UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

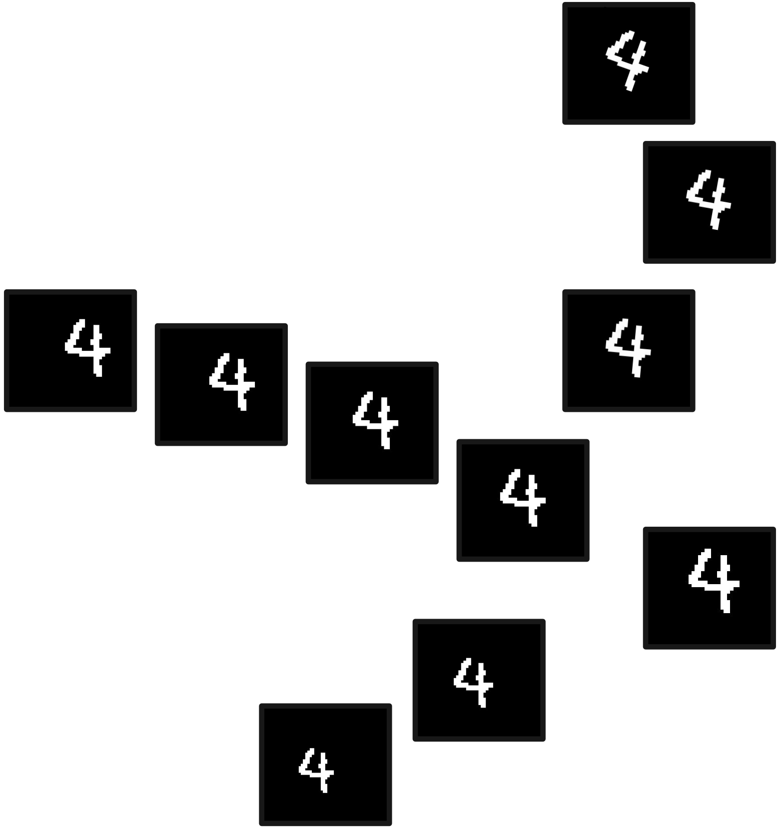
Intrinsic dimension estimation for locally undersampled data

Vittorio Erba[(1)], Marco Gherardi[(1)], Pietro Rotondo[(2)]

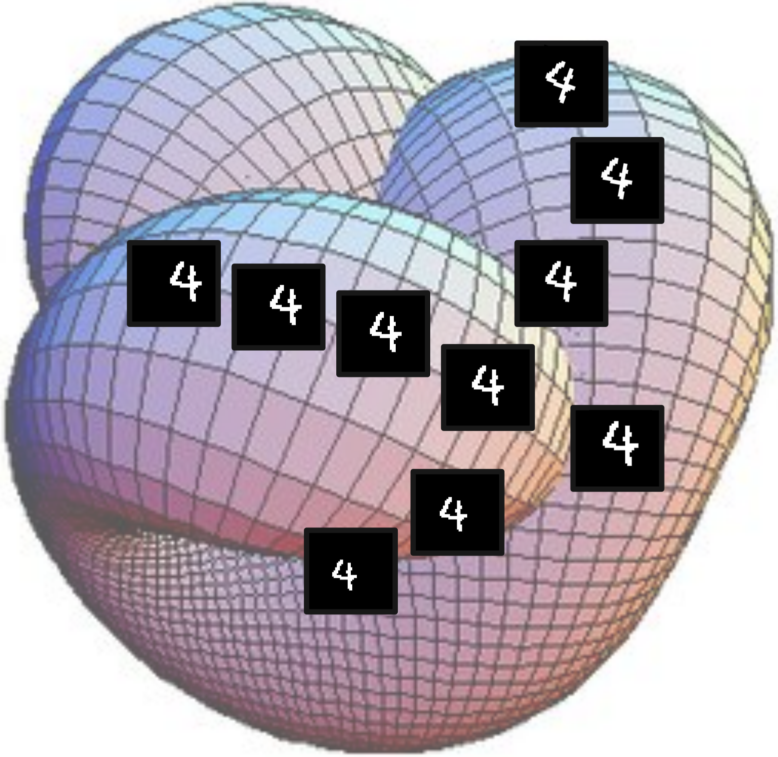
5th Workshop on Complex System, Università degli Studi di Milano, 31 October 2019

(1) Università degli Studi di Milano and INFN, Sezione di Milano
(2) School of Physics and Astronomy and Centre for the Mathematics and Theoretical Physics of Quantum Non-equilibrium Systems, Nottingham

