Astronomy and Computing 53 (2025) 100989

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full Length Paper

Intrinsic dimensionality estimation for the galaxy's distribution structure analysis

A. Chilingarian

Yerevan Physics Institute, Armenia

ARTICLE INFO

Keywords:
Multivariate data analysis
Intrinsic dimensionality estimation

ABSTRACT

The proposed local intrinsic dimensionality method (TIDIM algorithm) demonstrates significant potential in detecting specific filament-like and globular cluster-like structures. It provides a non-parametric, reproducible, and resolution-flexible framework for identifying complex structures within three-dimensional distributions. This approach offers a complementary perspective to traditional statistical tools by focusing on local features such as voids, filaments, and globular clusters. Additionally, it allows for the reconstruction of discrete tracers using the same flexible framework. Its capacity to localize anisotropies and apply TIDIM on a spatial grid makes it particularly useful for comparisons with simulations. The Sobol grid-based TIDIM analysis complements galaxy-based assessments by enabling the detection of structures in sparsely populated regions.

1. Introduction

Dimensionality estimation methods are crucial in data analysis, helping researchers discover the underlying structures in high-dimensional datasets. Early approaches, such as Principal Component Analysis (PCA, Pearson, 1901), mainly focus on linear dimensionality reduction by projecting data onto axes that maximize variance. The intrinsic dimensionality is estimated by counting the number of components needed to explain a certain proportion (usually 95 %) of the variance.

$$d = \textit{smallest integer such that sum}(\lambda i) \;/\; \textit{total variance} \geq \textit{threshold}$$

where λ_i are the eigenvalues of the covariance matrix.

PCA provides a global estimate; it is sensitive to scaling and cannot detect nonlinear structures. Therefore, PCA often fails to identify local manifold structures present in real-world datasets. In response to these limitations, Grassberger and Procaccia (1983) introduced the correlation dimension technique, rooted in fractal geometry, as a method for estimating local dimensionality. It is defined via the correlation integral:

$$C(r) = (2 / N(N-1)) \text{ number of point pairs with } x_i - x_j < r$$

By fitting log(C(r)) vs. log(r), the slope in the linear region gives the intrinsic dimension.

Braams (1974) proposed an algorithm that assesses dimensionality locally by examining the relationship between the radius of neighborhoods and the number of points contained within them, enabling the accurate mapping of nonlinear manifolds. Algorithms were based on the observation that if a set of points is distributed homogeneously in an I-dimensional subspace, then the number of neighbors K within a radius R from a given origin satisfies the scaling relation:

$$K \approx C * R$$

Taking the logarithm of both sides, this becomes:

$$\log(K) \approx \log(C) + I * \log(R)$$

This suggests that the slope of a linear fit to log(K) vs. log(R) provides an estimate of the intrinsic dimension I.

In 1989, A. Chilingarian revised and implemented this methodology to enhance the estimation of local dimensionality in empirical datasets (Chilingarian, 1992; 1989). His revised version eliminated the assumption of strict homogeneity, allowing the method to be applied to real-world data, such as high-energy cosmic rays or multiparticle production on colliders. The TIDIM algorithm was tested and optimized for numerical stability and was successfully utilized in high-dimensional physics applications, including feature selection.

It assumes that the number of neighboring points within radius R of a point O in a d-dimensional subspace scales as:

$$K(R) \approx C * R^d \rightarrow \log(K) \approx \log(C) + d * \log(R)$$

The slope d is determined via least-squares fitting of log(K) versus log (R) using k-nearest neighbors. The global dimension is obtained by

E-mail address: chili@aragats.am.

https://doi.org/10.1016/j.ascom.2025.100989
Received 21 May 2025; Accepted 2 July 2025
Available online 4 July 2025

taking the distribution median of the local dimensionality estimates at points uniformly covering the initial manifold. Additionally, the TIDIM algorithm was used to map the data points in the initial feature space to a lower-dimensional space, where most of the population is grouped due to conservation laws and other factors.

Beyond dimension estimation, a range of methods aim to reveal the actual coordinates of lower-dimensional manifolds embedded in high-dimensional space. These include:

- Isomap (Tenenbaum et al., 2000)**:** Combines geodesic distances with classical MDS to recover global manifold structure. Suitable for smoothly curved manifolds.
- Locally Linear Embedding (LLE, Roweis and Saul, 2000) preserves local neighborhood relationships using linear weights, thereby emphasizing local structure.
- Laplacian Eigenmaps**:** Uses the graph Laplacian to preserve local distances. Effective for discovering low-dimensional embeddings.
- Diffusion Maps (Coifman and Lafon, 2006): Constructs a diffusion process on the data graph and embeds using eigenvectors of the diffusion operator.
- Autoencoders (Hinton and Salakhutdinov 2006)**:** Neural networks that compress input into a low-dimensional bottleneck and reconstruct it, learning manifold coordinates.

