NJIT New Jersey Institute of Technology

Dimensionality-Aware Analysis of Local Intrinsic Structure in Multimedia Data

Michael E. Houle
michael.houle@njit.edu

MMI 2026
2nd NJIT Workshop on Multimedia Intelligence
May 21–22, 2026

Introduction

Locality in multimedia analysis:

- Similarity, indexing, retrieval, and relevance.
- Representations and choices of features.
- Multimodal fusion and alignment of representations.
- Anomalies and outliers.

Although the results of analysis vary across localities . . .

- Characterizations
- Complexity
- Semantics

. . . manifold models of data impose a uniform intrinsic dimensionality. (!!)

Introduction

Intrinsic Dimensionality (ID):

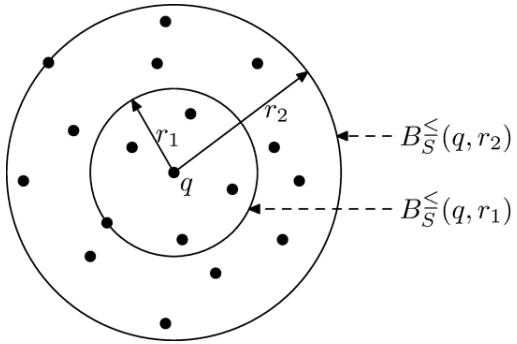
- Effective number of features needed to characterize data.
- Local vs global ID.
- Focus of this talk: Local Intrinsic Dimensionality (LID).

Early uses of LID variants in similarity-based analysis:

- Correlation between query 'difficulty' and high LID estimates. (Aumüller & Ceccarello, 2019)
- Analysis of computational complexity of search indices. (Cover Tree, Sash, RCT, ...)
- LID-based early termination heuristics for indexing. (H. et al, 2012; Casanova et al., 2017)
- Theoretical analysis of projection-based outlier detection. (de Vries et al., 2012)
- Correlation between 'outlierness' and high LID estimates. (H., Schubert, Zimek 2018)

Local Intrinsic Dimensionality

Generalized Expansion Dimension



Generalized Expansion Dimension (GED):

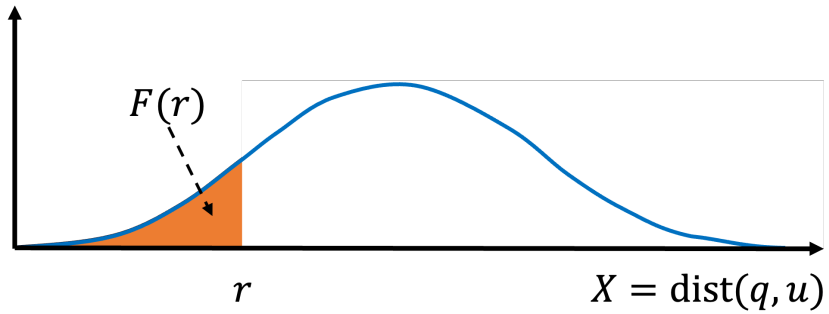
- Motivation: two balls of differing radii r_1 and r_2 in $\mathbb{R}^m$.
- Dimension m can be deduced from ratios of volumes:

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

- When $r_2 = 2r_1 \longrightarrow$ Expansion Dimension.
- V_1 and V_2 estimated by the numbers of points contained in the two balls.

Local Intrinsic Dimensionality

Motivation



Idea: replace GED volumes by probability measure.

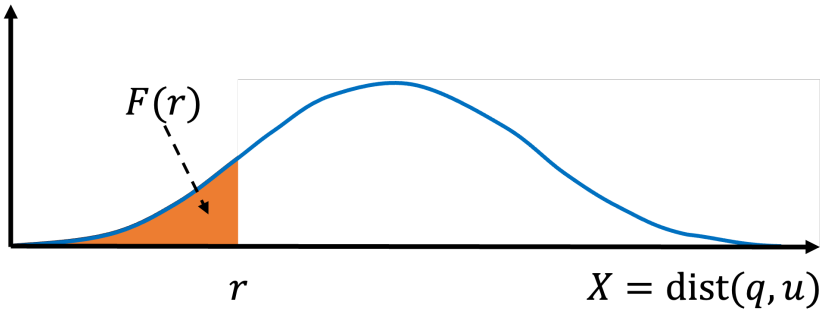
- Initial theoretical development: 2013–2017.
- Random variable $\mathbf{X}$ with support $[0, \infty)$.
- Cumulative distribution function F defined as

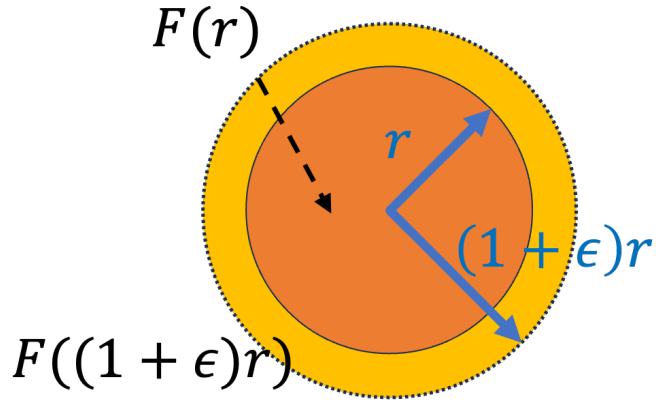
$$F(r) = \mathsf{Pr}[\mathbf{X} \leq r] \ .$$

- Smoothness condition: F is continuously differentiable over ranges of interest.

Local Intrinsic Dimensionality

Formal Definition





For continuous random distance variables:

- Volume is analogous to probability measure.
- r_1 and r_2 can be allowed to tend to a single value r.

Definition (Local Intrinsic Dimensionality)

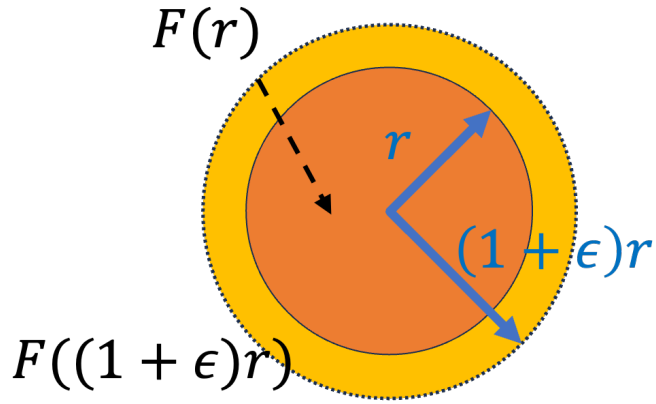
The *local intrinsic dimension* of $\mathbf{X}$ at distance r is

$$\mathrm{IntrDim}_{\mathbf{X}}(r) \triangleq \lim_{\epsilon \to 0^+} \frac{\ln\left(F\left((1+\epsilon)r\right)/F(r)\right)}{\ln(1+\epsilon)},$$

wherever the limit exists.

Local Intrinsic Dimensionality

Indiscriminability



Discriminative power of a distance measure:

- How does probability measure $F(r)$ change as r increases?
- Large relative change in probability measure $\Longrightarrow \mathbf{X}$ indiscriminative at distance r.

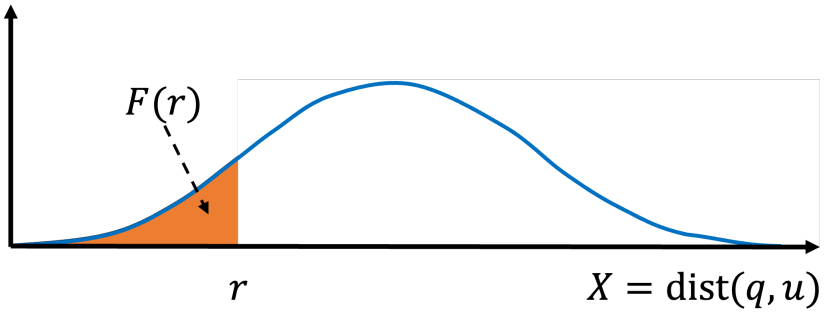
High indiscriminability implies:

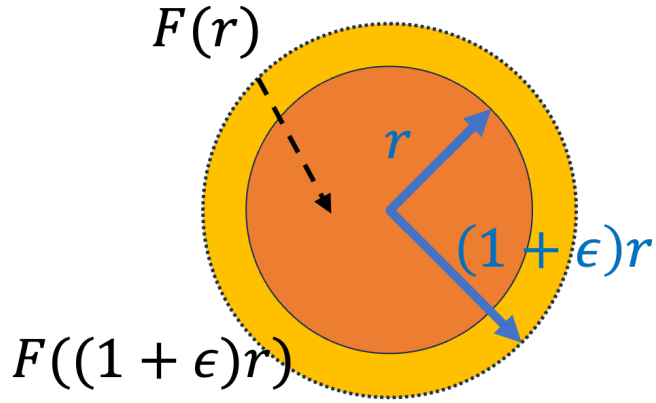
- Lack of robustness to measurement error in underlying features.
- Higher cost when expanding search radius.