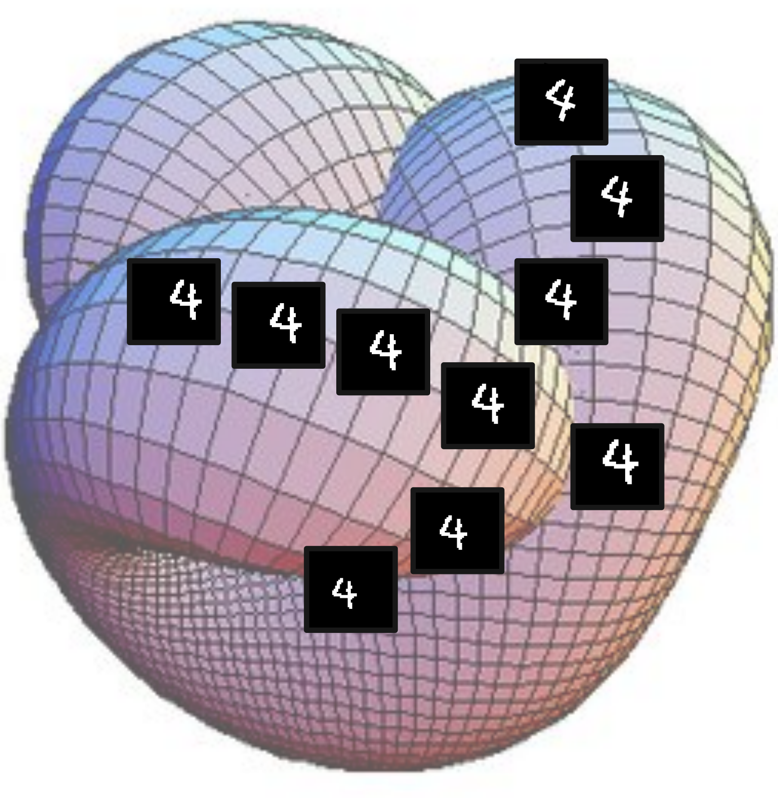
What is the structure underlying complex data?

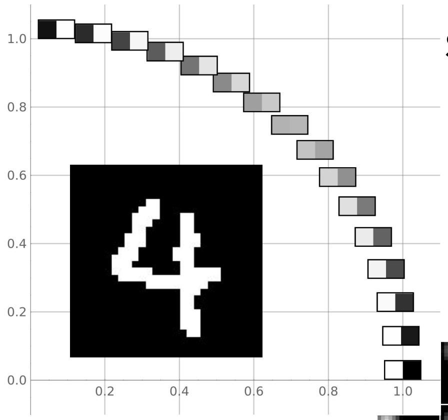


What is the structure underlying complex data?

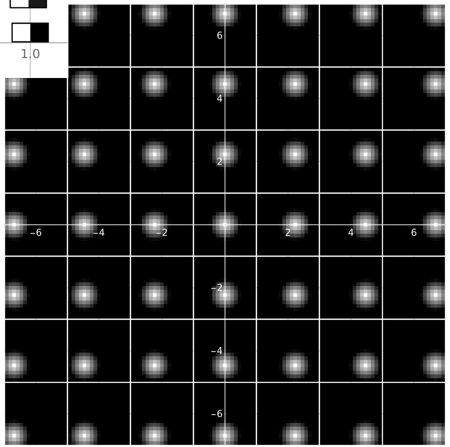


What is the structure underlying complex data?





Simple embeddings for humans are complex for maths

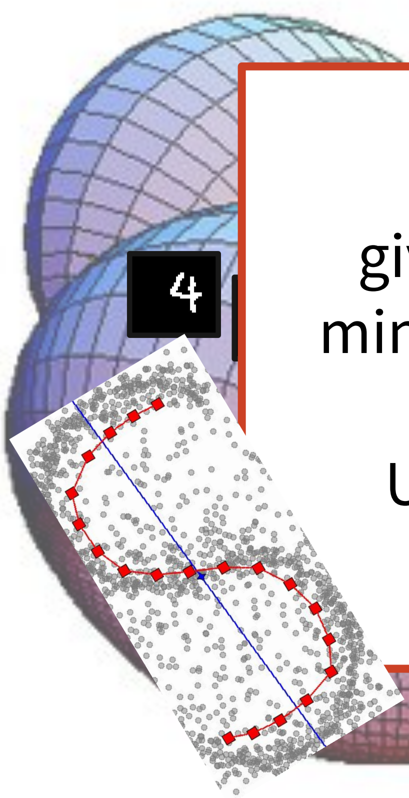


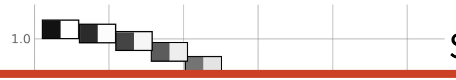
pixel space is huge!

$$\mathbb{R}^{28\times 28}$$

High dimensional embeddings are redundant: **Intrinsic Dimensionality**

What is the structure underlying complex data?



