(\frac{1}{\beta} \cdot \frac{r}{\delta}\right)^{\underline{\mathrm{ID}\square}_\delta}, \text{ and}$$

$$\mathbf{Pr}[\mathbf{L}_\perp^2 \le r_\perp] \ge \mathbf{Pr}[\mathbf{L}_\perp^2 \le \delta_\perp] \cdot \left(\frac{r_\perp}{\delta_\perp}\right)^{\overline{\mathrm{ID}\square}_{\delta_\perp}},$$

where $\underline{\mathrm{ID}\square}_\delta$ denotes the infimum $\underline{\mathrm{ID}\square}(0,\delta)$ of $\mathrm{ID}\square$ over the range $[0,\delta]$, and where $\overline{\mathrm{ID}\square}_{\delta_\perp}$ denotes the supremum $\overline{\mathrm{ID}\square}_\perp(0,\delta_\perp)$ of $\mathrm{ID}\square_\perp$ over the range $[0,\delta_\perp]$. Here, we assume that the variables have been chosen such $\delta_\perp = \frac{t}{m}\delta$, and furthermore that r/β, $r_\perp$, δ and $\delta_\perp$ are all strictly less than ϵ^2 (this latter condition will be enforced later, as more constraints on these variables are introduced).

Since by construction $\frac{r_\perp}{\delta_\perp} = \frac{r}{\delta}$, substituting the above inequalities into Inequality 1 yields

$$\mathbf{Pr}[\mathbf{L}_\perp^2 \le \delta_\perp] \cdot \left(\frac{r}{\delta}\right)^{\overline{\mathrm{ID}\square}_{\delta_\perp}} \le \mathbf{Pr}[\mathbf{L}^2 \le \delta] \cdot \left(\frac{1}{\beta} \cdot \frac{r}{\delta}\right)^{\underline{\mathrm{ID}\square}_\delta} + \beta^{\frac{t}{2}} e^{\frac{t}{2}} . \tag{2}$$

Consider now a new interpolation parameter c, whose role will be explained below. In terms of r, δ, and c, we fix the parameter β as follows:

$$\beta = \left(\frac{r}{\delta}\right)^c .$$

Note that under these conditions, for any $0 < c < 1$ and $\delta > 0$, choosing r such that $0 < r < \delta$ ensures that $0 < \beta < 1$.

Substitution into Inequality 2 gives

$$\mathbf{Pr}[\mathbf{L}_\perp^2 \le \delta_\perp] \cdot \left(\frac{r}{\delta}\right)^{\overline{\mathrm{ID}\square}_{\delta_\perp}} \le \mathbf{Pr}[\mathbf{L}^2 \le \delta] \cdot \left(\frac{r}{\delta}\right)^{(1-c)\cdot\underline{\mathrm{ID}\square}_\delta} + e^{\frac{t}{2}} \cdot \left(\frac{r}{\delta}\right)^{\frac{ct}{2}} .$$

Next, we balance the contributions of the terms on the right-hand side of the inequality, by choosing c such that $(1-c)\cdot\underline{\mathrm{ID}\square}_\delta = ct/2$. This produces

$$\mathbf{Pr}[\mathbf{L}_\perp^2 \le \delta_\perp] \cdot \left(\frac{r}{\delta}\right)^{\overline{\mathrm{ID}\square}_{\delta_\perp}} \le \left(\mathbf{Pr}[\mathbf{L}^2 \le \delta] + e^{\frac{t}{2}}\right) \cdot \left(\frac{r}{\delta}\right)^{\frac{\underline{\mathrm{ID}\square}_\delta\cdot(t/2)}{\underline{\mathrm{ID}\square}_\delta + t/2}} . \tag{3}$$