Data concentration effect!

Local Intrinsic Dimensionality

Formal Definition of Indiscriminability





Definition (Indiscriminability)

The *indiscriminability* of $\mathbf{X}$ at distance r is

$$\begin{aligned}
&\mathrm{InDiscr}_X(r) \\
&\quad \triangleq \lim_{\epsilon \to 0^+} \left[\frac{(F((1+\epsilon)r) - F(r))}{F(r)} \Big/ \frac{((1+\epsilon)r - r)}{r} \right] \\
&\quad = \lim_{\epsilon \to 0^+} \frac{F((1+\epsilon)r) - F(r)}{\epsilon \cdot F(r)},
\end{aligned}$$

wherever the limit exists.

Local Intrinsic Dimensionality

LID Formula

Theorem (LID Formula)

If F is positive and continuously differentiable over an open interval containing r, then

$$\mathrm{ID}_F(r) \triangleq \frac{rF'(r)}{F(r)} = \mathrm{IntrDim}_{\mathbf{X}}(r) = \mathrm{InDiscr}_{\mathbf{X}}(r)\ .$$

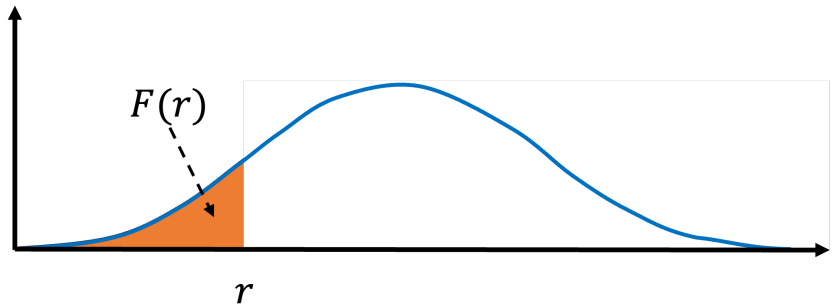
Local ID and indiscriminability are in fact the same!

- Can solve for limits using l'Hôpital's rule.
- If the limits exist, then $\mathrm{ID}_F(r) \geq 0$.

High LID $\Leftrightarrow$ Distance concentration
characterizes the
Curse of Dimensionality

Local Intrinsic Dimensionality

Interpretations of LID Formula



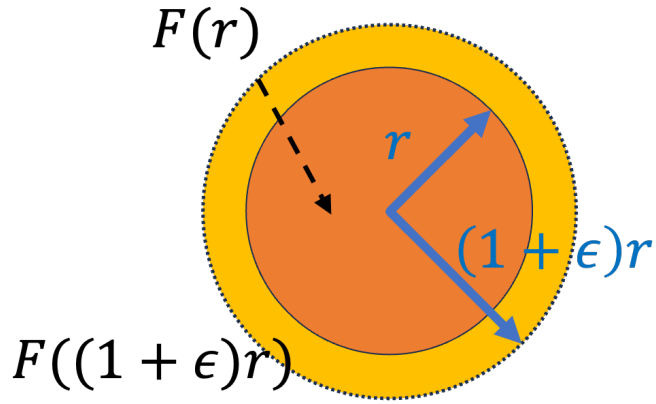
LID theory applies to any smooth growth function F.

- $F(0) = 0$.
- $F(r)$ increases monotonically over some neighborhood, $0 \leq r < \epsilon$.
- F is continuously differentiable over $[0, \epsilon)$.

F does not need to be the CDF of a probability distribution.

Local Intrinsic Dimensionality

Interpretations of LID Formula



LID as a normalized derivative:

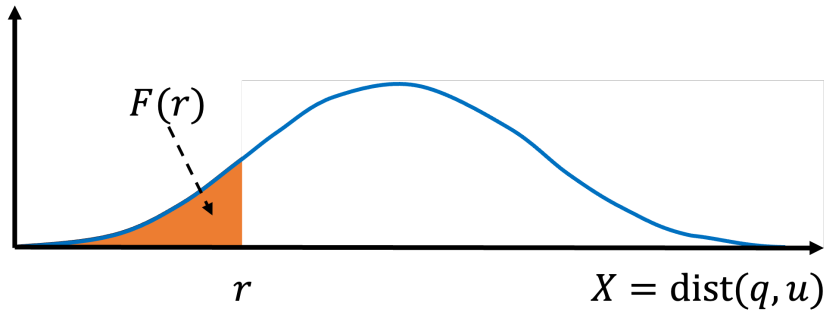
$$\mathrm{ID}_F(r) = F'(r) \Big/ \frac{F(r)}{r}\,.$$

Interpretation:

- Instantaneous rate of change $F'(r)\ldots$
- $\ldots$ normalized by cumulative rate of change $\frac{F(r)}{r}$.
- Measures the 'explosiveness' of growth.
- Unitless, but not scale-free. (Why?)

Local Intrinsic Dimensionality

Asymptotic LID

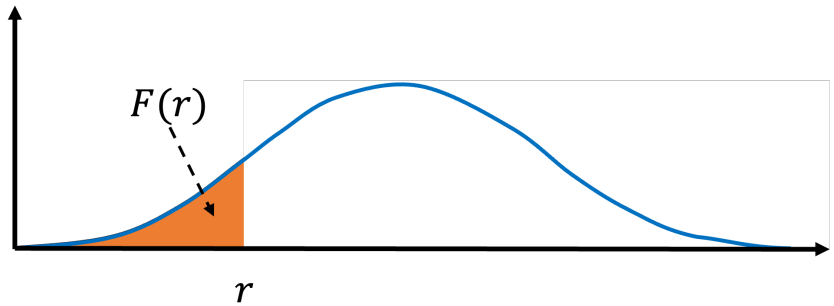


The question of scale:

- At which distance range should we assess ID_F?
- In practice, not much guidance for choosing r!
- Distributional interpretation: as the number of data samples increases, k-NN distances tend to zero.
- Given n samples, expected number in r-neighborhood is $n \cdot F(r)$.
- Letting $r \to 0$ leads to a local characterization of the behavior of F.

Local Intrinsic Dimensionality

Asymptotic LID



Asymptotic formulation of LID:

- Smooth growth tendency as the neighborhood radius vanishes:

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

- Can refer to ID_F^* simply as *intrinsic dimensionality*, and to $\mathrm{ID}_F(r)$ as *indiscriminability at distance* r.

Characterization of Distance Distributions

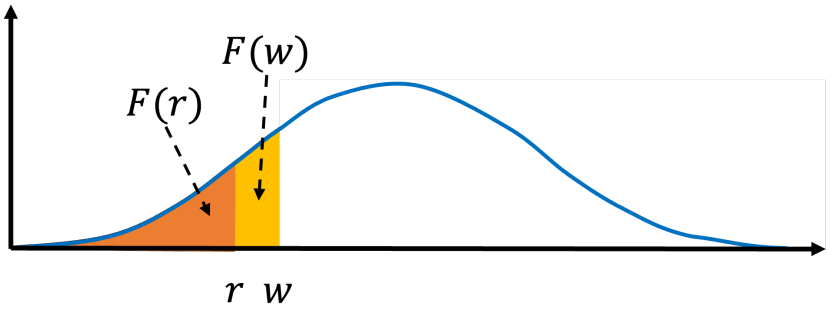
ID Representation Theorem

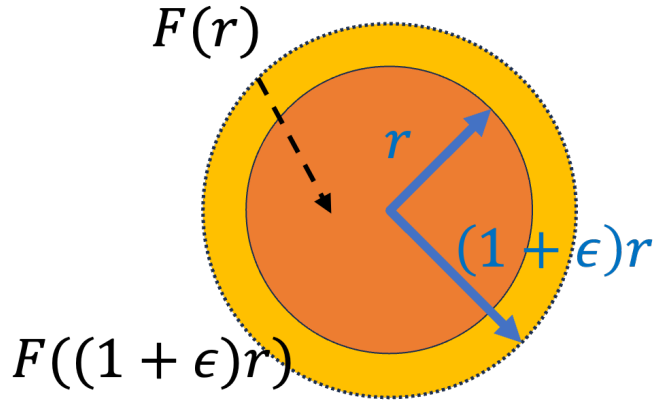
Intrinsic dimensionality at distance r:

- Order of magnitude (or scale) of the growth rate of F in the vicinity of r.
- Contains sufficient information about the distribution to allow its reconstruction.
- Can represent $F(r)$ as an adjustment of F at a distance w where $F(w)$ is known, together with a local measure of ID....