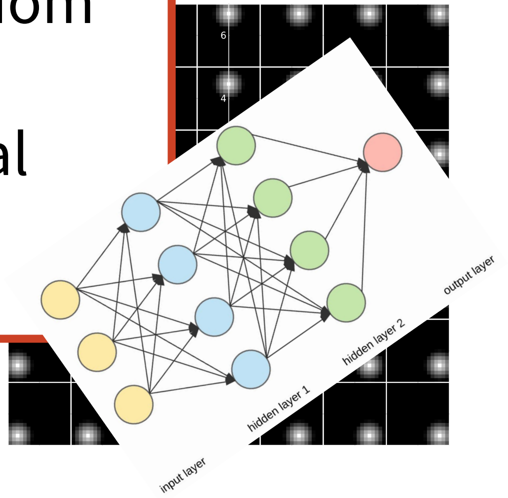


Simple embeddings for
humans are
plex for maths

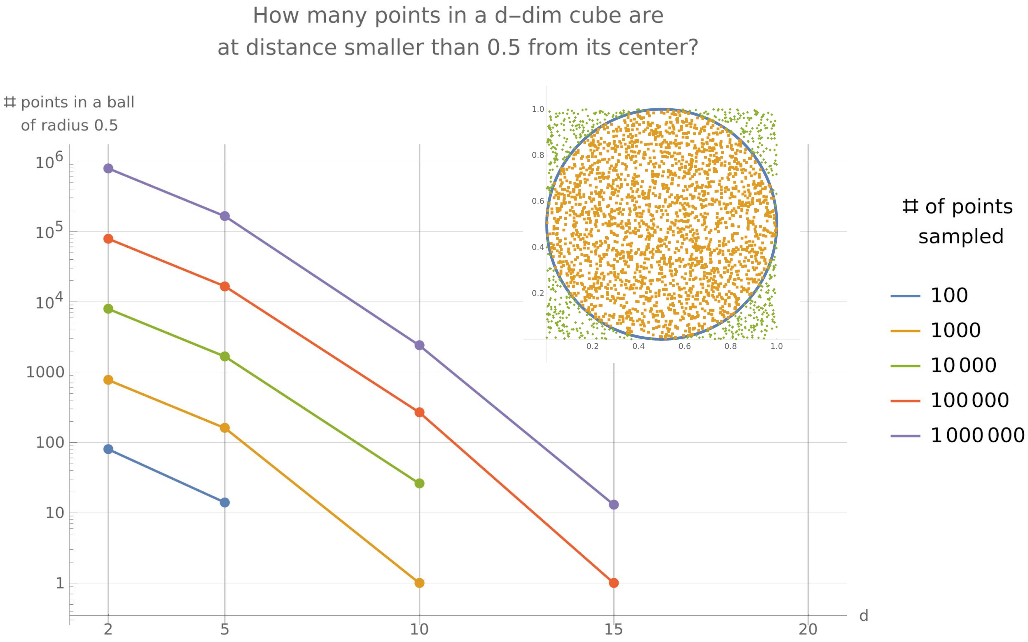
Intrinsic dimension estimation:
given the embedded data, retrieve the minimum number of degrees of freedom

Unsupervised learning, dimensional reduction, feature extraction

embeddings are redundant:
Intrinsic Dimensionality



Curse of dimensionality: exponential undersampling in high dimension

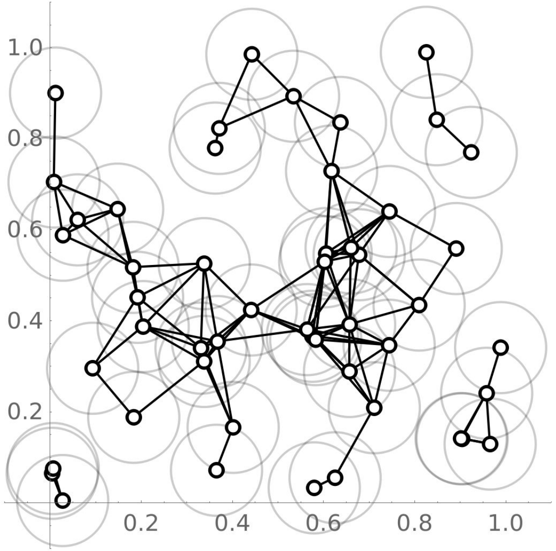


High dimension
d > 6

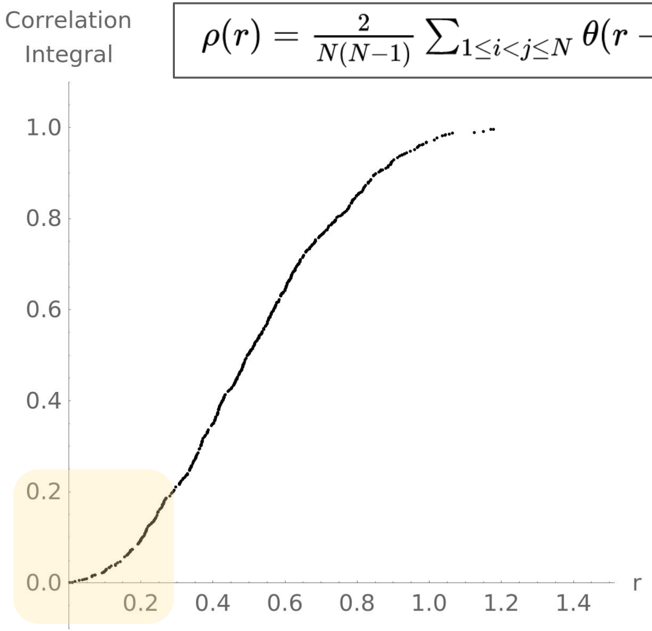
Eckmann, J-P., and David Ruelle. "Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems." *Physica D: Nonlinear Phenomena* 56.2-3 (1992): 185-187.