Note that from Theorem 1, the existence of the limits ID_F^* and $\mathrm{ID}_{F_\perp}^*$ implies that $F(\sqrt{\delta}) = \mathbf{Pr}[\mathbf{L} \le \sqrt{\delta}] > 0$ and $F_\perp(\sqrt{\delta_\perp}) = \mathbf{Pr}[\mathbf{L}_\perp \le \sqrt{\delta_\perp}] > 0$ whenever δ and $\delta_\perp$ are chosen to be sufficiently small. Taking the logarithms of both sides of Inequality 3, and dividing by $\ln(r/\delta)$, leads us to the following:

$$\overline{\mathrm{ID}\square}_{\delta_\perp} \ge \frac{\underline{\mathrm{ID}\square}_\delta\cdot(t/2)}{\underline{\mathrm{ID}\square}_\delta + t/2} - \frac{\ln(\mathbf{Pr}[\mathbf{L}^2 \le \delta] + e^{\frac{t}{2}})}{\ln(\delta/r)} + \frac{\ln \mathbf{Pr}[\mathbf{L}_\perp^2 \le \delta_\perp]}{\ln(\delta/r)} .$$

Fixing δ and $\delta_\perp$, and letting $r \to 0$, the inequality has the limit

$$\overline{\mathrm{ID}\square}_{\delta_\perp} \ge \frac{\underline{\mathrm{ID}\square}_\delta\cdot(t/2)}{\underline{\mathrm{ID}\square}_\delta + t/2} .$$

Next, letting $\delta \to 0$, we observe that $\delta_\perp \to 0$, $\underline{\mathrm{ID}\square}_\delta \to \mathrm{ID}\square^*$ and $\overline{\mathrm{ID}\square}_{\delta_\perp} \to \mathrm{ID}\square^*_\perp$, and thus

$$\mathrm{ID}\square^*_\perp \geq \frac{\mathrm{ID}\square^* \cdot (t/2)}{\mathrm{ID}\square^* + t/2} \,. \tag{4}$$

Up until now we have assumed that the quantities r/β, $r_\perp$, δ and $\delta_\perp$ were all within the interval $(0, \epsilon^2)$. Here, as $r \to 0$ and $\delta \to 0$, it can be verified that the aforementioned quantities all tend to 0 as well, and that this assumption is therefore eventually justified.

Finally, we transform the bound of Inequality 4 for the ID of the squared distance distributions (before and after projection) to one involving the original distributions. It follows from Theorem 2 that $\mathrm{ID}_F = 2 \cdot \mathrm{ID}\square$ and $\mathrm{ID}_{F_\perp} = 2 \cdot \mathrm{ID}\square_\perp$, from which we see that the lower bound

$$\mathrm{ID}^*_{F_\perp} \geq \frac{\mathrm{ID}^*_F \cdot t}{\mathrm{ID}^*_F + t}$$

holds as required.

We now turn our attention to the proof of the upper bound $\mathrm{ID}^*_{F_\perp} \leq \mathrm{ID}^*_F$. The proof is similar to (but much simpler than) that of the lower bound. Since $\|X\|^2 \geq \|X_\perp\|^2$, for any $r > 0$,

$$\mathbf{Pr}[\mathbf{L}^2 \leq r] \;=\; \mathbf{Pr}[\|X\|^2 \leq r] \;\leq\; \mathbf{Pr}[\|X_\perp\|^2 \leq r] \;=\; \mathbf{Pr}[\mathbf{L}_\perp^2 \leq r] \,.$$

Applying Theorem 3, and choosing $\delta_\perp = \delta > r$, we obtain

$$\mathbf{Pr}[\mathbf{L}^2 \leq \delta] \cdot \left(\frac{r}{\delta}\right)^{\underline{\mathrm{ID}\square}_\delta} \;\leq\; \mathbf{Pr}[\mathbf{L}_\perp^2 \leq \delta] \cdot \left(\frac{r}{\delta}\right)^{\overline{\mathrm{ID}\square}_{\delta_\perp}} \,,$$

Taking the logarithms of both sides, and dividing by $\ln(r/\delta)$, leads us to:

$$\underline{\mathrm{ID}\square}_\delta \;\geq\; \overline{\mathrm{ID}\square}_{\delta_\perp} + \frac{\ln \mathbf{Pr}[\mathbf{L}^2 \leq \delta]}{\ln(\delta/r)} - \frac{\ln \mathbf{Pr}[\mathbf{L}_\perp^2 \leq \delta_\perp]}{\ln(\delta/r)} \,.$$

Fixing δ and $\delta_\perp$, and letting $r \to 0$, the inequality has the limit $\overline{\mathrm{ID}\square}_{\delta_\perp} \leq \underline{\mathrm{ID}\square}_\delta$. Next, letting $\delta \to 0$, we observe that $\delta_\perp \to 0$, $\underline{\mathrm{ID}\square}_\delta \to \mathrm{ID}\square^*$ and $\overline{\mathrm{ID}\square}_{\delta_\perp} \to \mathrm{ID}\square^*_\perp$, and thus $\mathrm{ID}\square^*_\perp \leq \mathrm{ID}\square^*$, which in turn implies that $\mathrm{ID}^*_{F_\perp} \leq \mathrm{ID}^*_F$ as required.

