READ MORE
ON arχiv

Intrinsic dimension estimation for locally undersampled data

Vittorio Erba (Università degli Studi di Milano, INFN)
vittorio.erba@unimi.it
In collaboration with: Marco Gherardi, Pietro Rotondo

Problem: given a manifold $\mathcal{M}$ of dimension d embedded in $\mathbb{R}^D$, and a random sample X of N points from $\mathcal{M} \subset \mathbb{R}^D$ extracted with some probability distribution μ, can we recover the **Intrinsic Dimension (ID)** d? In other terms: which is the minimum number of parameters d that describe completely the dataset X?

Why should you care?

In the era of Big Data and Machine Learning, datasets are becoming increasingly redundant and intractable. Extracting the fundamental information from a dataset allows, for example, for aimed dimensional reduction and better comprehension of the structure and symmetries of the data.

State of the art

Current algorithms estimate effectively the ID of low dimensional manifolds ($d \lesssim 10$, see [1]) and of linearly embedded manifolds (using PCA, see [2]). In both cases, a large number of points is required for an effective estimation ($N \sim \exp d$ in the first case, $N \sim d \log d$ in the second case). Highly curved manifolds, multidimensional manifolds and high contrast images are considered challenging datasets for state-of-the-art estimators [3].

Smoothness $\implies$ local approximations

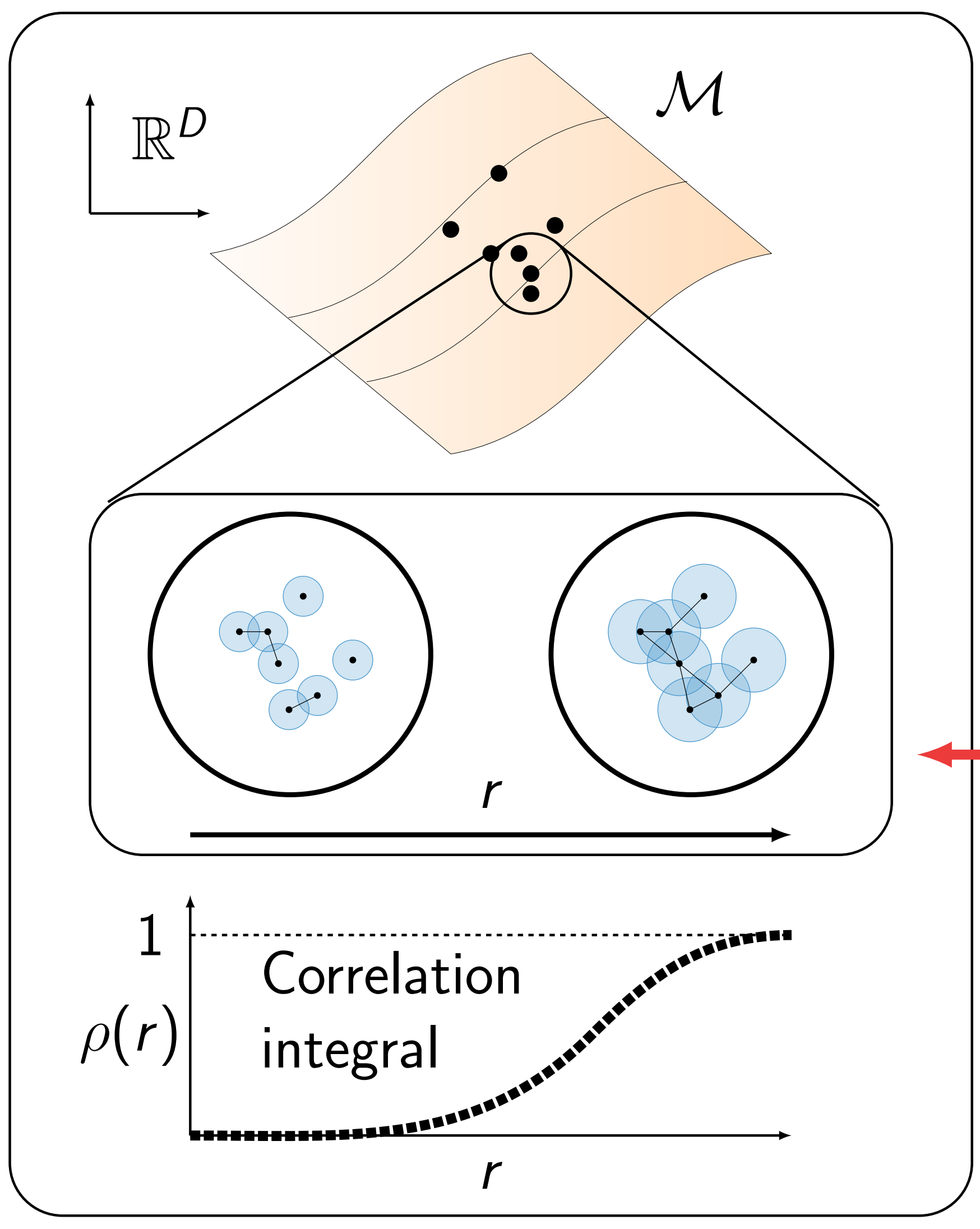
- manifold $\to$ tangent hyperplane
- embedding $\to$ differential (linear map)
- probability measure $\to$ uniform measure