Geometric estimators: look at the local structure of data

Locally, datasets are linear ⇒ Local number of neighbours scales as r^d

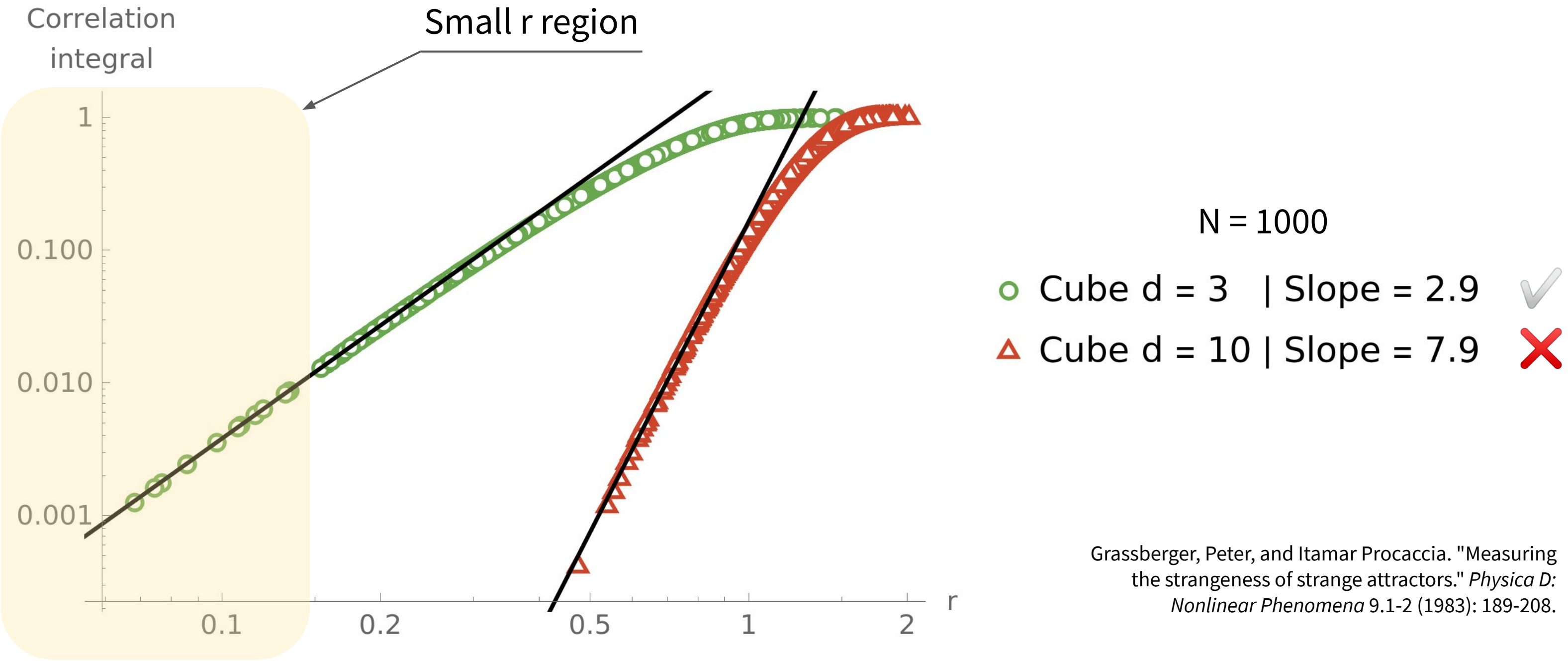


$$\rho(r) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \theta(r - ||\vec{x}_i - \vec{x}_j||)$$



Data undersampling ⇒ ID underestimation

The intrinsic dimension can be extracted by a linear fit on the log-log plot in the region of small r



Grassberger, Peter, and Itamar Procaccia. "Measuring the strangeness of strange attractors." *Physica D: Nonlinear Phenomena* 9.1-2 (1983): 189-208.

A tradeoff between non-linearity and undersampling

	Geometric local estimators	Projective global estimators	???
Non linear	✓	✗	✓
High dimension	✗ (exp d)	✓ (d log d)	✓

Grassberger, Peter, and Itamar Procaccia. "Measuring the strangeness of strange attractors." *Physica D: Nonlinear Phenomena* 9.1-2 (1983): 189-208.

Ceruti, Claudio, et al. "Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration." *Pattern recognition* 47.8 (2014): 2569-2581.

Hein, Matthias, and Jean-Yves Audibert. "Intrinsic dimensionality estimation of submanifolds in R d." *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005.

