Intrinsic Dimension, Degrees of Freedom, Odds and Uniformity: A Unified Perspective

James Bailey, Ricardo Campello and Michael Houle

1 INTRODUCTION

Intrinsic dimensionality serves as a metric for data complexity. A unitless value intuitively assessing the number of hidden factors needed to represent the data. It is widely used in areas such as similarity search, anomaly detection, submanifolds and deep learning representations. **Local intrinsic dimensionality (LID)** assesses the intrinsic dimension when restricted to the neighborhood of an individual reference location

Let F be a CDF of the local distance distribution around a reference point. When F is differentiable at $r > 0$,

$$\mathrm{ID}_F(r) \equiv \frac{rF'(r)}{F(r)} \quad \text{and} \quad \mathrm{LID}(F) \equiv \lim_{r\to 0} \mathrm{ID}_F(r) =: \mathrm{ID}_F^{\star}.$$

- Representation theorem (tail form). For small r, w in the lower tail,

$$\frac{F(r)}{F(w)} = \left(\frac{r}{w}\right)^{\mathrm{ID}_F^{\star}} \cdot A_F(r,w), \quad A_F(r,w) \to 1 \text{ as } r, w \to 0.$$

- Hence the limit tail distribution has CDF and PDF

$$F_w^*(r) = \left(\frac{r}{w}\right)^{\mathrm{ID}_F^{\star}}, \qquad f_w^*(r) = \frac{\mathrm{ID}_F^{\star}}{w}\left(\frac{r}{w}\right)^{\mathrm{ID}_F^{\star}-1}.$$

2 OBJECTIVE & CONTRIBUTIONS

As more applications are found for LID analysis and estimation, the need for a more unified and accessible understanding of LID grows. We explore new interpretations of LID in relation to fundamental concepts in statistics and its applications

Our Contributions:

1. Show LID can be understood as an expected number of degrees of freedom.

2. Establish LID admits rank-based and density-based interpretations as a statistical odds.

3. Use these to reinterpret AUC and Bayes factors via "AUC-odds" and dimensionality gain.

4. Highlight the special role of the uniform distribution as a canonical baseline

3 Statistical Degrees of Freedom (DOF) and LID

Degrees of freedom is a key concept from statistics. However, it is regarded as being difficult to explain. We show its connection to LID for both integer and non-integer scenarios.

- Draw k i.i.d. samples $X_1, \ldots, X_k \sim \mathrm{Uniform}[0, w]$.
- The maximum $X_{(k)} = \max_i X_i$ has

$$F_{X_{(k)}}(x) = \left(\frac{x}{w}\right)^k, \qquad f_{X_{(k)}}(x) = \frac{k}{w}\left(\frac{x}{w}\right)^{k-1}.$$

- Compare with $f_w^*(r)$: this matches when $\mathrm{ID}_F^{\star} = k$.
- Interpretation: $\mathrm{ID}_F^{\star}$ counts the number of i.i.d. uniform draws (DOF) whose max yields the observed tail shape.

- Let K be a positive-integer-valued r.v. (e.g., degenerate, Poisson, geometric, or a mixture).
- Conditioned on $K = k$, $X = \max\{X_1, \ldots, X_K\}$ with $X_i \sim \mathrm{Uniform}[0, w]$ has $f(x \mid K = k) = \frac{k}{w}\left(\frac{x}{w}\right)^{k-1}$.
- Consensus distribution via weighted geometric averaging (logarithmic pooling) over K yields

$$f_{w,K}(x) = \frac{\mu_K}{w}\left(\frac{x}{w}\right)^{\mu_K - 1}, \qquad F_{w,K}(x) = \left(\frac{x}{w}\right)^{\mu_K},$$

where $\mu_K = \mathbb{E}[K]$.

- Therefore $\mathrm{ID}_F^{\star} = \mu_K$ can be non-integer: an expected DOF.

4 Rank based odds interpretation of LID

Given two random variables X and Y, we can test the probability and odds that X is larger than Y. This assesses the degree to which X 'dominates' Y, and can be expressed in terms of their LIDs.

Let $X \sim F_w^*$ and $Y \sim G_w^*$ on $[0, w]$ with $F(x) = \left(\frac{x}{w}\right)^{\alpha}$ and $G(x) = \left(\frac{x}{w}\right)^{\beta}$, where $\alpha = \mathrm{ID}_F^{\star}$, $\beta = \mathrm{ID}_G^{\star}$.

$$\mathbb{P}[X > Y] = \frac{\alpha}{\alpha + \beta},$$
$$\mathrm{Odds}(X > Y) = \frac{\alpha}{\beta}.$$

Uniform baseline: if $Y \sim \mathrm{Uniform}[0, w]$ $(\beta = 1)$, then

$$\mathbb{P}[X > Y] = \frac{\alpha}{\alpha + 1}, \qquad \mathrm{Odds}(X > Y) = \alpha = \mathrm{ID}_F^{\star}.$$

5 Density based odds interpretation of LID

We can also compare the local densities of F and G at a boundary point w. If G is uniform then the LID is interpretable as the posterior odds that a sample at the tail boundary w was selected from F instead of G.

With equal priors $\mathbb{P}(F) = \mathbb{P}(G)$ and densities f, g,

$$\mathrm{Odds}(F \mid r) = \frac{\mathbb{P}(F \mid r)}{\mathbb{P}(G \mid r)} = \frac{f(r)}{g(r)} = \frac{\mathrm{ID}_F^{\star} F(r)/r}{\mathrm{ID}_G^{\star} G(r)/r} = \frac{\alpha}{\beta} \cdot \frac{F(r)}{G(r)}.$$

At the tail boundary $r = w$ where $F(w) = G(w) = 1$,

$$\mathrm{Odds}(F \mid w) = \frac{\alpha}{\beta}; \quad \text{with G uniform, } \mathrm{Odds}(F \mid w) = \alpha.$$

Thus, LID equals the posterior odds at the tail boundary against a uniform reference.

Re-interpreting area under the ROC Curve (AUC) using LID

- Classic interpretation: $\mathrm{AUC} = \mathbb{P}[X > Y]$ for positive score X and negative score Y.
- Canonical pair: take $Y \sim \mathrm{Uniform}[0, w]$ and $X \sim F_w^*$ with $\alpha = \mathrm{ID}_F^{\star}$.
- Then $\mathrm{AUC} = \frac{\alpha}{\alpha + 1}$ so $\alpha = \frac{\mathrm{AUC}}{1 - \mathrm{AUC}}$.

Model 1 AUC	Model 2 AUC	$\mathrm{ID}_2^* / \mathrm{ID}_1^*$
0.500	0.505	1.020
0.500	0.665	1.985
0.900	0.948	2.026
0.990	0.995	2.010

Gains in AUC can be interpreted as changes in intrinsic dimensionality.

When AUC is high, a minor change in AUC can produce a relatively large change in dimensionality.

So minor gains at high AUC may reflect large changes in discriminative performance!