$\implies$ study algorithms for the ID estimation of uniformly sampled, linearly embedded hyperplanes, and apply it locally on the dataset.

High ID $\implies$ local sparse sampling

To sample a manifold densely enough to be able to exploit local approximations, one needs a number of datapoints $N \sim \exp(d)$.

$\implies$ study algorithms that are effective in the undersampled regime.



Full Correlation Integral (FCI) estimator

Study the scaling of the fraction of pairs of datapoints at distance less then r (the **correlation integral**).

$$\rho(r) = \frac{2}{N(N-1)} \sum_{i<j} \theta(r - ||x_i - x_j||_{\mathbb{R}^D})$$

Its average value can be analytically computed for the uniformly sampled, linearly embedded hypersphere $\mathcal{S}^d$, giving an explicit function $\overline{\rho}(r; d)$ that depends parametrically only on d.

FCI estimator

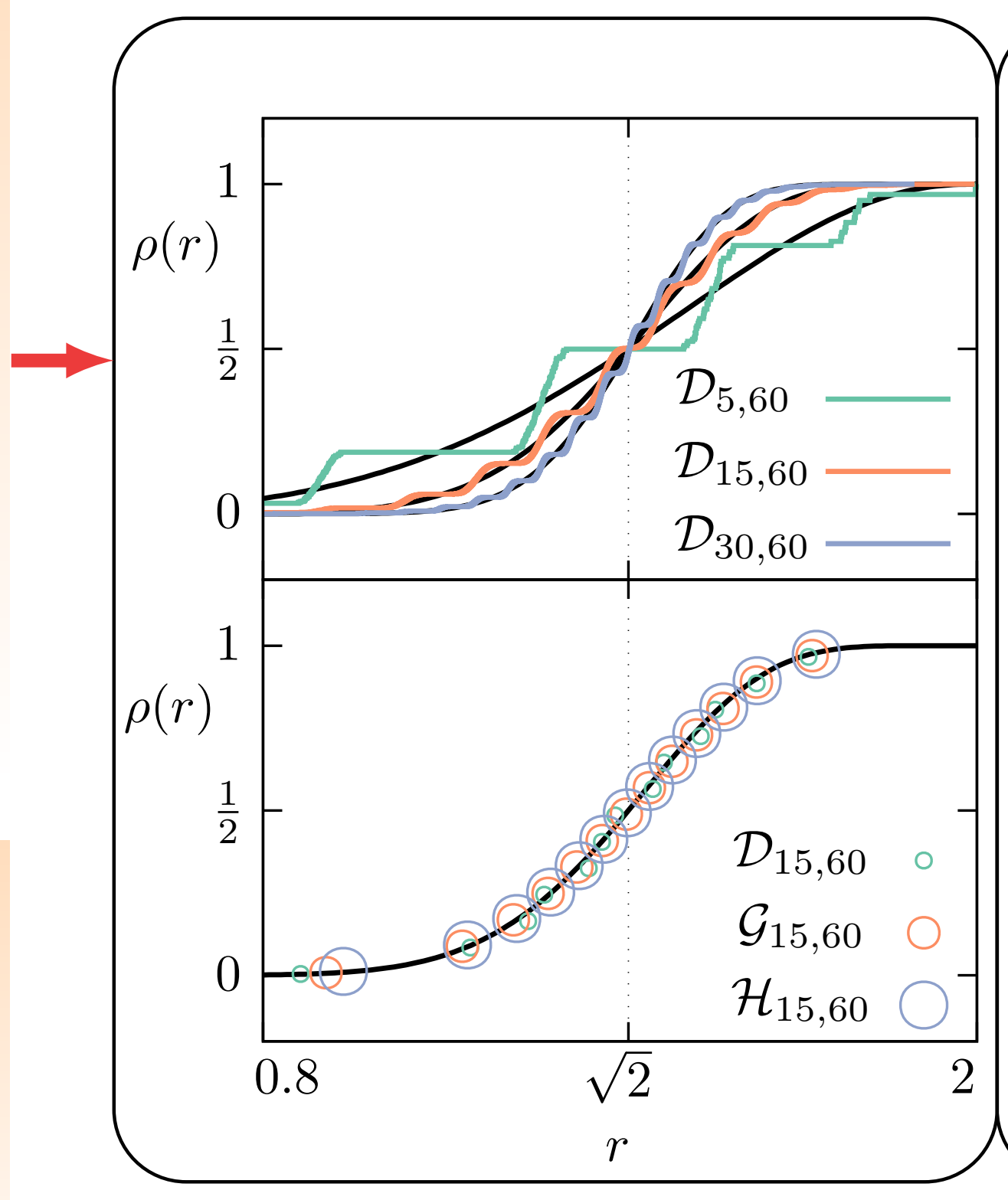
1. Center datapoints on their baricenter $N^{-1} \sum_i x_i$;
2. Normalize the centered datapoints;
3. Compute the empirical correlation integral $\rho(r)$;
4. Fit $\rho(r)$ with $\overline{\rho}(r; d)$ and find d.