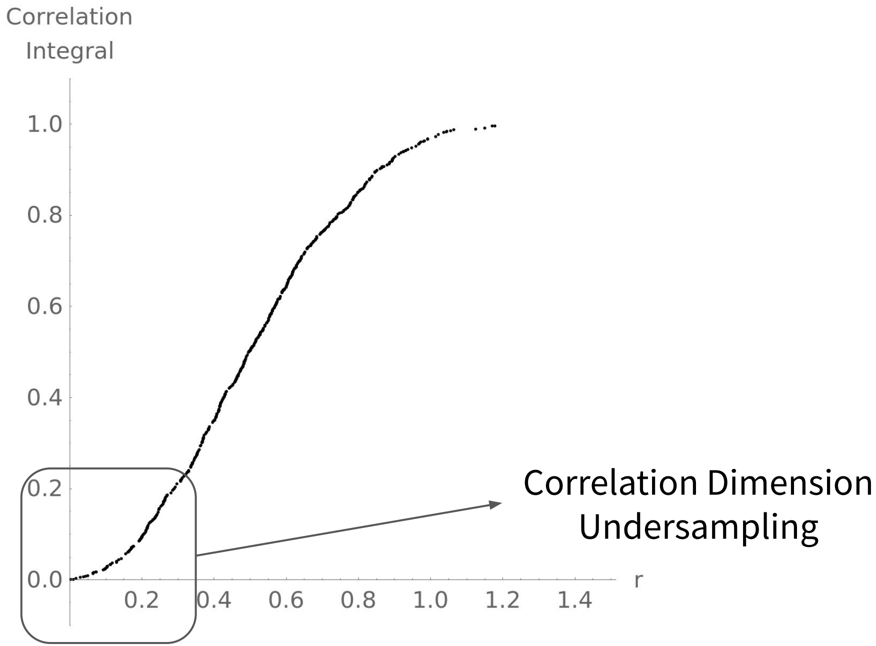
Correlation dimension
DANCO
Hein
...

Principal Component Analysis

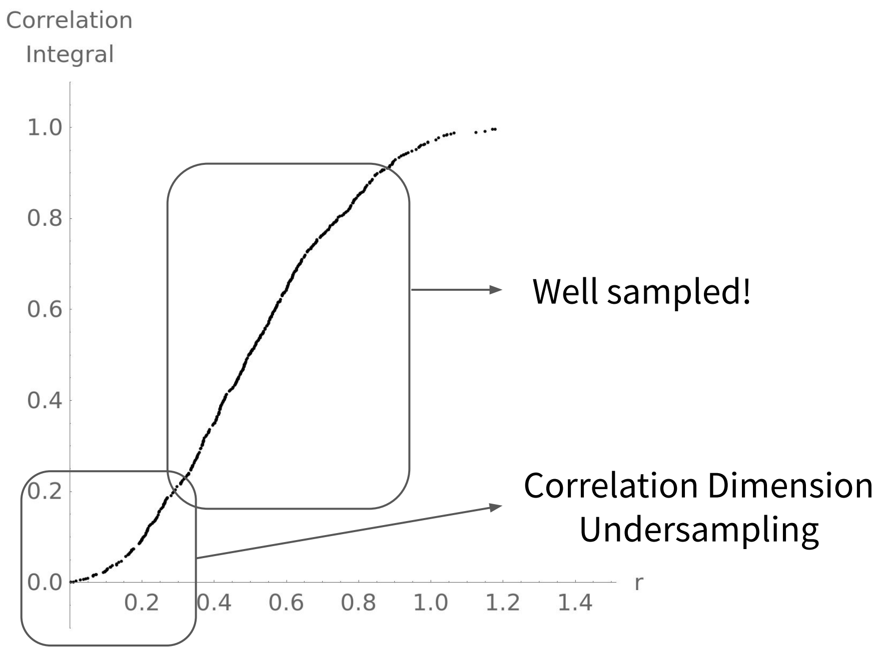
Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901): 559-572.

Anna V. Little, Jason Lee, Yoon-Mo Jung, and Mauro Maggioni. Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD. In 2009 IEEE/SP 15th Workshop on Statistical Signal Processing. IEEE, 2009.

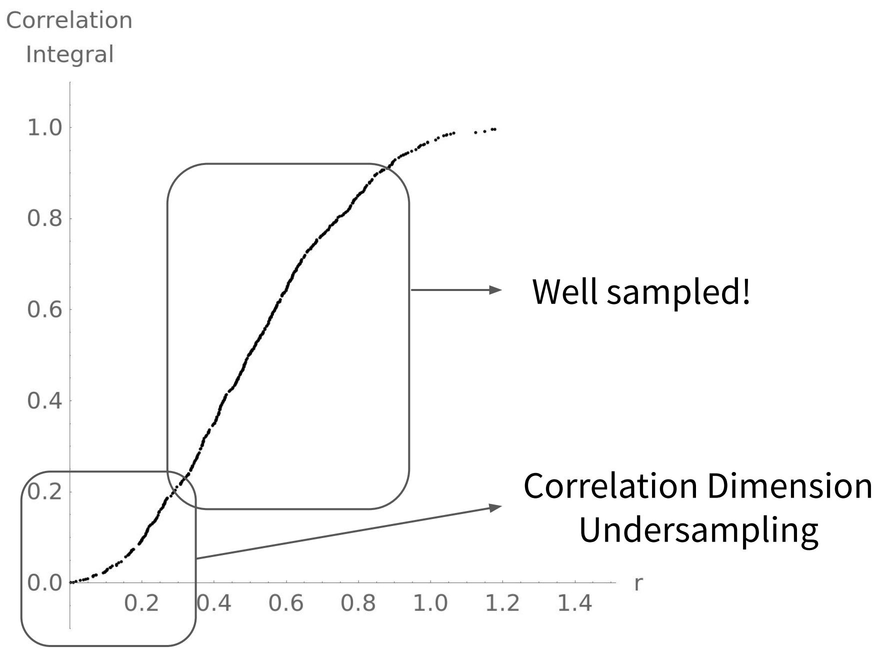
The Full Correlation Integral Estimator (FCI)



