Angle-Based Intrinsic Dimensionality

Erich Schubert

LS8 Künstliche Intelligenz, Fakultät für Informatik, Technische Universität Dortmund
erich.schubert@tu-dortmund.de

Data Science Colloquium, 2020-11-26

Angle-based Intrinsic Dimensionality

The following results have been published as:

Erik Thordsen and Erich Schubert
"ABID: Angle Based Intrinsic Dimensionality"
In: *Proceedings of the 13th International Conference on Similarity Search and Applications, SISAP*. 2020, pp. 218–232.
DOI: 10.1007/978-3-030-60936-8_17

(and won the "best student paper award", congratulations, Erik Thordsen!)

An extended version is invited to a special issue of Information Systems, Elsevier.

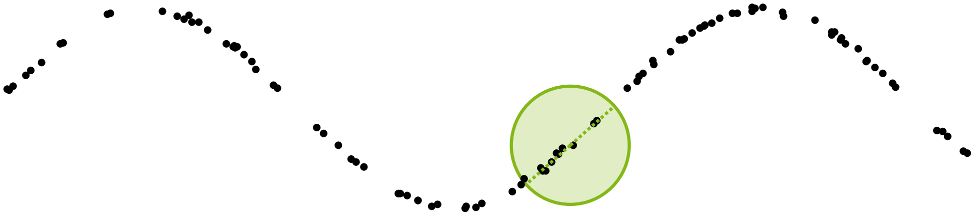
Intrinsic Dimensionality

Data may be *d*-dimensional, but *locally* behave like lower-dimensional data!

- this data set is $\subset \mathbb{R}^2$
- *local* density growth is linear in the radius, i.e., behaves like $\mathbb{R}^1$
- one notion of intrinsic dimensionality is based on the "expansion rate":
  how fast the amount of data growths with increasing distance: expect $|N_\varepsilon| \sim a^d$

Intrinsic Dimensionality

Data may be *d*-dimensional, but *locally* behave like lower-dimensional data!



- this data set is $\subset \mathbb{R}^2$
- *local* density growth is linear in the radius, i.e., behaves like $\mathbb{R}^1$
- one notion of intrinsic dimensionality is based on the “expansion rate”:
  how fast the amount of data growths with increasing distance: expect $|N_\varepsilon| \sim a^d$

Estimating Intrinsic Dimensionality [HKN12; Hou17a; Hou17b]

Intuitively: if we double the radius, a 1-dimensional interval doubles the length,
a 2-dimensional disc has 4 times the area, and a 3-dimensional ball has 8 times the volume.

We estimate the intrinsic volume by the number of points n_r at radius r.

The volume of a m-sphere is proportional to R^m, and hence $\log n_r \propto m \log r$.

$$\frac{\text{Volume}(r_1)}{\text{Volume}(r_2)} = \frac{r_1^m}{r_2^m} = \left(\frac{r_1}{r_2}\right)^m \approx \frac{n_{r_1}}{n_{r_2}}$$

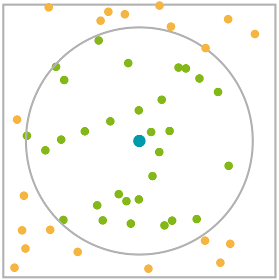
$$m \approx \frac{\log n_{r_1} - \log n_{r_2}}{\log r_1 - \log r_2}$$

For a more robust estimate, we can average this over many radiuses.

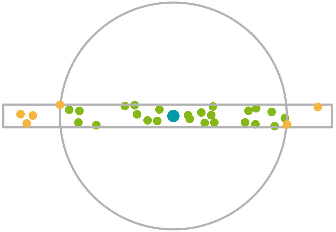
Many different estimators based on this idea exist [HKN12; Ams+15; Ams+18; Ams+19].

Angle-Based Intrinsic Dimensionality Intuition [TS20]

Consider the distribution of angles between neighboring points:



approximately uniform (in 2d)

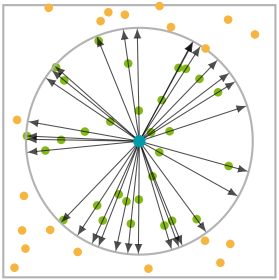


more small and large angles

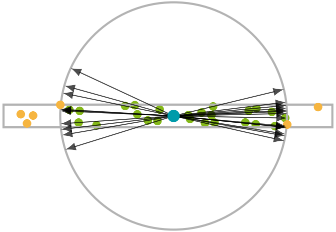
What is the expected angle distribution in a uniform m-dimensional neighborhood?