Each method has strengths and limitations. Isomap and diffusion maps capture global structure, while LLE and Laplacian Eigenmaps emphasize local geometry. Autoencoders provide flexible, learnable embeddings that scale to complex data and support out-of-sample mapping. However, none of these techniques recover analytic forms of the manifold, and most require dense, well-sampled data.

We examine the TIDIM method through extensive simulation experiments, comparing it directly with PCA and correlation dimensionality methods. The focus is on situations where local structure is essential, and where global methods, such as PCA, fall short.

The final illustration of the TIDIM algorithm involves analyzing galaxy distribution in the nearby Universe ($z < 0.2$) using data from the Sloan Digital Sky Survey. We also test TIDIM on Gaussian and real-world datasets (Digits, Wine, and Breast Cancer). We evaluate TIDIM through extensive experiments, comparing it directly with PCA and assessing its usefulness alongside the above nonlinear manifold learning techniques. The focus is on situations where local structure is crucial and global methods, such as PCA, are insufficient. Each section provides methodological details, comparative analysis, and visual illustrations that demonstrate the advantages of TIDIM for local manifold detection and mapping, particularly in complex and nonlinear contexts.

2. Synthetic Gaussian dataset

We generated 10,000 points in a 10-dimensional space, where 3 dimensions were sampled from N (0, 1) and the remaining 7 dimensions from N(0, 0.01), resulting in an effective intrinsic dimension of 3. Yield global dimensionality values from 2.9803 to 3.0073, see Table 1. PCA identified three significant components, accounting for 95 % of the variance. The corrected correlation dimension, computed on a subset of 1000 points, yielded an estimate of approximately 3.03, see Table 2.

3. Real-world dataset analysis

Table 3 presents a comparison of two local intrinsic dimensionality estimation methods—TIDIM and Correlation dimensionality, and global estimate PCA—applied to three real-world datasets: Digits (64-dimensional data), Wine (13-dimensional data), and Breast Cancer (30-dimensional data). TIDIM estimates are computed using $k = 30$ neighbors, while PCA estimates are based on the number of components required to explain 95 % of the variance

Table 1
Summary Table of global dimensionality estimates($k = 5$ to 500).

k (Neighbors)	Mean Estimated Dimension	Variance
5	3.2496	5.4115
10	2.9363	1.5813
15	2.8988	0.9829
20	2.8931	0.7308
25	2.8971	0.5857
30	2.9024	0.4919
35	2.9083	0.4278
40	2.9144	0.3811
45	2.9203	0.3455
50	2.9261	0.3176
55	2.9313	0.2941
60	2.9361	0.2745
65	2.9405	0.2579
70	2.9448	0.2440
75	2.9488	0.2321
80	2.9526	0.2217
85	2.9561	0.2124
90	2.9595	0.2042
95	2.9626	0.1969
100	2.9656	0.1903
130	2.9803	0.1619
160	2.9916	0.1444
190	3.0003	0.1329
220	3.0073	0.1252
300	3.0204	0.1143
400	3.0304	0.1096
500	3.0364	0.1087

Table 2
Comparisons of dimensionality estimation methods($k = 300$).

Method	Estimated Dimension ± MSD	Comment
TIDIM ($k = 300$)	3.043 ± 0.0002	Local estimator
Correlation Dimension ($k = 300$)	3.030 ± 0.0001	Fractal estimator
PCA (95 % variance)	3	Global linear structure

Table 3
Comparison of dimensionality estimation for the real-world data.

Dataset	TIDIM Mean ± MSD	PCA Dim (95 %)	PCA Total Variance	Correlation Dim ± MSD
Digits (64)	6.81 ± 0.0036	29	95 %	5.81 ± 0.0006
Wine (13)	1.40 ± 0.0014	1	95 %	1.34 ± 0.0003
Breast Cancer (30)	2.45 ± 0.0008	1	95 %	2.01 ± 0.0002

3.1. Particle interaction simulation (3N-Dimensional space)

fParticle collisions are usually described as a distribution in 3N-dimensional impulse space. However, particles born via short-lived resonances, and 3 N impulse space can be shrunk to lower dimensions. The modeling scheme includes production of short-lived resonances, i.e., correlated groups that decay into multiple particles (Chilingarian and Harutunyan, 1989).

Let RRR be the number of resonances per event, each resonance producing 2–3 final particles. These particles are momentum-correlated, e.g., their momenta sum to the resonance momentum. The remaining particles are uncorrelated and drawn from a 3D Gaussian distribution. For each event:

- Generate RRR random resonance 3-momenta.