The Full Correlation Integral Estimator (FCI)



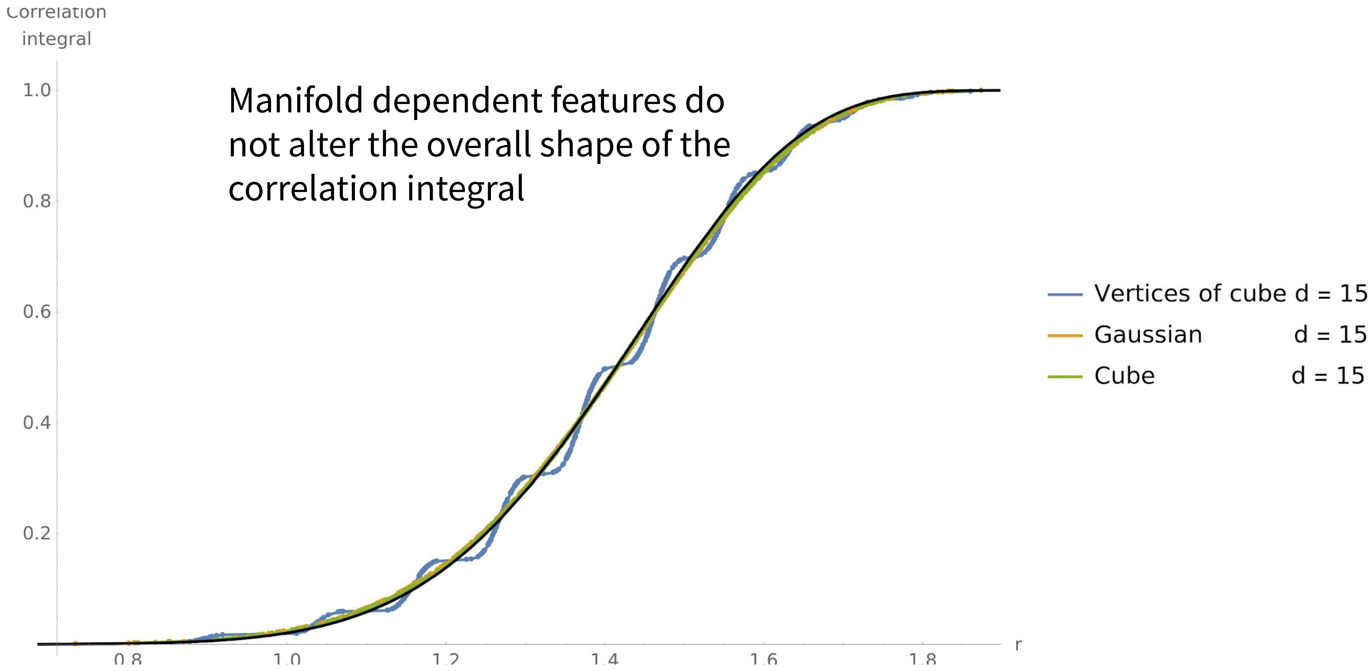
The Full Correlation Integral Estimator (FCI)