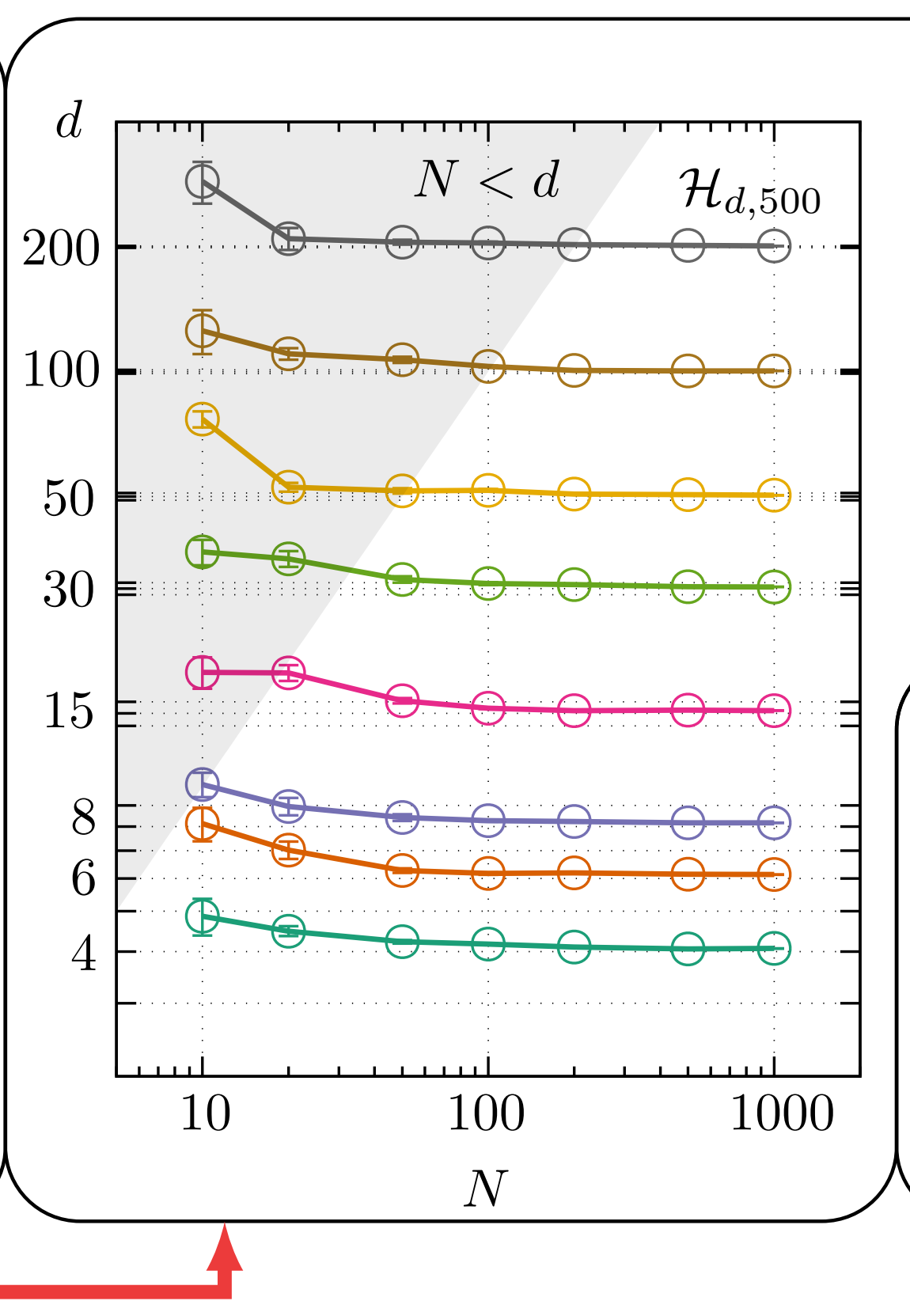
The estimator is **exact** for hyperplanes sampled with rotationally invariant measure, linearly embedded. It's an extension of the Correlation Dimension introduces by Grassberger and Procaccia [4].