- For each resonance: Split its 3-momentum into 2–3 correlated fragments with small random deviations.
- $N = 10$ particles
- $R = 3$ resonances per event
- 2–3 particles per resonance
- 10,000 events

Each event will be represented as a vector in R^{30} (10×3), but the effective dimensionality will be ~15–20 due to correlations between born particles.

Using the TIDIM algorithm, we find the mean estimated dimension to be approximately 12.07 with a variance of about 0.36. The TIDIM histogram shows a clear peak around 12, indicating a reduction in dimensionality from the original 30. Additionally, the PCA cumulative variance plot demonstrates that roughly 12 components are sufficient to account for 95 % of the data structure. Thus, PCA (a global linear method) confirms TIDIM results.

3.2. Astrophysical data analysis. Local dimensionality estimation of the SDSS galaxy sample

We use a random sample of 10,000 spectroscopically confirmed galaxies from the Sloan Digital Sky Survey (SDSS, 2020), selected in the redshift range $z = 0.01$ to 0.2. The dataset includes Right Ascension (RA), Declination (Dec), and redshift (z) values for each galaxy. The TIDIM algorithm was used to evaluate the intrinsic dimensionality of the spatial distribution of galaxies. In Fig. 1, we show the distributions of these parameters in the selected data sample.

In Fig. 2, we present the map of local dimensionalities estimated for the 10,000 points across the RA-Dec plot at the galaxy location points. The plot contains several empty areas due to the non-uniformity of the selected galaxy sample shown in Fig. 1.

We use a subsample of SDSS galaxies in the region with Right Ascension between 130° and 230°, and Declination between 10° and 30° Dimensionality is estimated using the median of local slope estimates over the full k-range from 7 to 51. In Fig. 3, we show the histogram of the obtained intrinsic dimensionalities. We pay special attention to the tails of the histogram ($d > 3.5$ and $d < 1$) to gain insight into the special conditions surrounding galaxies with extreme estimates of dimensionality.

Galaxies with high estimated local dimension are highlighted, and the spatial distribution of dimensionality is visualized. We compare the spatial distribution and inferred environment of two extreme classes of galaxies from the SDSS subsample with RA 130–230° and Dec 10–30° Local intrinsic dimensionality was estimated using the median slope across k-nearest neighbor distances ($k = 7$ to 51). The analysis focuses on two categories: 1) galaxies with TIDIM > 3.5 and 2) galaxies with TIDIM < 1.

In Fig. 4, we observe that high-dimensional galaxies ($d > 3.5$) are distributed in underdense regions, consistent with the presence of cosmic voids. Their sparse local neighborhoods yield higher local slope estimates in the TIDIM algorithm. This class includes potential void galaxies, where neighbor sparsity inflates the dimensionality estimate.

Galaxies with TIDIM < 1 shown in Fig. 5 are concentrated in narrow filaments (Tempel et al. 2014) or the cores of superclusters. Their local geometry is highly constrained, resulting in nearly linear arrangements of points. These galaxies may be embedded in known structures such as the Coma or Hercules Superclusters. Their placement within dense environments suggests physical proximity to cosmic filaments or galaxy walls.

Figs. 4–5 illustrate the analysis of the galaxy neighborhood in the Universe; however, for a comprehensive galaxy density analysis of the entire Universe, we need to populate the whole space with test points, regardless of whether galaxies are present at those points or not. For this, we applied the TIDIM estimator to a uniform Sobol grid (Sobol, 1979) of 5000 quasi-random points in the RA–Dec plane (RA: 130°–230°, Dec: 10°–30°), using the distribution of real SDSS galaxies in the same region. Each grid point was analyzed by computing the local slope of log–log neighbor counts. This time, we observe many more events in the right tail of the TIDIM estimates distribution, with 48 instead of 1 having $d > 4$ (void-like structure). In contrast, the number of events with $d < 1$ (filament-like structure) shrinks from 30 to 2, proving that the Universe is mostly empty, and galaxies cover a very small share of the Universe's entire volume.

The distribution of local dimensionality values across the uniform grid shows a clear bias toward higher local dimension in sparsely populated regions. Compared to galaxy-based estimates, the Sobol grid method identifies void-like zones (TIDIM > 4) more frequently.. This reflects the grid's independence from galaxy clustering and confirms the method's utility in tracing large-scale underdense environments. Thus, the Sobol grid-based local dimensionality analysis complements galaxy-based evaluations by enabling the detection of structures in sparsely populated or unsampled regions. It provides a physically interpretable and reproducible tool for mapping structures and identifying voids in the Universe.