Characterization of Distributions

ID Representation Theorem





Theorem (ID Representation)

Assume that ID_F^* *exists. Then for any* r, w *such that* F *is positive and continuously differentiable everywhere over an open interval containing* $[\min\{r,w\}, \max\{r,w\}]$,

$$\frac{F(r)}{F(w)} = \left(\frac{r}{w}\right)^{\mathrm{ID}_F^*} \cdot A_{F,w}(r), \text{ where}$$

$$A_{F,w}(r) \triangleq \exp\left(\int_r^w \frac{\mathrm{ID}_F^* - \mathrm{ID}_F(t)}{t}\,\mathrm{d}t\right).$$

Characterization of Distance Distributions

ID Representation Theorem

Theorem (ID Representation)

Assume that ID_F^* *exists. Then for any* r, w *such that* F *is positive and continuously differentiable everywhere over an open interval containing* $[\min\{r,w\}, \max\{r,w\}]$,

$$\begin{aligned} \frac{F(r)}{F(w)} &= \left(\frac{r}{w}\right)^{\mathrm{ID}_F^*} \cdot A_{F,w}(r), \text{ where} \\ A_{F,w}(r) &\triangleq \exp\left(\int_r^w \frac{\mathrm{ID}_F^* - \mathrm{ID}_F(t)}{t}\,\mathrm{d}t\right). \end{aligned}$$

Interesting when $w \to 0$ and r/w is bounded.

- If so, then $A_{F,w}(r) \to 1$.
- Remaining factors give an approximation of $F(r)$ in the vicinity of w.

Characterization of Distance Distributions

Convergence of ID Representation

$A_{F,w}(r)$ accounts for the discrepancy between the true distribution and the limit distribution.

- As $w \to 0$, with r chosen in the vicinity of w, the ID Representation Theorem essentially states that F behaves as

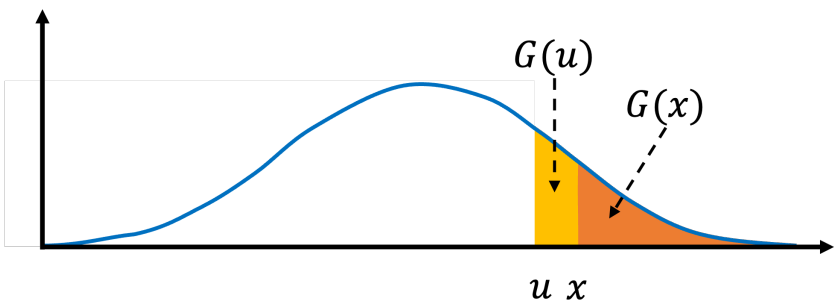
$$F(r) \sim F(w) \cdot \left(\frac{r}{w}\right)^{\mathrm{ID}_F^*}.$$

- Variant of the statistical theory of extreme values ....

ID and Extreme Value Theory

Extreme Value Theory

- Modeling the extreme behavior of stochastic processes.
- Applications in civil engineering, operations research, risk assessment, material sciences, bioinformatics, geophysics, ...
- Extreme events occur in the distributional tails.
- For many decades, almost all attention has been given to distributions with unbounded upper tails (heavy-tailed).

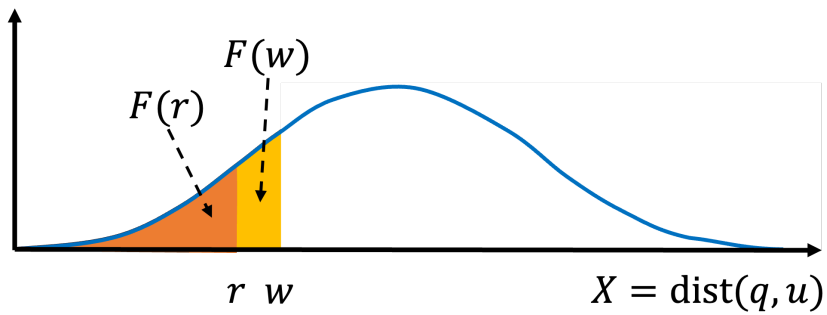


ID and Extreme Value Theory

Query Result Distribution

Query results are extreme events.

- The probability of appearance within the query result is that of falling within the lower tail of the distance distribution (short-tailed).
- EVT modeling was (first?) applied to the distribution of query results in multimedia indexing in 2013 (Furon & Jégou).
- In the context of distance distributions, EVT can be derived in terms of LID.



LID Estimation in Practice

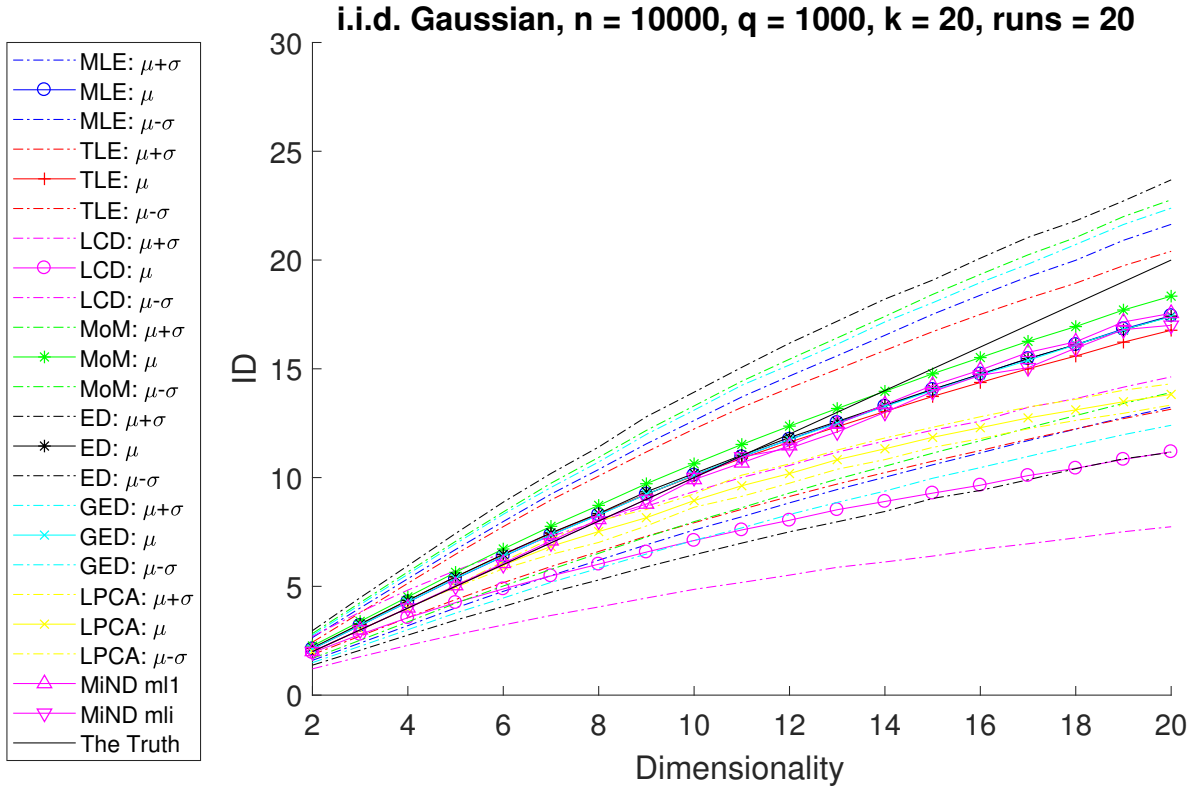
Practical estimators of LID exist.