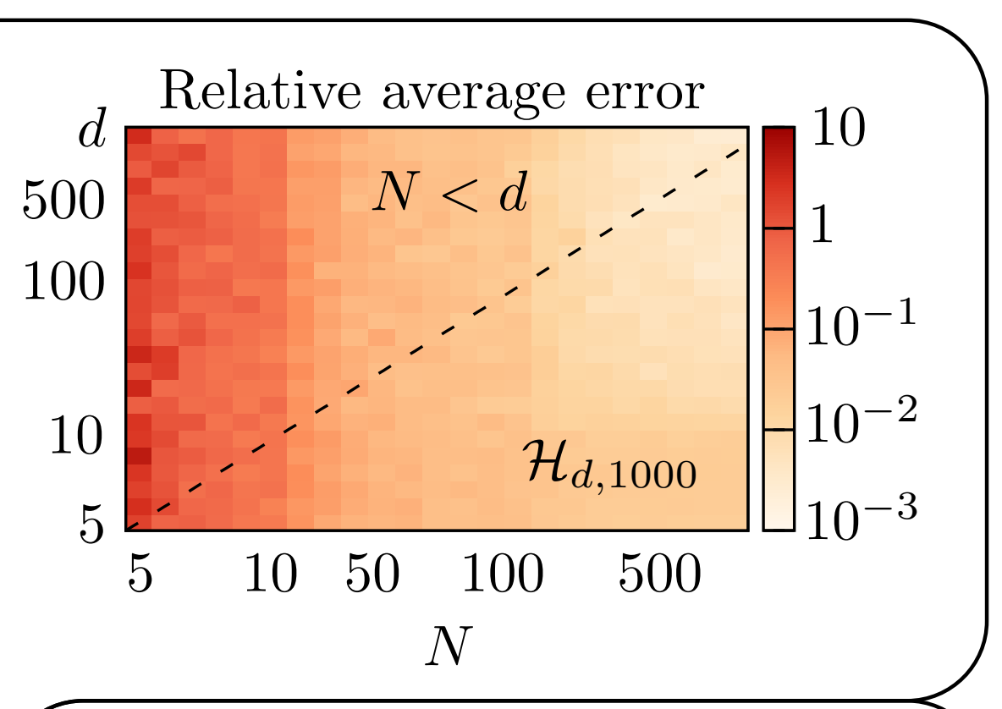
The FCI estimator is effective even for non isotropically sampled datasets that show empirical correlation integrals with non-trivial manifold-dependent features.