3.3. Intrinsic dimensionality-based detection of discrete structures in 3D spaces

Understanding the structural composition of the universe—from filamentary bridges in the cosmic web to compact globular clusters—requires robust analytical tools capable of uncovering such structures without requiring prior labeling or clustering assumptions.

In this section, we present a comparative study of structure detection using the TIDIM method, which estimates local intrinsic dimension from the scaling behavior of point distributions. The TIDIM framework is applied to both simulated globular clusters and cosmic filaments superimposed on a uniform background. We also formulate decision rules to evaluate classification accuracy.

The TIDIM method estimates local dimensionality at each point by analyzing how the neighborhood volume scales with distance. Specifically, for each point:

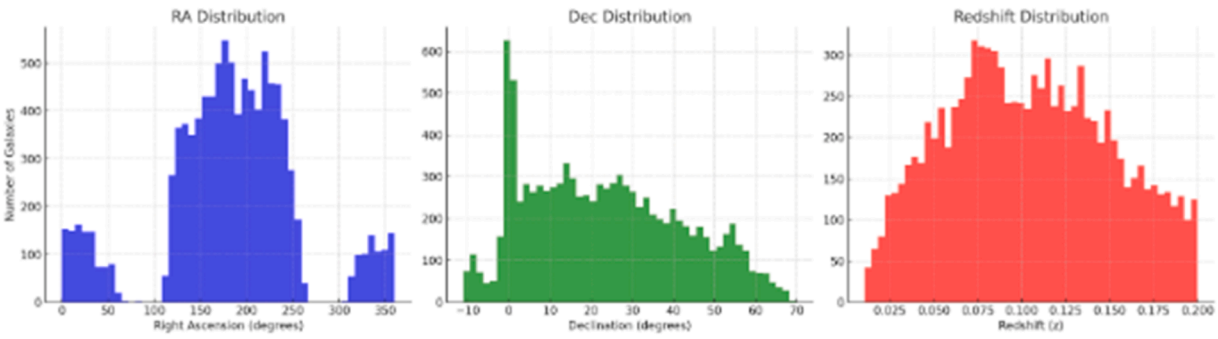

Fig. 1. Local Dimensionality Analysis of SDSS Galaxies in Selected Sky Region.

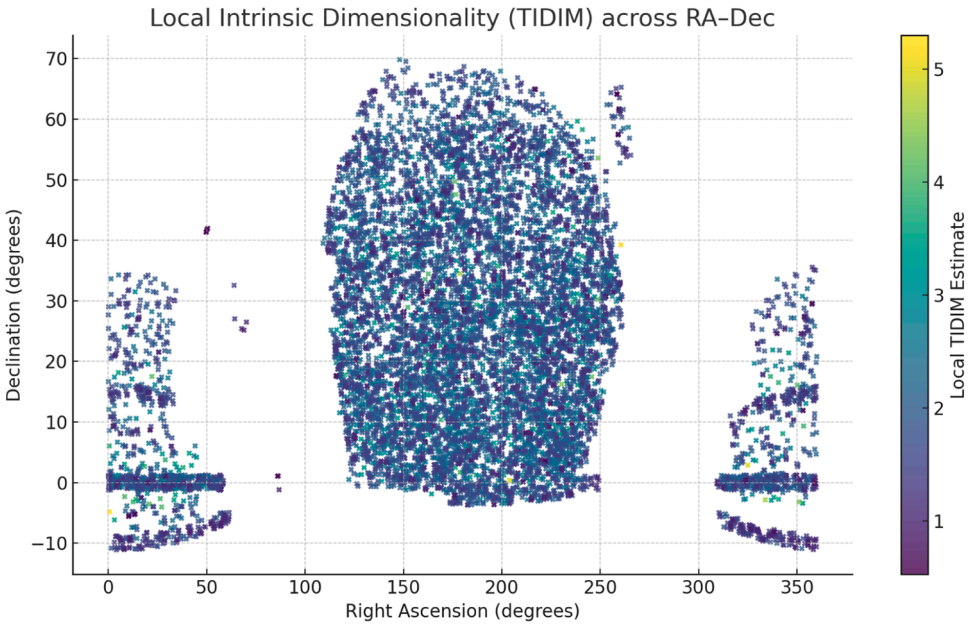

Fig. 2. Two-Dimensional Sky Map Colored by Local TIDIM.