- MLE estimator:

$$\widehat{\mathrm{ID}}_F^* = -\left(\frac{1}{k}\sum_{i=1}^{k}\ln\frac{r_i}{r_k}\right)^{-1}.$$

- Here, r_i is the i-th smallest distance in the sample.
- Given a neighborhood, very fast to calculate.
- Originally discovered by Hill (1975) in the setting of EVT.
- Rediscovered many times from different starting assumptions (MLE on the LID framework, Poisson point process...).
- Many other estimators of LID have been developed (e.g. ED, GED, 2-NN, MoM, TLE, ...).

LID Estimation in Practice

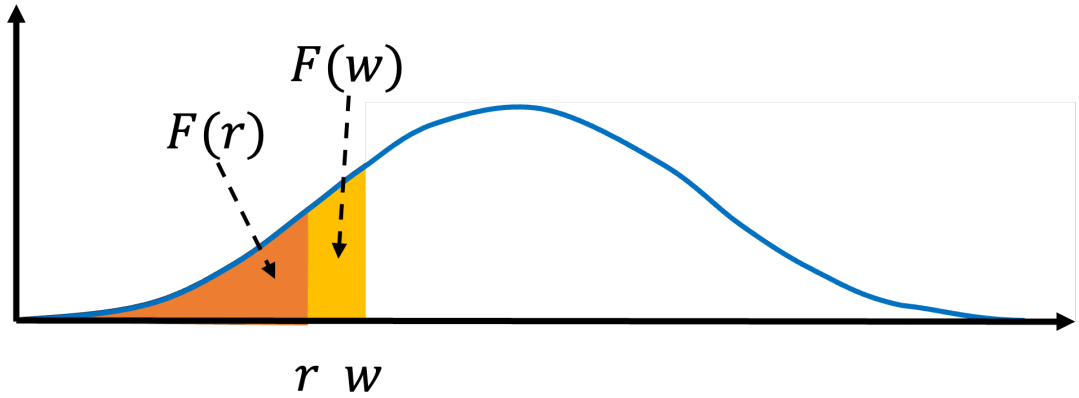


Comparison of various LID estimators (Amsaleg et al., 2022)

- Gaussian mixtures (i.i.d.) of varying dimensionalities.
- Means and standard deviations of LID estimates.

Note negative bias trend as dimensionality increases!

LID Estimation and Scale



General considerations for LID estimation:

- LID is a property of smooth growth functions F.
- For convenience, we assume that F is the CDF of a distribution on $[0, \infty)$.
- Data points are treated as samples from this distribution.
- In practice, we assume a threshold $w > 0$, the tail boundary or neighborhood radius.
- We are interested in those samples that fall within the (lower) distributional tail, $[0, w]$.

LID Estimation and Scale

Asymptotic Tail Distribution

Tail distribution:

- Consider F_w, the restriction of F to the tail $[0, w]$:

$$F_w(r) \triangleq F(r|\, 0 \le r \le w) = \frac{F(r)}{F(w)} .$$

- $F_w(w) = 1 \implies F_w$ satisfies the conditions of a distribution.
- When r and w tend 'nicely' to 0, the ID Representation theorem states that

$$F_w(r) \approx \left(\frac{r}{w}\right)^{\mathrm{ID}^*_F} .$$

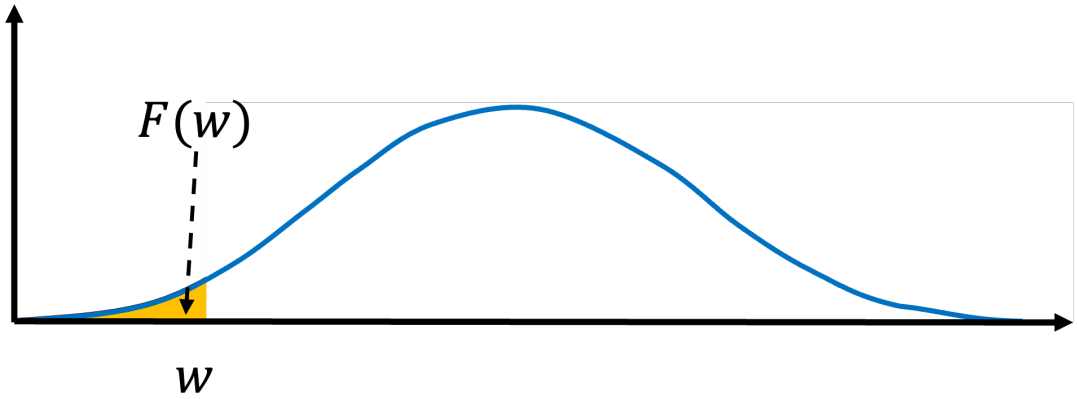
- We focus on the asymptotic tail distribution:

$$F^*_w(r) \triangleq \left(\frac{r}{w}\right)^{\mathrm{ID}^*_F} .$$

- Note that $\mathrm{ID}^*_{F^*_w} = \mathrm{ID}^*_{F_w} = \mathrm{ID}^*_F$.

LID Estimation and Scale

Choice of Tail Boundary

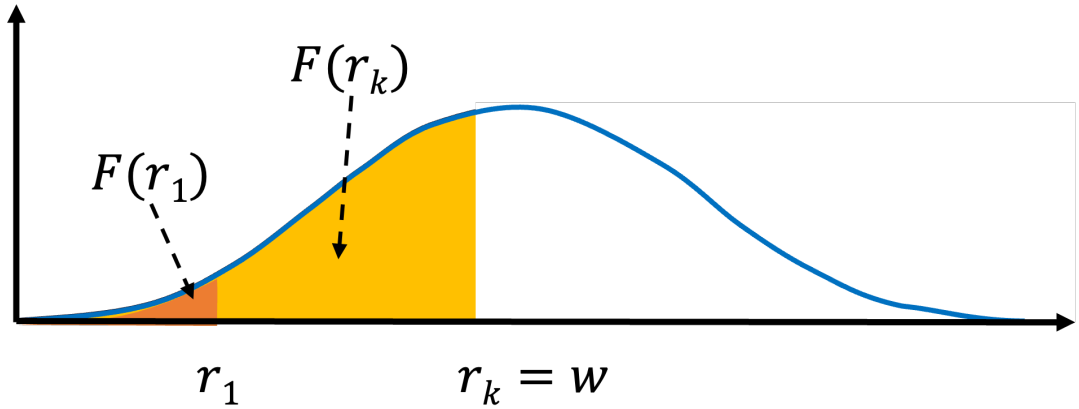


Fix the tail boundary, w?

- Dataset usually is taken with respect to the overall distribution F.
- If w is chosen too small, we may have too few data samples in the tail.
- Estimator becomes more unstable.
- We may not get any samples at all!

LID Estimation and Scale

Choice of Tail Boundary

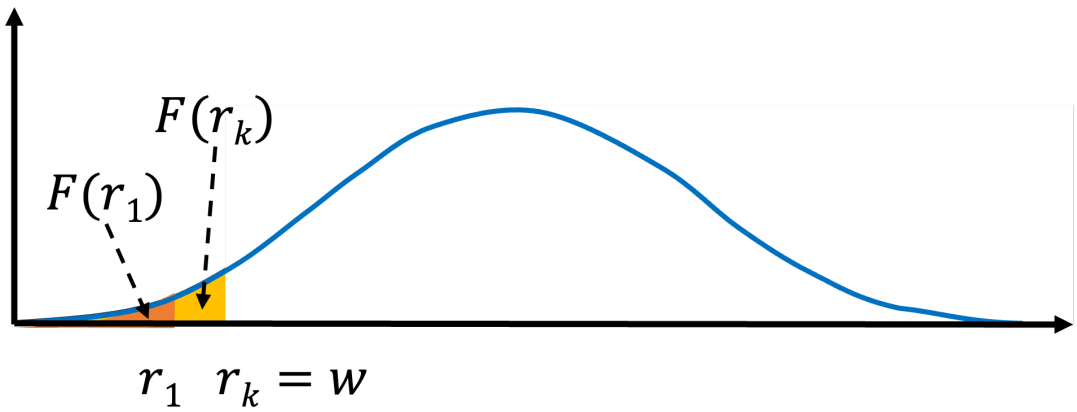


Fix the number of samples, k?