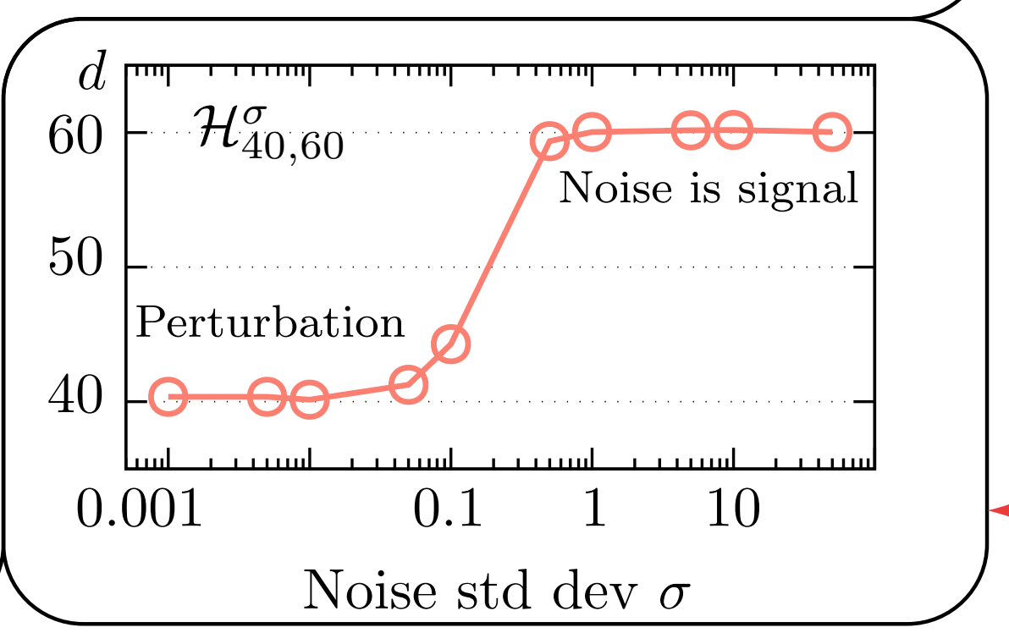
$\mathcal{D}_{d,D}$ Uniformly sampled points from $\{0,1\}^d$
$\mathcal{G}_{d,D}$ Gaussian (variance=1) points from $\mathbb{R}^d$
$\mathcal{H}_{d,D}$ Uniformly sampled points from $[0,1]^d$
* all linearly embedded in $\mathbb{R}^D$.

The FCI estimator works well in extremely high dimensions (vs state-of-the-art estimators that fail for ID$\gtrsim$ 20) and in the extremely undersampled regime $N < d$.