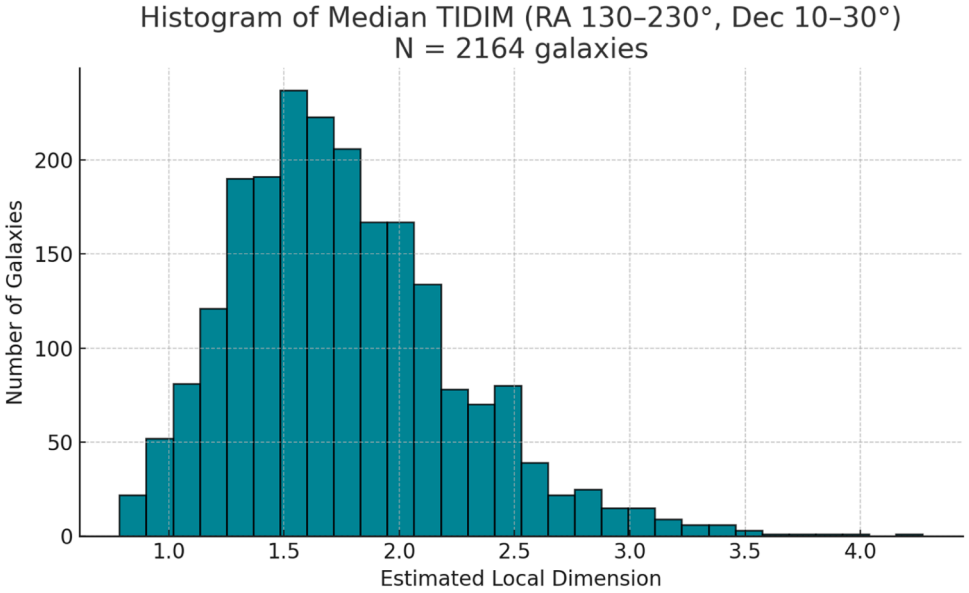

Fig. 3. Local dimensionality histogram of SDSS galaxies in selected sky region.

- A range of nearest-neighbor sizes K ∈ {20, 30, 40, 50, 60} or {100, 400, 600} is used.
- For each K, the log-log slope of distance vs. cumulative neighbor count is computed.
- The inverse slope provides a local ID estimate.
- The final Ti-DIM value is taken as the median across K.

This ensemble approach smooths fluctuations and enhances stability in regions of varying density.

The simulation setup for the globular clusters includes 1000 background points uniformly distributed within the unit cube $[0, 1]^3$, along with 10 globular clusters, each containing a randomly chosen number of points (between 20 and 30) distributed normally around a randomly selected center. The cluster points are tightly concentrated using a Gaussian distribution with a standard deviation of 0.01. We used a simple classification rule: points with TIDIM $<$ 1.5 were considered cluster candidates. Ten independent simulation trials were performed, and for each, the performance of the classification scheme was evaluated. Clusters were matched using spatial overlap criteria and nearest-neighbor linkage, and statistical measures were computed across the 10 randomized simulation trials. Fig. 6 shows the classification of clusters for one of the ten independent simulation trials. We first apply decision rules for each point based on the estimated local dimensionality. Then, detected clusters are matched to the true simulated clusters by requiring that at least 6 points from a detected group overlap with the ground-truth cluster.

Performance Evaluation:

Across 10 independent trials containing 10 globular clusters each:

True positives (mean detected clusters): 9.8 ± 0.42

False negatives (mean missed clusters): 0.1 ± 0.32

False positives (erroneously accepted): 0.4 ± 0.49

The results confirm high sensitivity of TIDIM, with minimal false

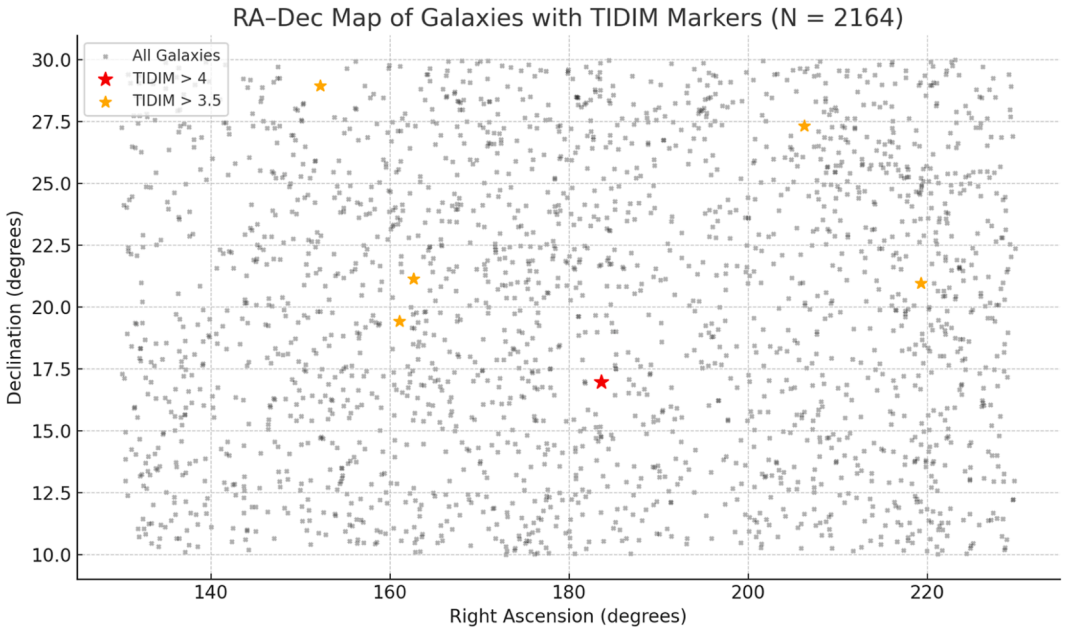

Fig. 4. High TIDIM Galaxies Marked by red and rose Asterisks on the RA–DEC Galaxy Map.

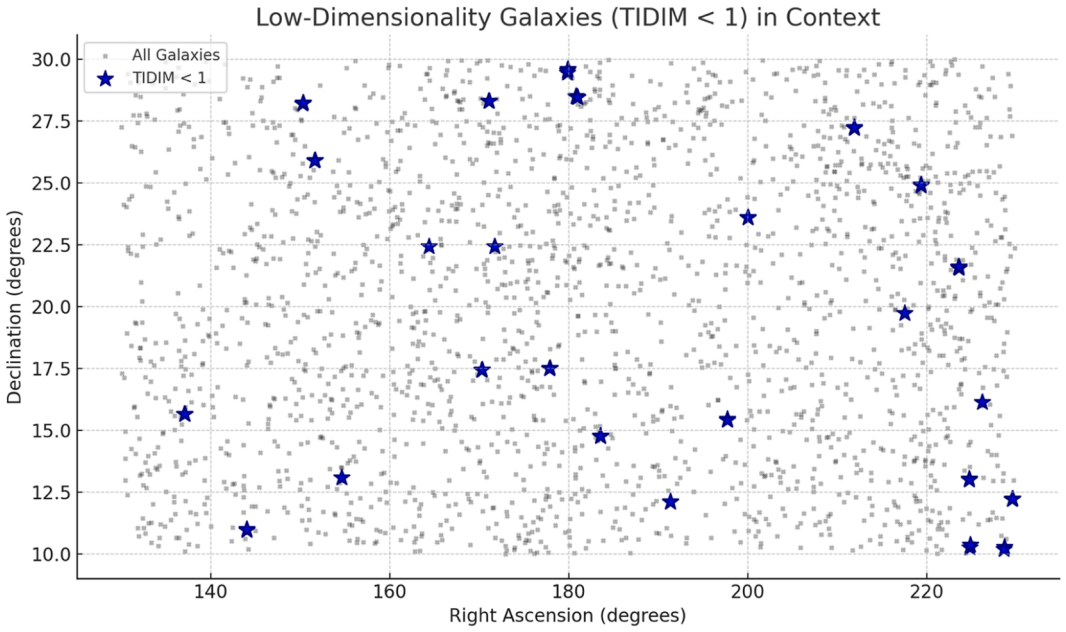

Fig. 5. Low-dimensionality galaxies (TIDIM < 1) are indicated by blue asterisks on the RA–DEC galaxy map.

positives and consistent full-cluster detection across multiple trials. This affirms TIDIM as a reliable tool for identifying compact structures.

In a more complex scenario for detecting linear structures, ten non-intersecting filamentary structures (each with 10–20 points) were embedded in a background of 5000 uniformly distributed points. The filaments were extended linearly and sparsely populated. A point was classified as part of a filament if at least 4 of its 10 nearest neighbors had TIDIM < 1.5, and if their average inter-point distance was < 0.0293. This combined dimensionality and proximity criterion was crucial for capturing the linearity and compactness of filamentary structures (Fig. 7).

Across 10 independent trials containing 10 filaments each:

True positives (TP, mean detected filaments): 8.4 ± 0.84

False negatives (FN, mean missed filaments): 0.1 ± 0.32

False positives (FP, erroneously accepted): 1.3 ± 0.46

Although slightly less accurate than for compact clusters, the detection rule proved effective for identifying elongated 1D filaments amid larger noise.

The used structure detection schemes provide high detection efficiency, requiring no a priori clustering model, making TIDIM a robust unsupervised method for structure recovery in noisy, high-dimensional astrophysical datasets. False positives remained low in both compact and extended scenarios, suggesting the method's robustness against background noise.