Linear manifolds
Isotropic sampling measure
Linear embeddings
$\Rightarrow$

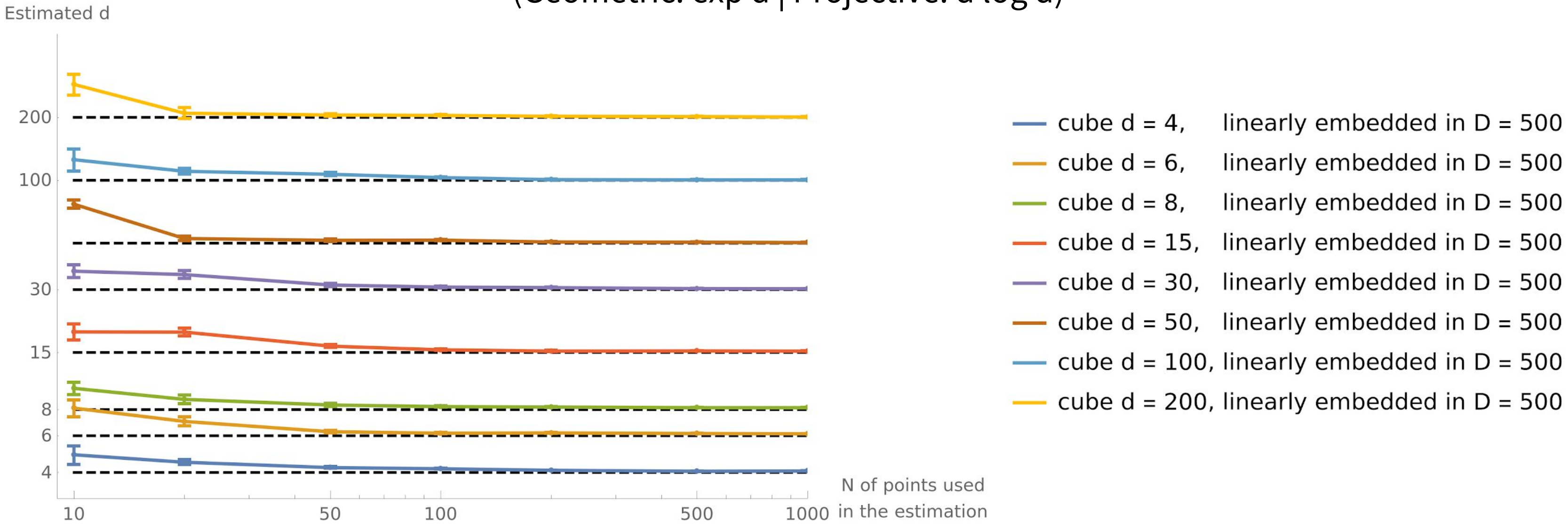
$$\rho(r;d) = \frac{1}{2} + \frac{\Omega_{d-1}}{\Omega_d}(r^2-2)\,{}_2F_1\left(\begin{matrix}\frac{1}{2}, 1-\frac{d}{2} \\ \frac{3}{2}\end{matrix}\middle|(r^2-2)^2\right)$$

The FCI estimator is robust to non idealities



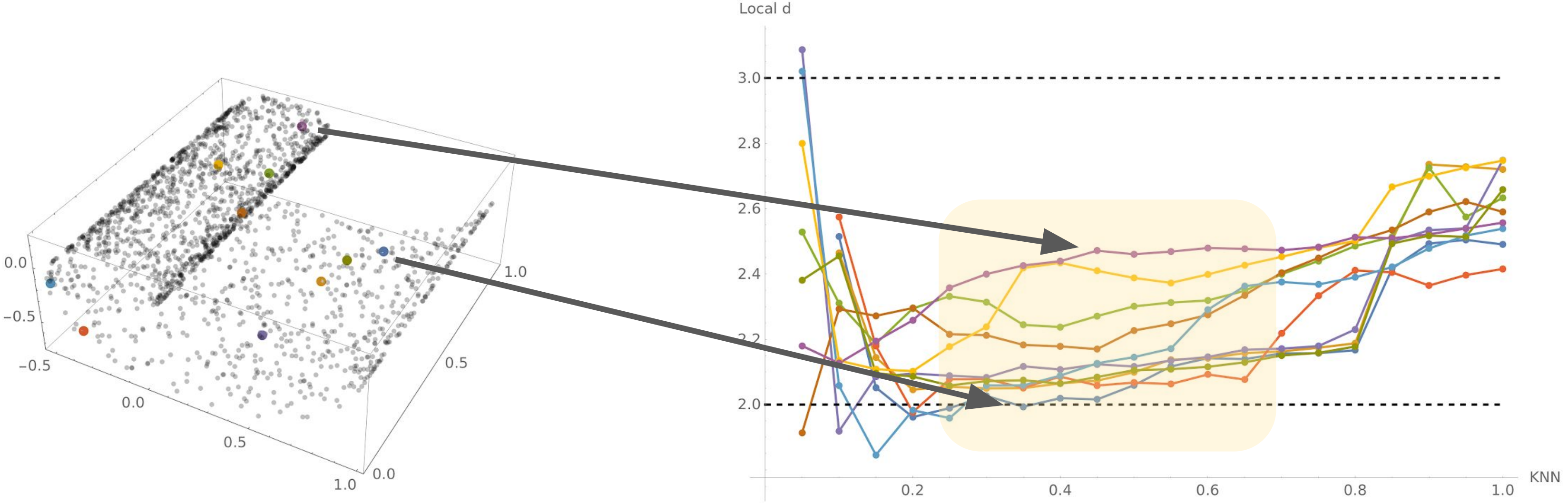
The FCI estimator is robust to undersampling

Able to estimate in the extreme
undersampled regime N < d
(Geometric: exp d | Projective: d log d)