- Sample values $r_1 \leq r_2 \leq \ldots \leq r_k \longrightarrow$ 'distances'.
- Typically, tail threshold is set to $w = r_k$.
- If $r_k = w$ is too large, the tail distribution F_w may not be well-approximated by its asymptotic distribution $F_w(r)$.

LID Estimation and Scale

Sampling Rate

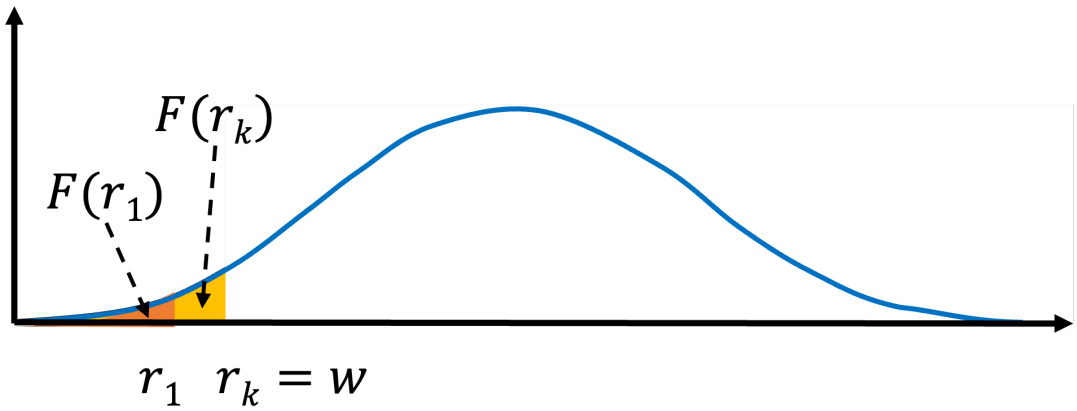


Sampling rate determines the tail probability.

- Assume: m data subsamples taken from dataset of size n.
- Tail boundary w is associated with probability $F(w)$ of a subsample falling in the tail.
- Expected number of subsamples in the tail: $m \cdot F(w)$.
- Alternatively, if the tail is determined by k smallest distance values, then this probability is approximately $F(r_k) \approx k/m$.

LID Estimation and Scale

Local vs Global



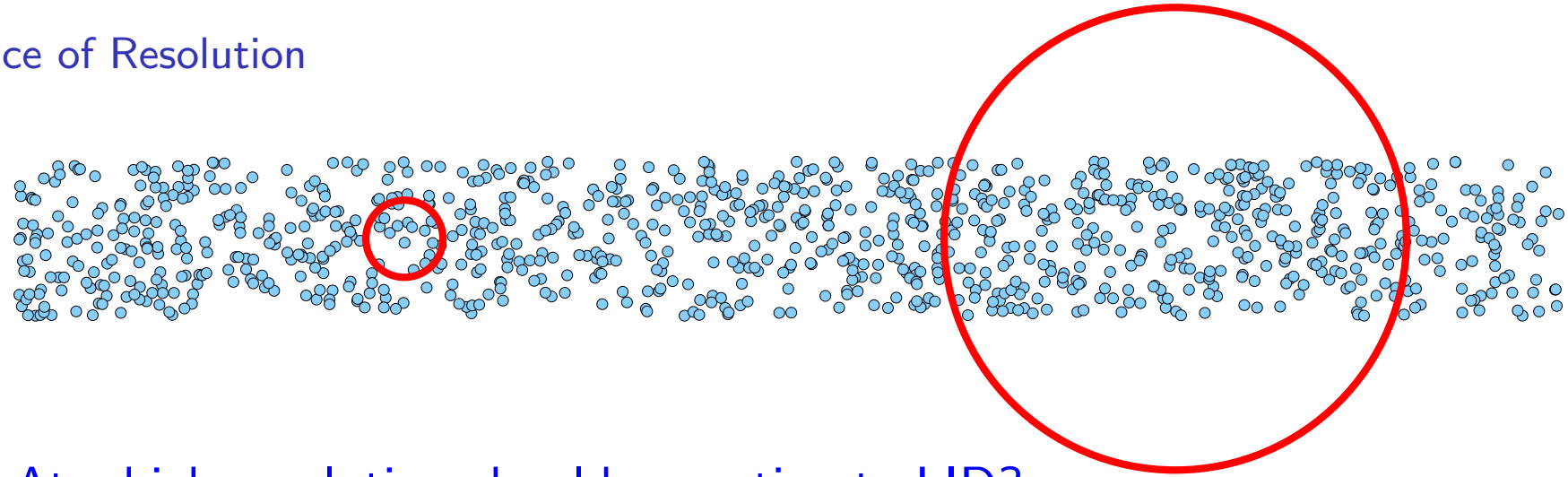
Implications for LID estimation when k is fixed:

- Truly local estimation requires larger amounts of data.
- Tail estimates from small samples produce estimates reflecting global characteristics.

The number of data samples
determines the **resolution** (or **scale** or the **degree of locality**)
of the LID estimate.

LID Estimation and Scale

Choice of Resolution

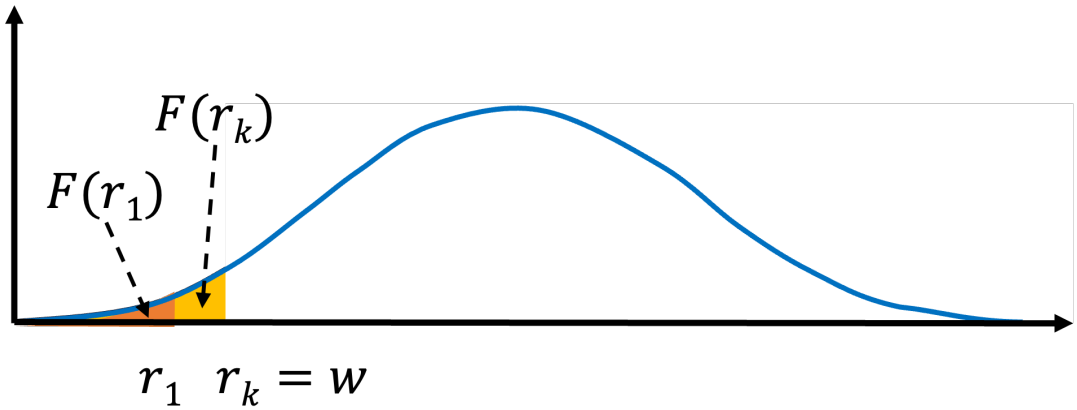


At which resolution should we estimate LID?

- For this example:
  - High resolution $\longrightarrow \widehat{\mathrm{ID}}_F^* \approx 2$.
  - Low resolution $\longrightarrow \widehat{\mathrm{ID}}_F^* \approx 1$.
  - Asymptotically $\longrightarrow \mathrm{ID}_F^* = 2$?
- We cannot know the theoretical LID unless we know F. (In practice, we rarely do!)