4. Discussion and conclusions

The intrinsic dimensionality method provides a robust, local, and

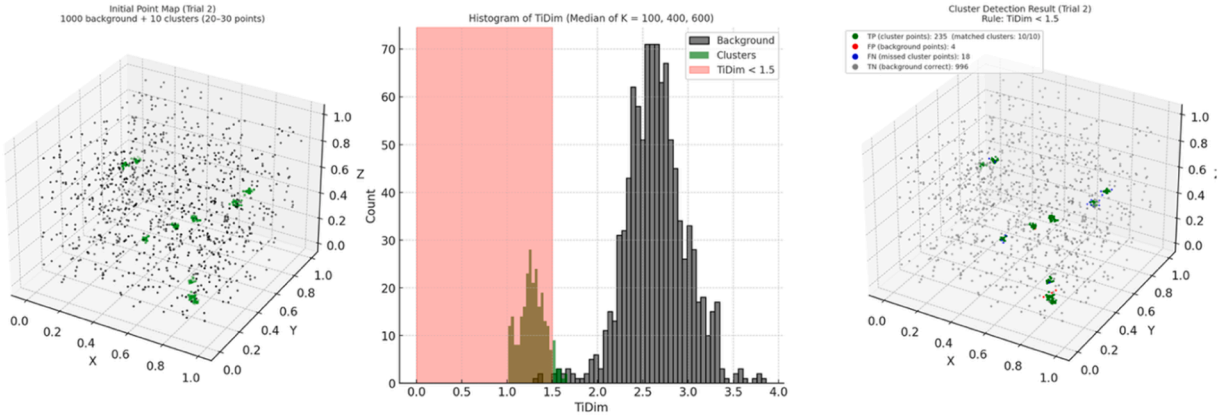

Fig. 6. Three-panel visualization of one of 10 simulation trials. Left: The initial 3D distribution of points. Black dots represent 1000 uniformly distributed background points. Green dots denote the 10 injected compact globular clusters (each 20–30 points). Center: Histogram of TIDIM values for background (gray) and cluster (green) points. The red-shaded area highlights the detection threshold TIDIM < 1.5. Right: Final detection result using the TIDIM-based rule. Green points are true positives (cluster points correctly identified), red points are false positives (background points incorrectly identified as clusters), blue points are false negatives (missed cluster points), and gray points are true negatives (correctly identified background).

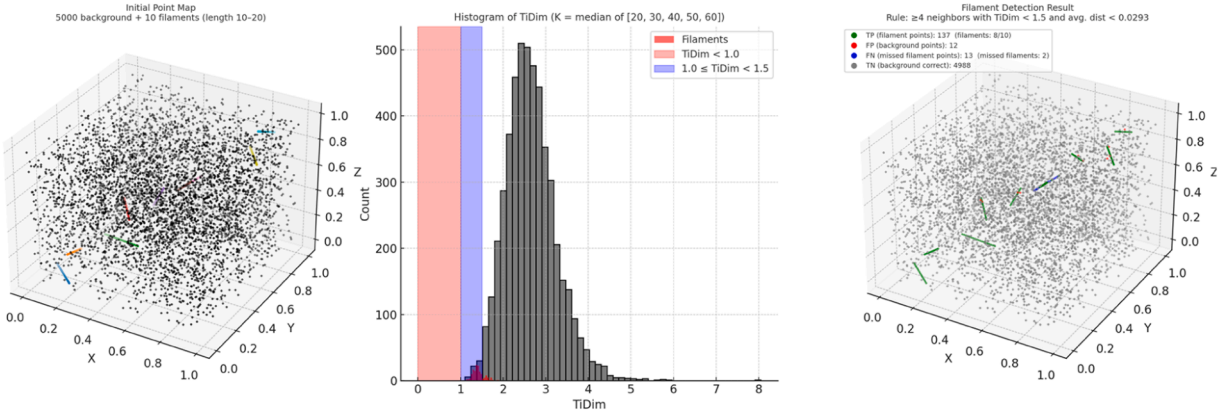

Fig. 7. Three-panel visualization of one simulation trial. Left: 3D view of background with superimposed filaments. Center: Histogram of TIDIM values for background and filament points. Right panel: Green = TP, Red = FP, Blue = FN, Gray = TN.

unsupervised approach to uncovering geometrical structures in noisy 3D environments. Its successful use on synthetic Gaussian data, as well as real-world problems, including both compact globular clusters and sparse filaments, demonstrates its versatility across different contexts. Key advantages include:

- No prior assumptions about shape or size.
- Sensitivity to both dense and sparse structures.
- Adjustable decision rules that incorporate both dimensionality and spatial proximity.

These features make TIDIM a potentially valuable tool for astrophysical data analysis, particularly in characterizing the complex geometry of large-scale galaxy distributions.