4 Conclusion

Theorem 4 has important implications for the theory and practice of databases, machine learning, data mining, and other areas in which similarity information plays a role. Under a reasonable assumption of the continuity of the local data distribution, random projection in Euclidean vector spaces cannot be relied upon to significantly improve the discriminability of a distance measure as the number of data samples tends to infinity, nor can it be counted upon to greatly alleviate the asymptotic effects of the curse of dimensionality.

To see this, let us assume that we have a reference point Q within the domain of a global data distribution, whose distance distribution has a local intrinsic dimensionality of ID^*_F. For a random projection to a subspace of dimension $t \gg \mathrm{ID}^*_F$, the bounds of Theorem 4 become almost tight, showing that the local ID of the distribution is essentially unchanged after projection. As increasingly larger data samples are drawn from the distribution, the k-nearest neighbor distance r_k tends to 0, and thus the discriminability of the distance measure at r_k tends to ID^*_F. Thus, under this scenario, as the data set size scales, the discriminability of the distance measure over fixed-cardinality neighborhoods is less and less affected by random projection.

On the other hand, when the projection dimension t is of the same order as ID^*_F (or of lower order), Theorem 4 implies that the local ID of the projected distribution is no smaller than t, and no greater than ID^*_F. However, the information loss associated with projection to dimensionalities below that of the intrinsic dimension would make any improvements in discriminability a moot point, as the distance distribution would no longer be well-preserved in the vicinity of Q.

These implications together are of particular importance when randomly projecting data drawn from a mixture of distributions, where the local intrinsic dimensionality can vary greatly from location to location. Theorem 4 indicates that the target dimension for projection should be chosen to be substantially larger than the LID estimate at locations of particular interest or importance. Using existing estimators of LID [3, 17], appropriate target dimensions for dimensional reduction can be determined locally, without the need to construct an explicit embedding of the data, or an explicit representation of the projective subspace.

These conclusions should not be taken to mean that random projection is never capable of reducing the number of latent variables of a user-supplied data set, or improving the discriminability of distances within the set. Theorem 4 addresses only the *asymptotic* effect of random projection on the LID of continuous Euclidean distance distributions, as the number of data instance rises. In applications where the number of data instances can scale into the billions or more, it is possible that these asymptotic effects could become more and more evident. However, any attempt to empirically verify the predictions of Theorem 4 would face significant obstacles, due to the limits in precision and stability exhibited by all existing estimators of ID (not just LID) [3, 17], and due to the difficulty in determining an appropriate locality size — if too large, locality is violated; if too small, there are too few samples for the estimators to converge. For this reason, the development of more effective ID estimation methods for small local data samples is an important topic for future research.

Acknowledgments

Michael E. Houle acknowledges the financial support of JSPS Kakenhi Kiban (B) Research Grant 18H03296. Ken-ichi Kawarabayashi is supported by JSPS Kakenhi Research Grant JP18H05291.