LID Estimation and Scale

Choice of Resolution



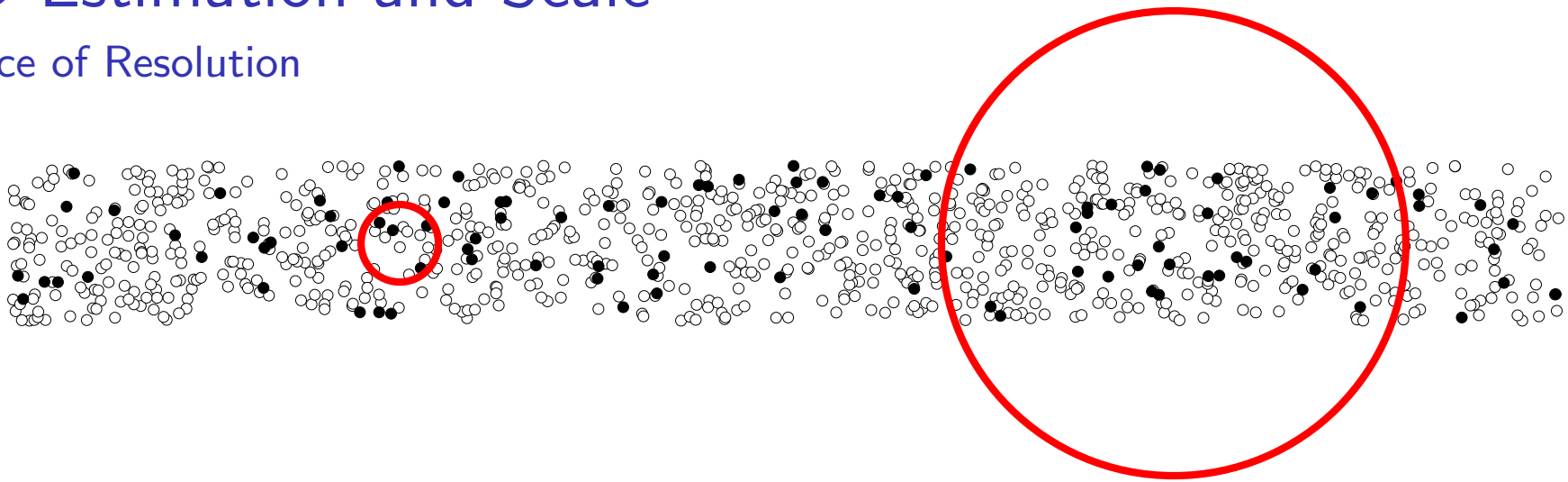
The choice of resolution depends on the practical application.

Choose subsample size m and/or number of tail samples k to determine the scale of interest?

- To characterize a group of size s, should choose $\frac{k}{m} \lesssim \frac{s}{n}$, assuming uniform subsampling. (Why?)
- For ID of individual small clusters or query results: $s \approx$ cluster size / query result size.
- For ID of classes or groups of clusters: $s \approx$ class size / group size.

LID Estimation and Scale

Choice of Resolution



Choose subsample size m so that k-NN computation is affordable?

- Common in (for example) deep learning applications $\longrightarrow m =$ mini-batch size.
- Assume k is fixed, so that LID estimation is stable.
- Low $m \longrightarrow \frac{k}{m}$ is high $\longrightarrow \frac{s}{n}$ is high $\longrightarrow$ low resolution!
- The estimate risks reflecting global characteristics rather than local characteristics.

Density-Ratio-Based Outlier Detection

Introduction

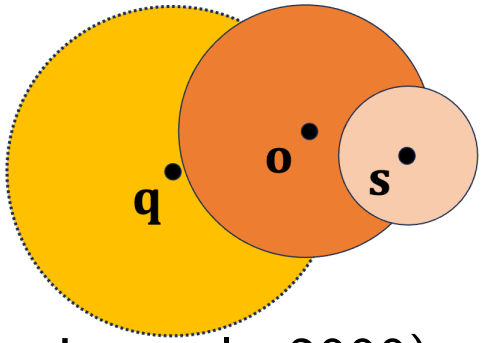
Nonparametric outlier detection

- Low densities near the query point $\longrightarrow$ outlierness.
- Early methods:
  - kNN: density as reciprocal of k-th NN distance.
  - LOF: average of ratio of densities at query & its neighbors.
- Many variants have been proposed over the past decades ...
- ... but kNN and LOF have consistently been top performers!

None take local variation in ID explicitly into account.

Density-Ratio-Based Outlier Detection

SLOF



Simplified LOF (SLOF) (Schubert et al., 2014):

- Variant of Local Outlier Factor (LOF) (Breunig et al., 2000).
- Average of ratios of two 'densities':
  based at query $\mathbf{q}$, & based at neighbor $\mathbf{o}$.

$$\mathrm{SLOF}_k(\mathbf{q}) \triangleq \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k(\mathbf{q})} \frac{\mathrm{slrd}_k(\mathbf{o})}{\mathrm{slrd}_k(\mathbf{q})},$$

- *Inverse k-NN distance* (slrd):
  simplification of LOF's local reachability distance, lrd.

$$\mathrm{slrd}_k(\mathbf{p}) \triangleq \frac{1}{k\_\mathrm{dist}(\mathbf{p})}.$$

- lrd and slrd are not actually densities!
  (Units should be of mass over volume?)

$$\frac{k}{c(k\_\mathrm{dist}(\mathbf{p}))^m} \quad ??$$

Dimensionality-Aware Outlier Model

Introduction

Dimensionality-Aware Outlier Detection (Anderberg et al., 2024):

- DAO: first known method for outlier detection that adapts to variations in LID.
- Derived by applying the theory of LID to local density ratios.
- Can be regarded as a reformulation of (simplified) LOF.
- Verification through experimental analysis on > 800 datasets.

Dimensionality-Aware Outlier Model

ALDR Outlierness

Probability-density-based outlierness:

- Given query $\mathbf{q}$, its probability density within radius ϵ is probability over volume:

$$F_{\mathbf{q}}(\epsilon)/V_{\mathbf{q}}(\epsilon).$$

- Next, consider a neighbor $\mathbf{o}$ drawn at random.
- Take expected ratio of densities. (Volumes cancel!)

$$\mathop{\mathbb{E}}_{\mathbf{o}\in B_{\mathbf{q}}(\epsilon)}\left[\frac{F_{\mathbf{o}}(\epsilon)}{F_{\mathbf{q}}(\epsilon)}\right]$$

- Let $\epsilon \to 0$, obtaining a ratio of infinitesimals.

Asymptotic local expected density ratio (ALDR):

$$\mathrm{ALDR}(\mathbf{q}) \triangleq \lim_{\epsilon\to 0^+} \mathop{\mathbb{E}}_{\mathbf{o}\in B_{\mathbf{q}}(\epsilon)}\left[\frac{F_{\mathbf{o}}(\epsilon)}{F_{\mathbf{q}}(\epsilon)}\right].$$

Dimensionality-Aware Outlier Model

Outlierness and LID

Reformulation of ALDR:

- Decouple radius of selection from radius of density ratio.

$$\mathrm{ALDR}'(\mathbf{q}) \triangleq \lim_{\epsilon \to 0^+} \mathop{\mathbb{E}}_{\mathbf{o} \in B_{\mathbf{q}}(\epsilon)} \left[\lim_{\gamma \to 0^+} \frac{F_{\mathbf{o}}(\gamma)}{F_{\mathbf{q}}(\gamma)} \right].$$

- Sparse outlierness $\longrightarrow$ $\mathrm{ALDR}, \mathrm{ALDR}' > 1$.

- Using the LID Representation Theorem, can prove that

$$\mathrm{ALDR}'(\mathbf{q}) \triangleq \lim_{\epsilon \to 0^+} \mathop{\mathbb{E}}_{\mathbf{o} \in B_{\mathbf{q}}(\epsilon)} \left[\lim_{p \to 0^+} \left(\frac{\delta_{\mathbf{q}}(p)}{\delta_{\mathbf{o}}(p)} \right)^{\mathrm{ID}^*_{F_{\mathbf{o}}}} \right],$$

where $\delta_{\mathbf{q}}(p)$ and $\delta_{\mathbf{o}}(p)$ are the radii at which $F_{\mathbf{q}}$ and $F_{\mathbf{o}}$ achieve probability p.