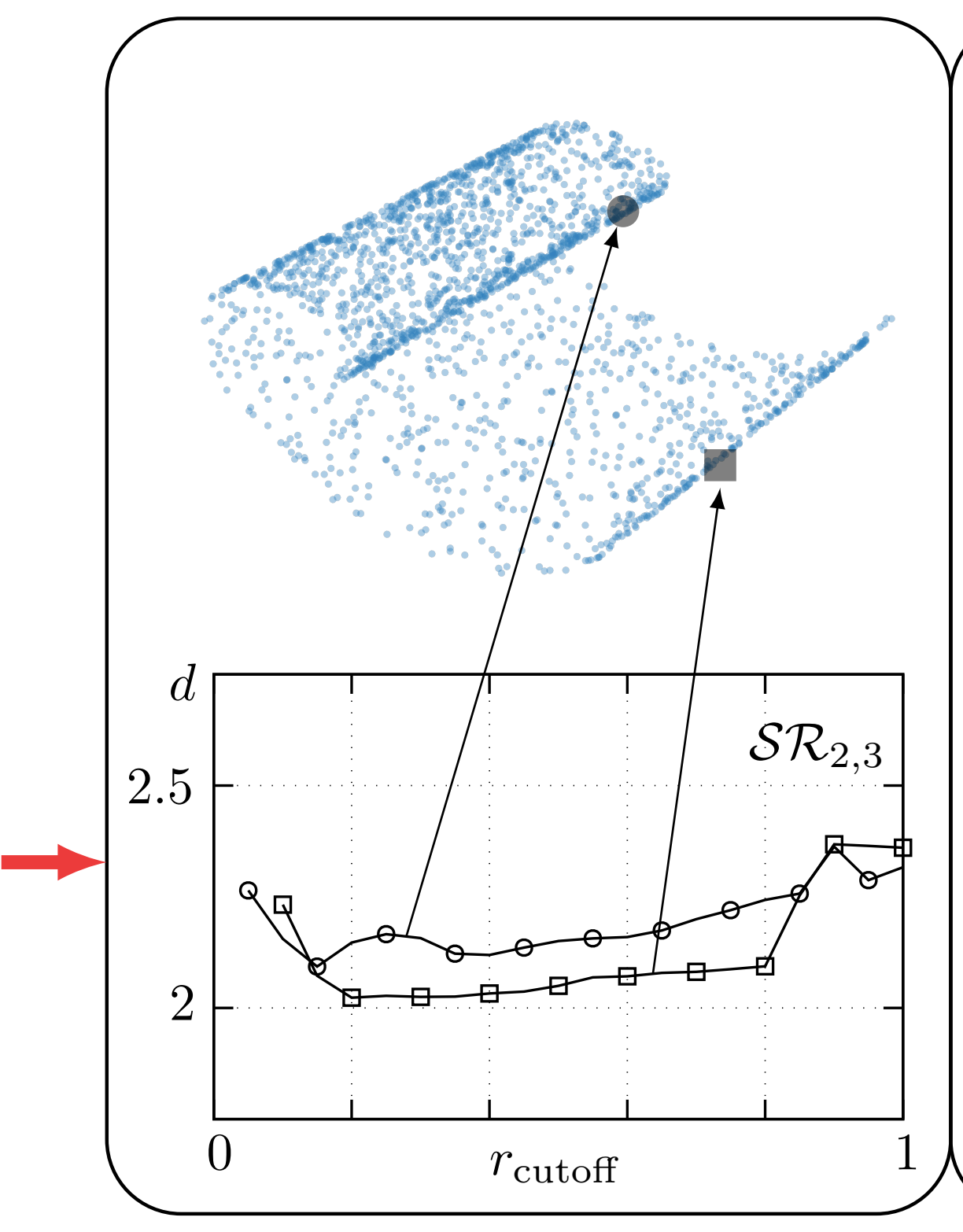
The FCI estimator is robust in the presence of noise, showing a sharp transition between a regime in which the noise is a perturbation and a phase in which the noise covers the signal.