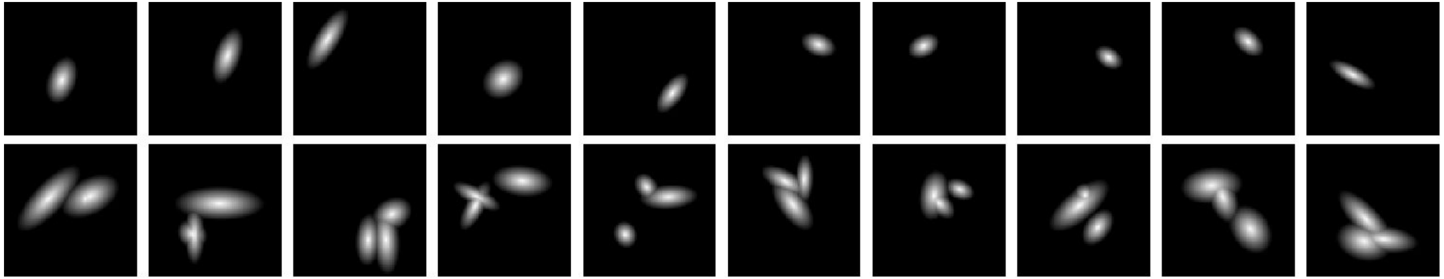
A multiscale generalization of the FCI

Thanks to robustness + extreme undersampling



Use the “most persistent” minimum as the estimator of the Intrinsic Dimension

The multiscale FCI estimator can deal with complex bitmap images

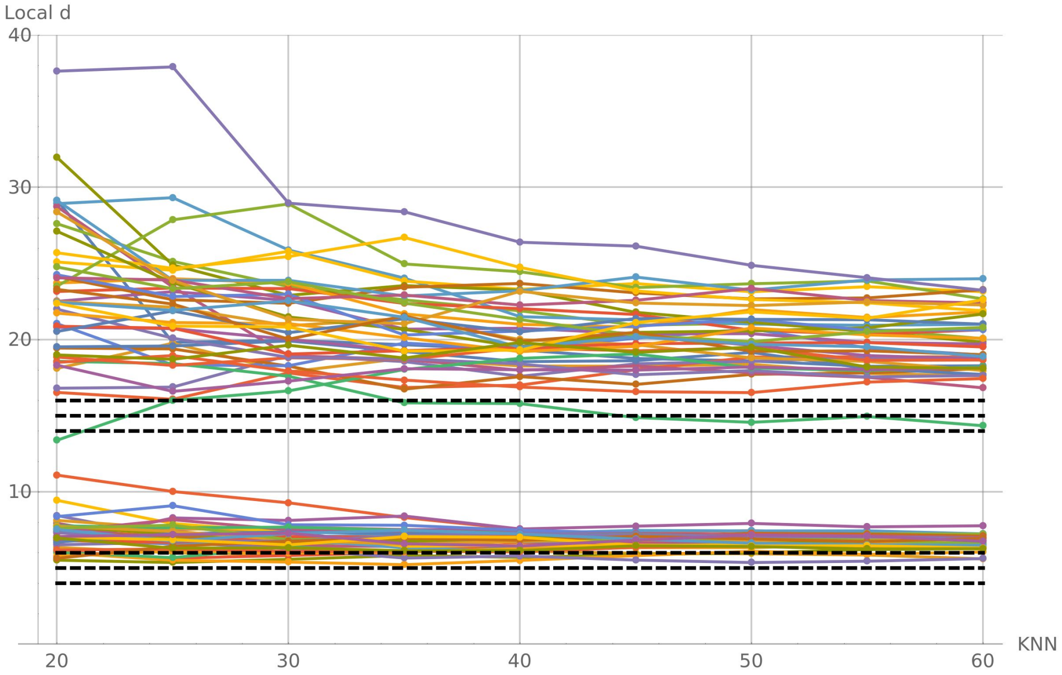


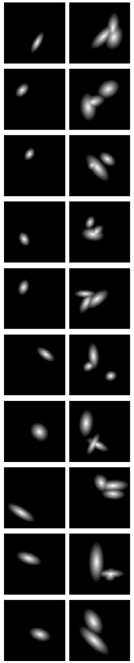
d=5
$D=81^2$

d=15
$D=81^2$

5 degrees of freedom per blob:
translation x, translation y, eccentricity,
scale, tilt

The multiscale FCI estimator can deal with complex bitmap images





The multiscale FCI is more versatile than other estimators

Estimator	$\mathcal{SR}_{2,3}$	$\mathcal{H}_{20,50} \cup \mathcal{H}_{30,50}$	$\mathcal{C}_{6,12}$	$\mathcal{B}_{5,81^2}$	$\mathcal{B}_{15,81^2}$
CorrDim [8]	1.98	12.53	5.93	5	13.5
Takens [10]	1.97	12.01	5.77	N.A.	N.A.
Hein et al. [13]	2	13	6	N.A.	N.A.
PCA	3	20 & 30	12	40	40
mPCA [24]	3	20 & 30	[9,12]	[2,10]	[6,31]
Multiscale FCI	2	20 & 30	6	5	15
Non linear	✓	✗	✓	✓	✓
High dimension	✗	✓	✗	✗	✓
Multidimensional	✗	✓	✗	✗	✗

Hybrid between local geometric methods and projective global methods

Easy multiscale generalization

Performant in a wide variety of situations

Thank you for your attention!

Learn more:

V.E., Marco Gherardi, Pietro Rotondo, "Intrinsic dimension estimation for locally undersampled data"
https://arxiv.org/abs/1906.07670 (accepted in Scientific Reports)