Dimensionality-Aware Outlier Model

DAO Estimator

$$\mathrm{ALDR}(\mathbf{q}) \triangleq \lim_{\epsilon \to 0^+} \mathop{\mathbb{E}}_{\mathbf{o} \in B_{\mathbf{q}}(\epsilon)} \left[\frac{F_{\mathbf{o}}(\epsilon)}{F_{\mathbf{q}}(\epsilon)} \right],$$

$$\mathrm{ALDR}'(\mathbf{q}) \triangleq \lim_{\epsilon \to 0^+} \mathop{\mathbb{E}}_{\mathbf{o} \in B_{\mathbf{q}}(\epsilon)} \left[\lim_{p \to 0^+} \left(\frac{\delta_{\mathbf{q}}(p)}{\delta_{\mathbf{o}}(p)} \right)^{\mathrm{ID}^*_{F_{\mathbf{o}}}} \right].$$

Estimator of ALDR':

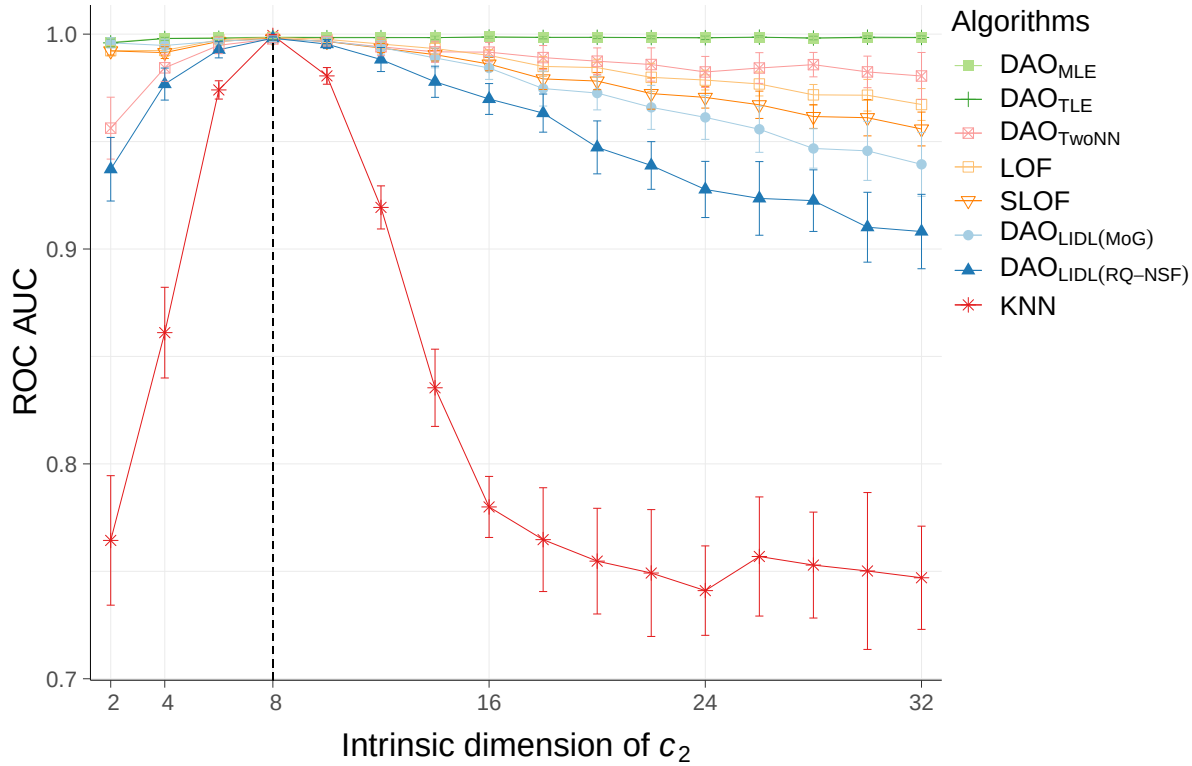
$$\mathrm{DAO}_k(\mathbf{q}) \triangleq \frac{1}{k} \sum_{\mathbf{o} \in \mathrm{NN}_k \mathbf{q}} \left(\frac{k\_\mathrm{dist}(\mathbf{q})}{k\_\mathrm{dist}(\mathbf{o})} \right)^{\widehat{\mathrm{ID}^*_{F_{\mathbf{o}}}}}.$$

Notes:

- DAO estimates ALDR as well! (Set $\epsilon = \delta_{\mathbf{q}}(p)$ and $p = k/n$.)
- SLOF implicitly assumes that all neighbors have $\mathrm{ID}^*_{F_{\mathbf{o}}} = 1$.

Experimental Analysis

Dimensionality Awareness



Outlier detection task on 480 synthetic 2-cluster datasets in $\mathbb{R}^{32}$.

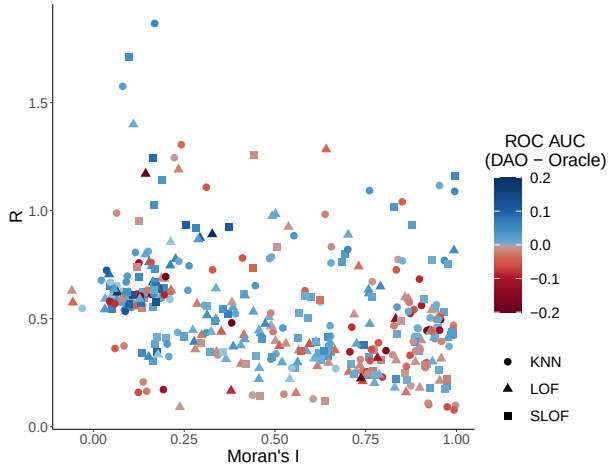
- Cluster c_1: Gaussian in a randomly-oriented 8-dim subspace.
- Cluster c_2: Gaussian of dimension between 2 and 32.
- For each dimension, 30 datasets of 1600 points each.
- Extreme points labeled as outliers (5% of total).

Experimental Analysis

Relating Performance with Variation in LID

Outlier detection task on 393 real datasets.

Repository	Features	Size	Outliers	Datasets
Campos *et al.* [2016]	[5, 259]	[50, 49534]	[3%, 36%]	15
Marques *et al.* [2020]	[10, 649]	[100, 910]	[1%, 10%]	3
Rayana [2016]	[6, 274]	[129, 7848]	[2%, 36%]	11
Goldstein & Uchida [2016]	[27, 400]	[367, 49534]	[2%, 3%]	3
Emmott *et al.* [2016]	[7, 128]	[992, 515129]	[9%, 50%]	11
Kandanaarachchi *et al.* [2020]	[2, 649]	[72, 9083]	[1%, 3%]	350
Overall	[2, 649]	[50, 515129]	[1%, 50%]	393

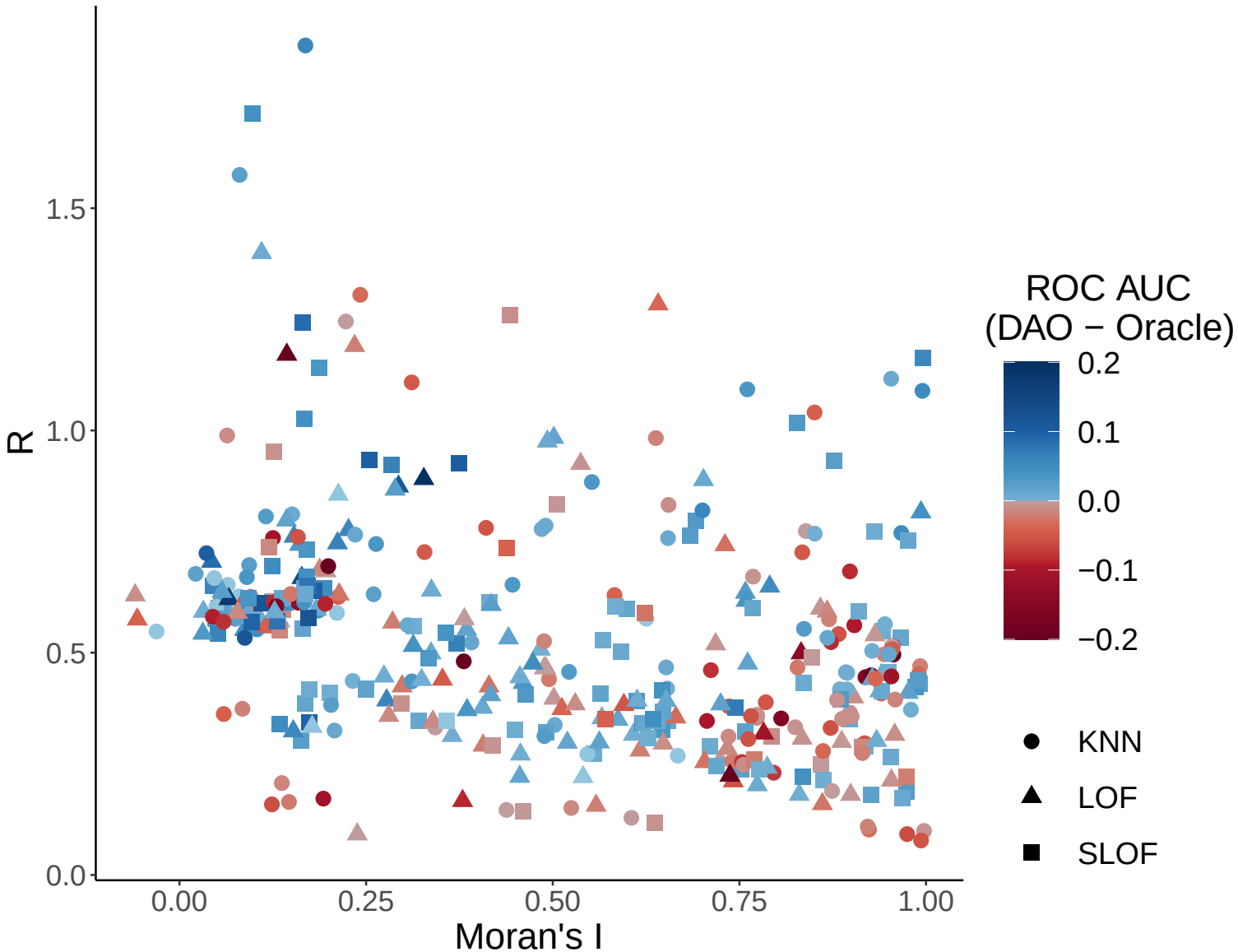


Plots show measures of variation in log-LID values.