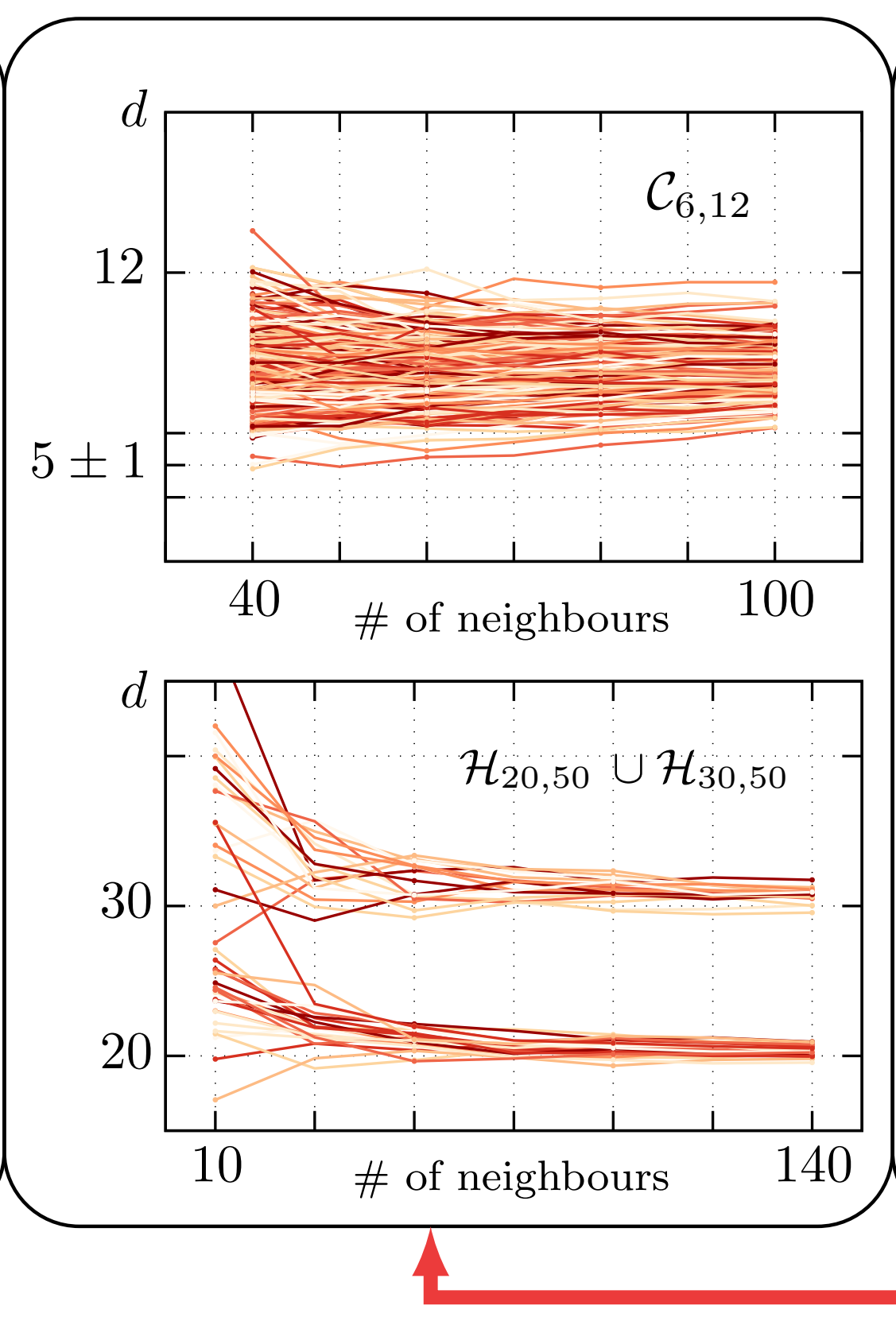
$\mathcal{H}^{\sigma}_{d,D}$: Uniformly sampled points from $[0,1]^d$, linearly embedded in $\mathbb{R}^D$, with D-dimensional Gaussian noise of variance σ^2