Unlike PCA and other global linear methods, TIDIM provides insights into how dimensionality varies across different regions of a dataset, which is essential for feature engineering and anomaly detection. TIDIM's ability to handle both global and local nonlinearities makes it well-suited for complex scientific data such as in particle physics, genomics, and astrophysics. In practical use, TIDIM offers not only the estimated dimensionality but also a framework for localizing and interpreting the geometric structure of the underlying manifold. The TIDIM algorithm enables estimating the number of degrees of freedom in a small neighborhood, mapping the local geometry of the manifold embedded in high-dimensional space, and detecting structures such as folds, branches, resonances, filaments, or jets that lie on curved or nonlinear submanifolds.

Although many methods exist for analyzing large-scale structure in the universe, the TIDIM approach offers a novel and practical way to quantify local geometric complexity in galaxy distributions. Its robustness to noise and independence from prior physical assumptions make it a promising tool for structure recovery in the era of large-scale galaxy surveys.

TIDIM algorithm aligns with ideas found in modern structure-finding and field reconstruction approaches. Key parallels include:

- The Bayesian Origin Reconstruction from Galaxies (BORG, Jasche and Wandelt, 2013) algorithm, which infers the density field from sparse sampling;
- Voronoi and Delaunay tessellation-based methods for void and wall detection (Van de Weygaert and Schaap, 2009);
- Regular-grid KDE and spline-based smoothing used in simulations to interpolate mass or galaxy distributions (Silverman, 1986);
- Machine learning latent-space embeddings that implicitly fill configuration space (Ravanbakhsh et al., 2017).

However, these methods typically require dense sampling or a physical model prior (e.g., gravitational dynamics in BORG). At the same time, TIDIM offers a purely data-driven and non-parametric alternative that can operate in sparse or observationally incomplete

settings.

We demonstrate intrinsic dimensionality methods across various problems, including astrophysical data and simulated astronomical structures. However, additional work with astronomical catalogs of observational and simulated data is necessary to endorse these methods for astronomical applications fully. Only by analyzing astronomical data sets and comparing results with established methods can new techniques be reliably integrated into astrophysical research.

CRediT authorship contribution statement

A. Chilingarian: Writing – original draft, Visualization, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Braams, B.J., 1974. Intrinsic dimensionality estimation. Internal Report, KVI, Groningen.
Chilingarian, A., Harutunyan, S., 1989. On the possibility of a multidimensional kinematic information analysis using nearest-neighbour estimation of dimensionality. Nucl. Instrum. Methods A 281, 388.
Chilingarian, A. 1989. Dimensionality analysis of multiparticle production at high energies, Yerevan Physics Institute Preprint N I-1213(90)-89 .
Chilingarian, A., 1992. Dimensionality analysis of multiparticle production at high energies. Comput. Phys. Commun. 69 (1992), 347–359.
Coifman, R.R., Lafon, S., 2006. Diffusion maps. Appl. Comput. Harmon. Anal. 21 (1), 5–30.
Jasche, J., Wandelt, B.D., 2013. Bayesian physical reconstruction of initial conditions from large-scale structure surveys. Mon. Not. R. Astron. Soc. 432 (2), 894–913. https://doi.org/10.1093/mnras/stt485.
Grassberger, P., Procaccia, I., 1983. Measuring the strangeness of strange attractors. Phys. D: Nonlinear Phenom. 9 (1–2), 189–208.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science (1979) 313 (5786), 504–507.
Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2 (11), 559–572.
Ravanbakhsh, S., Oliva, J.B., Fromenteau, S., Poczos, B., Schneider, J., Wasserman, L., 2017. Estimating cosmological parameters from the dark matter distribution. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 2407–2416. In: https://proceedings.mlr.press/v70/ravanbakhsh17a.html.
Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science (1979) 290 (5500), 2323–2326.
SDSS Collaboration, 2020. The sixteenth data release of the sloan digital sky survey: first release from the APOGEE-2 southern survey and full release of eBOSS spectra. Astrophys. J. Suppl. Ser. 249 (3), 3.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.
Sobol, I.M., 1979. On the systematic search in a hypercube. SIAM J. Number. Anal 16, 790.
Tempel, E., Stoica, R.S., Martínez, V.J., Liivamägi, L.J., Castellan, G., Saar, E., 2014. Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS. Mon. Not. R. Astron. Soc. 438 (4), 3465–3482.
Tenenbaum, J.B., De Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science (1979) 290 (5500), 2319–2323.
Van de Weygaert, R., Schaap, W., et al., 2009. The Cosmic Web: geometric analysis. In: Martínez, V., et al. (Eds.), Data Analysis in Cosmology, Data Analysis in Cosmology, 665. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-44767-2_13. Lecture Notes in Physics.