References

1. D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Sys. Sci.*, 66(4):671–687, 2003.
2. N. Ailon and B. Chazelle. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.
3. L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, and M. Nett. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Data Mining and Knowledge Discovery*, 32(6):1768–1805, 2018.
4. R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. *Found. Comput. Math.*, 9(1):51–77, 2009.
5. Y. Bartal, B. Recht, and L. J. Schulman. Dimensionality reduction: Beyond the johnson-lindenstrauss bound. In *Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, pages 868–887, 2011.
6. J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(5):572–575, 1998.
7. F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, 2002.
8. G. Casanova, E. Englmeier, M. E. Houle, P. Kröger, M. Nett, and A. Zimek. Dimensional testing for reverse k-nearest neighbor search. *PVLDB*, 10(7):769–780, 2017.
9. S. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer, 2001.
10. A. Dasgupta, R. Kumar, and T. Sarlos. A sparse Johnson-Lindenstrauss transform. In *STOC*, pages 341–350, 2010.
11. S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.
12. T. de Vries, S. Chawla, and M. E. Houle. Density-preserving projections for large-scale local anomaly detection. *Knowl. Inf. Syst.*, 32(1):25–52, 2012.
13. D. P. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms.* Cambridge University Press, 2009.
14. K. Falconer. *Fractal Geometry: Mathematical Foundations and Applications.* John Wiley & Sons, 2003.
15. C. Faloutsos and I. Kamel. Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In *PODS'94*, pages 4–13, 1994.
16. P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *J. Combinatorial Theory Ser. B*, 44(3):355–362, 1988.
17. M. I. Gomes, L. Canto e Castro, M. I. Fraga Alves, and D. Pestana. Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. *Extremes*, 11:3–34, 2008.
18. P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9(1–2):189–208, 1983.
19. A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS'03*, pages 534–543. IEEE Computer Society, 2003.
20. M. E. Houle. Dimensionality, discriminability, density & distance distributions. In *Proc. ICDMW'13*, pages 468–473, 2013.

21. M. E. Houle. Local intrinsic dimensionality I: An extreme-value-theoretic foundation for similarity applications. In *International Conference on Similarity Search and Applications*, pages 64–79, 2017.
22. M. E. Houle. Local intrinsic dimensionality II: Multivariate analysis and distributional support. In *International Conference on Similarity Search and Applications*, pages 80–95, 2017.
23. M. E. Houle, H. Kashima, and M. Nett. Generalized expansion dimension. In *Proc. ICDMW'12*, pages 587–594, 2012.
24. M. E. Houle, X. Ma, M. Nett, and V. Oria. Dimensional testing for multi-step similarity search. In *ICDM'12*, pages 299–308, 2012.
25. M. E. Houle, X. Ma, and V. Oria. Effective and efficient algorithms for flexible aggregate similarity search in high dimensional spaces. *IEEE TKDE*, 27(12):3258–3273, 2015.
26. M. E. Houle, X. Ma, V. Oria, and J. Sun. Efficient algorithms for similarity search in user-specified projective subspaces. *Information Systems*, 59:2–14, 2016.
27. M. E. Houle and M. Nett. Rank-based similarity search: Reducing the dimensional dependence. *IEEE TPAMI*, 37(1):136–150, 2015.
28. R. Huisman, K. G. Koedijk, C. J. M. Kool, and F. Palm. Tail-index estimates in small samples. *Journal of business and economic statistics*, 19(2):208–216, 2001.
29. P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC*, pages 604–613, 1998.
30. W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *AMS Conference in Modern Analysis and Probability*, pages 189–206, 1982.
31. D. M. Kane and J. Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):4:1–4:23, 2014.
32. D. R. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *STOC'02*, pages 741–750, 2002.
33. K. G. Larsen and J. Nelson. The Johnson-Lindenstrauss lemma is optimal for linear dimensionality reduction. *arXiv.org*, cs.IT, Nov 2014.
34. E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, 2004.
35. P. Mattila. Hausdorff dimension, orthogonal projections and intersections with planes. *Ann. Acad. Sci. Fenn. A Math.*, 1:227–244, 1975.
36. G. Navarro, R. Paredes, N. Reyes, and C. Bustos. An empirical evaluation of intrinsic dimension estimators. *Inf. Syst.*, 64:206–218, 2017.
37. S. Romano, O. Chelly, V. Nguyen, J. Bailey, and M. E. Houle. Measuring dependency via intrinsic dimensionality. In *ICPR*, pages 1207–1212, 2016.
38. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.
39. A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli. Novel high intrinsic dimensionality estimators. *Machine Learning Journal*, 89(1-2):37–65, October 2012.
40. B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.
41. J. Tenenbaum, V. D. Silva, and J. Langford. A global geometric framework for non linear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
42. J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19(6–7):889–899, 2006.
43. E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In *NIPS'02*, pages 505–512, 2002.