Angle-Based Intrinsic Dimensionality Intuition [TS20]

Consider the distribution of angles between neighboring points:



approximately uniform (in 2d)



more small and large angles

What is the expected angle distribution in a uniform *m*-dimensional neighborhood?

Angle-Based Intrinsic Dimensionality (ABID) [TS20] /2

The distribution of angles in a $(m-1)$-sphere (or m-ball) is:

$$P(\theta) = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{d-1}{2})} \cdot \sin(\theta)^{d-2}$$

Fortunately, the distribution of Cosines is simpler (and more efficient to compute):

$$P(C) = \tfrac{1}{2} B\left(\tfrac{1+C}{2}; \tfrac{d-1}{2}, \tfrac{d-1}{2}\right)$$

which yields the following estimator of intrinsic dimensionality [TS20]:

$$m \approx 1/\mathrm{Var}(C)$$

where C are the pairwise cosines between any two neighbors.

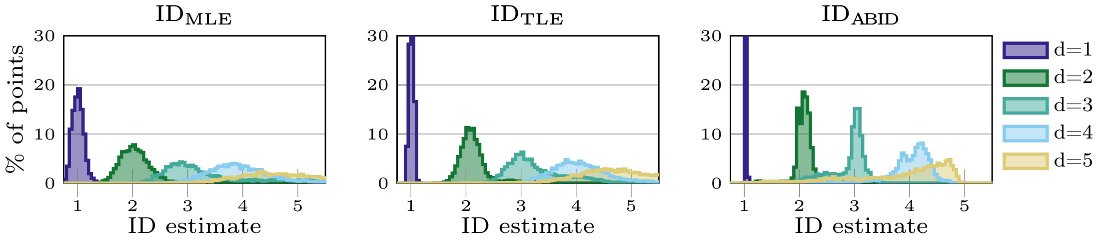
A better estimator can be derived by using the second non-central moment,
and including the self-angles (cosine 1) of each neighbor [TS20]:

$$m \approx n^2 \Big/ \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\langle x_i - c, x_j - c \rangle}{\|x_i - c\| \, \|x_j - c\|} \right)^2$$

Experimental Results

While our focus is a new theoretical approach, we evaluated the method with different data sets.

A single example, manifolds embedded in $\mathbb{R}^5$:



More experiments & details in the paper [TS20]!

Bibliography I

[Ams+15] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. "Estimating Local Intrinsic Dimensionality". In: *SIGKDD*. 2015, pp. 29–38. DOI: 10.1145/2783258.2783405.

[Ams+18] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayashi, and Michael Nett. "Extreme-value-theoretic estimation of local intrinsic dimensionality". In: *Data Min. Knowl. Discov.* 32.6 (2018), pp. 1768–1805. DOI: 10.1007/s10618-018-0578-6.

[Ams+19] Laurent Amsaleg, Oussama Chelly, Michael E. Houle, Ken-ichi Kawarabayashi, Milos Radovanovic, and Weeris Treeratanajaru. "Intrinsic Dimensionality Estimation within Tight Localities". In: *SDM*. 2019, pp. 181–189. DOI: 10.1137/1.9781611975673.21.

[HKN12] Michael E. Houle, Hisashi Kashima, and Michael Nett. "Generalized Expansion Dimension". In: *ICDM Workshops*. 2012, pp. 587–594. DOI: 10.1109/ICDMW.2012.94.

[Hou17a] Michael E. Houle. "Local Intrinsic Dimensionality I: An Extreme-Value-Theoretic Foundation for Similarity Applications". In: *SISAP*. 2017, pp. 64–79. DOI: 10.1007/978-3-319-68474-1_5.

[Hou17b] Michael E. Houle. "Local Intrinsic Dimensionality II: Multivariate Analysis and Distributional Support". In: *SISAP*. 2017, pp. 80–95. DOI: 10.1007/978-3-319-68474-1_6.

[TS20] Erik Thordsen and Erich Schubert. "ABID: Angle Based Intrinsic Dimensionality". In: *Proceedings of the 13th International Conference on Similarity Search and Applications, SISAP*. 2020, pp. 218–232. DOI: 10.1007/978-3-030-60936-8_17.