- y-axis: Global dispersion (Mean Absolute Difference R).
- x-axis: Maximum local autocorrelation (Moran's I) over neighborhoods of size $k \in [5, 100]$.
- Greatest uniformity of LID: lower-right corner.
- Color indicates ROC-AUC difference (DAO's minus competitor's).

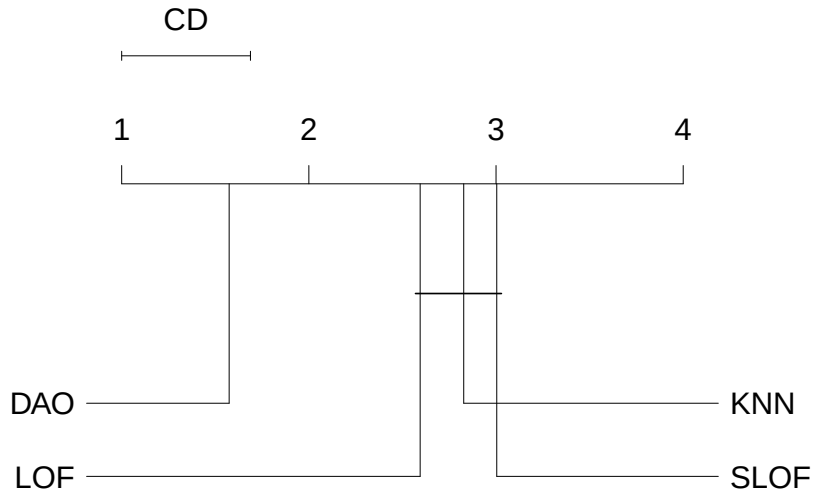
Experimental Analysis

DAO versus Best Competitor



Experimental Analysis

Critical Difference Diagram of Average Ranks



- DAO (with MLE) against baseline competitors.
- Average ranks are shown (between 1 and 4).
- 393 real datasets, with significance level $\alpha = 1e{-}16$.
- No statistically-significant differences among the competitors.

Conclusion

Variation of LID across data does exist!

- Conventional approaches are susceptible to these variations.
- Some of the applications of LID theory to date:
  - More 'principled' treatment of local outlierness (DAO) ...
  - ...and out-of-distribution detection in ML.
    (Wang et al., SDM 2021)
  - Characterization of susceptibility of a point to adversarial attack.
    (Ma et al., ICLR 2018; Amsaleg et al., IEEE TIFS 2021)
  - LID-based regularization for encouraging alignment in GANs.
    (Barua et al., arXiv 2019; Li et al., IJCAI 2019)
  - Explanability in ML through improved perturbation.
    (Jia et al., KDD 2019)
- All involve interpretation of the local interplay between distances, probability, and features (ID).

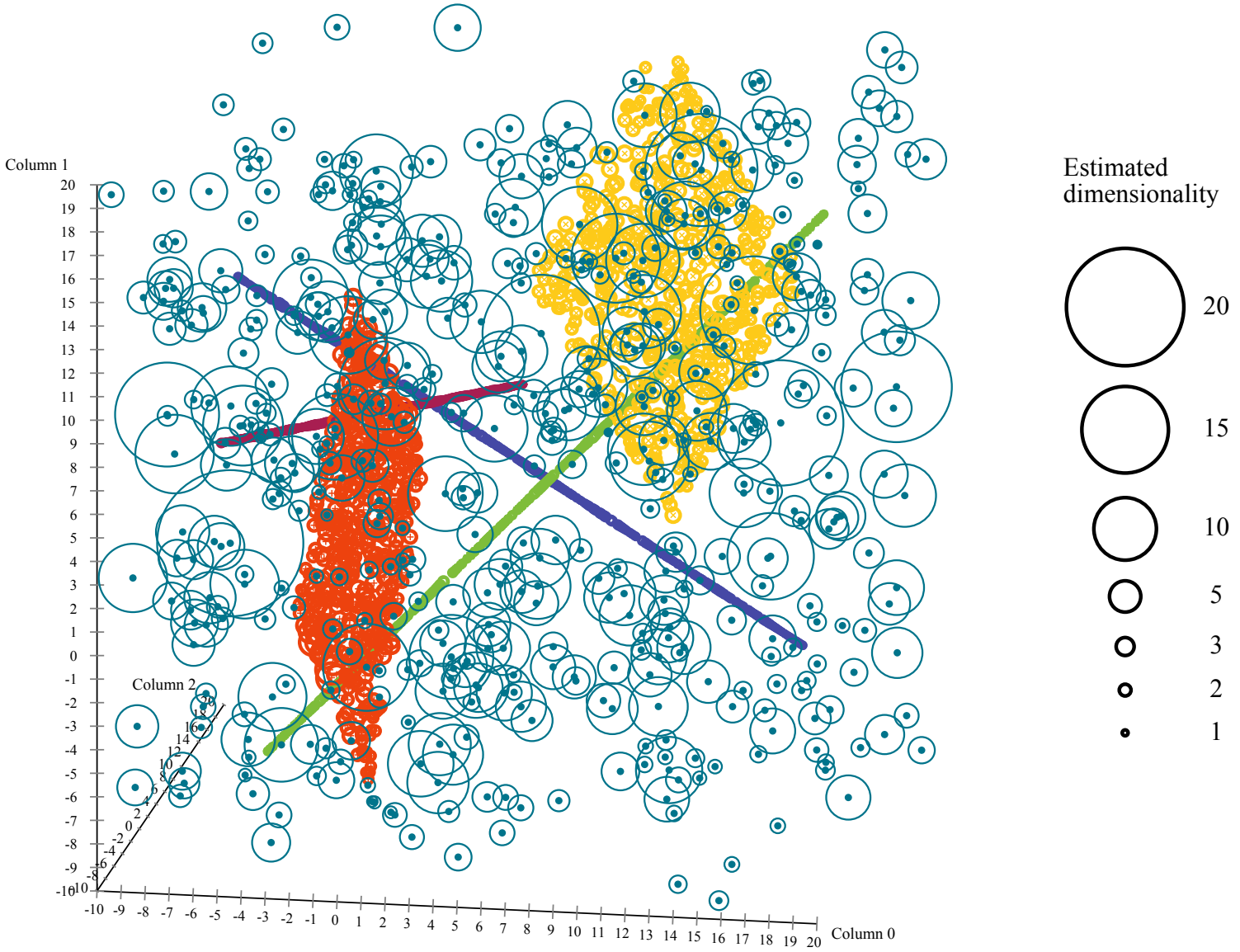
Conclusion

Exploiting LID in multimedia learning applications?

- Representations with better local ID properties?
- Distributional models within localities (neighborhoods)?
- Neighborhood-based alignment of evolving representations?
- LID-based regularization of learning process?
- Characterizing the convergence of learning?
- Locally adaptive thresholding and decision making?
- ...

Success requires effective LID estimation:

- Appropriate resolution (scale).
- Efficiency.
- Robustness.
- Smoothed estimation.


Column 1
Column 2
Column 0
Estimated dimensionality
20
15
10
5
3
2
1

Thank You!