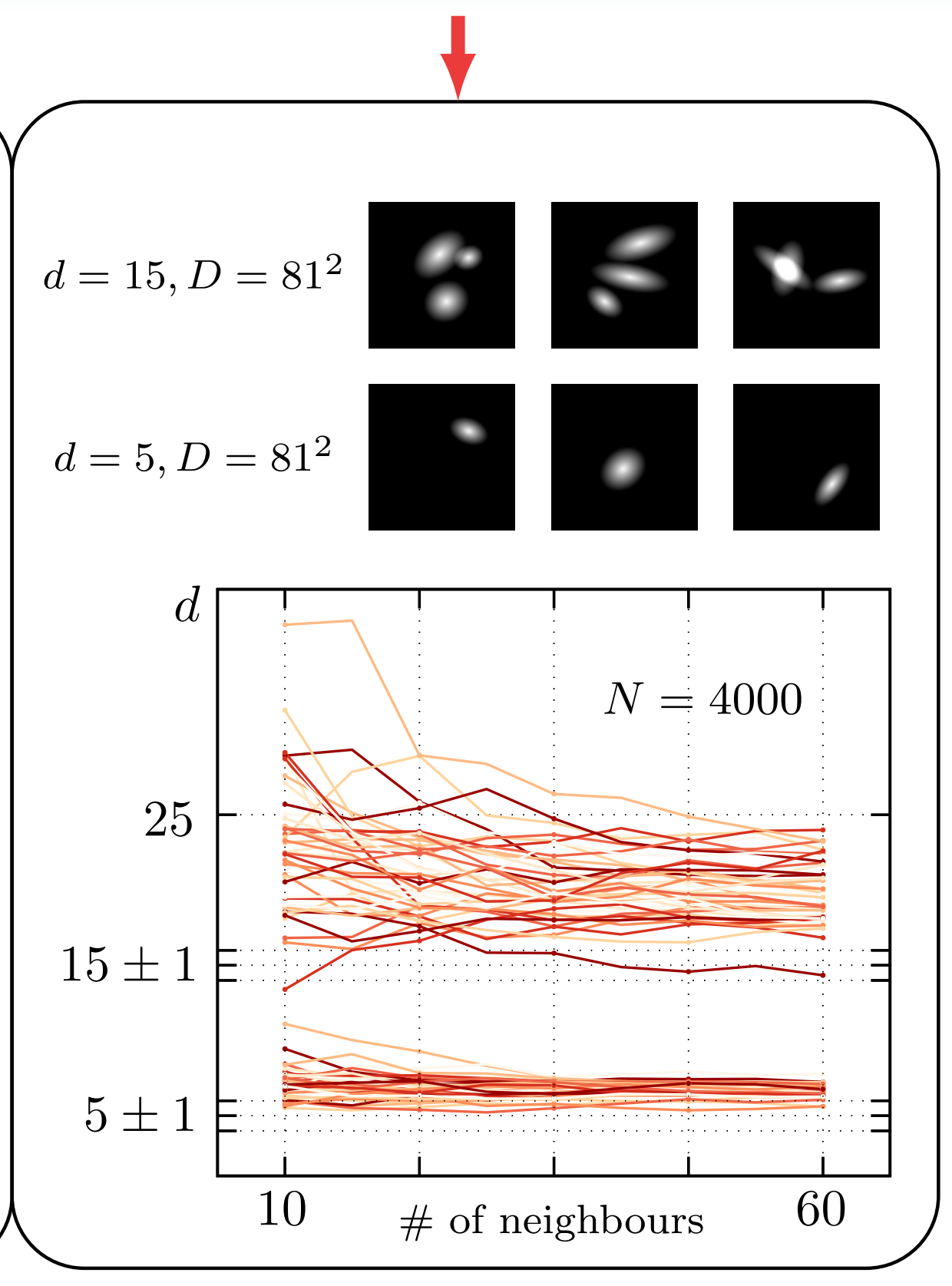
The FCI estimator can be **localized** by applying it on small patches of the dataset, such as neighbourhoods of points at max-distance r_{cutoff} or by selecting the first k nearest neighbours. This gives a local estimate of the ID as a function of r_{cutoff} or k.

Example of local ID analysis for the Swiss Roll dataset. The ID is overestimated where the curvature is high, suggesting that the real ID is given by the minimum plateaus of the curves.

The local FCI analysis allows to treat complex datasets such as bitmap images. We generated images of *blobs* with $d = 5$ (2 translation, 1 rotation, 1 scale parameter and 1 eccentricity) by transforming a default circular blob. In both cases of 1 blob and 3 blob, we are able to estimate the correct ID.







The local FCI analysis succeeds to identify the correct ID of datasets that are considered challenges in the literature: $\mathcal{C}_{6,12}$ is an extremely curved manifold, and the union $\mathcal{H}_{20,50} \cup \mathcal{H}_{30,50}$ displays multidimensionality.

$\mathcal{C}_{d,D}$: Uniformly sampled points from $[0, 2\pi]^d$, embedded with the map

$\phi(x1 \dots x_d) =$
$(x_2 \cos(x_1), x_2 \sin(x_1) \dots x_1 \cos(x_d), x_1 \sin(x_d))$

[1] Ceruti, Claudio, et al. "Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration." Pattern recognition 47.8 (2014): 2569-2581.
[2] Little, Anna V., Yoon-Mo Jung, and Mauro Maggioni. "Multiscale estimation of intrinsic dimensionality of data sets." 2009 AAAI Fall Symposium Series. 2009.
[3] Bengio, Yoshua, Martin Monperrus, and Hugo Larochelle. "Nonlocal estimation of manifold structure." Neural Computation 18.10 (2006): 2509-2528.
[4] Grassberger, Peter, and Itamar Procaccia. "Measuring the strangeness of strange attractors." Physica D: Nonlinear Phenomena 9.1-2 (1983): 189-208.