Estimating the effective dimension of large biological datasets using Fisher separability analysis

Luca Albergante*
Institut Curie, PSL Research University, Mines ParisTech, INSERM, U900
Paris, France
luca.albergante@curie.fr

Jonathan Bac*
Institut Curie, PSL Research University, Mines ParisTech, INSERM, U900
Centre de Recherches Interdisciplinaires, Paris Diderot University
Paris, France
jonathan.bac@cri-paris.org

Andrei Zinovyev†
Institut Curie, PSL Research University, Mines ParisTech, INSERM, U900
Paris, France
Lobachevsky University
Nizhni Novgorod, Russia
andrei.zinovyev@curie.fr

***Abstract*—Modern large-scale datasets are frequently said to be high-dimensional. However, their data point clouds frequently possess structures, significantly decreasing their intrinsic dimensionality (ID) due to the presence of clusters, points being located close to low-dimensional varieties or fine-grained lumping. We introduce and test a dimensionality estimator, based on analysing the separability properties of data points, on several benchmarks and real biological datasets. We show that the introduced measure of ID has performance competitive with state-of-the-art measures, being efficient across a wide range of dimensions and performing better in the case of noisy samples. Moreover, it allows estimating the intrinsic dimension in situations where the intrinsic manifold assumption is not valid.**

***Index Terms*—high-dimensional data, intrinsic dimensionality, separability, cancer mutation, single cell RNA-Seq**

I. Introduction

High-dimensional data are becoming increasingly available in real-life problems across many disciplines. Multiple research efforts in the field of machine learning are currently focused on better characterising, analysing, and comprehending them. A key feature related to data complexity, which is still largely unexplored, is the *intrinsic dimensionality (ID)*, sometimes also called effective dimensionality, of the cloud of points. Informally, ID describes the effective number of variables needed to approximate the data with sufficient accuracy. ID can be measured both globally and locally (i.e., by segmenting the data cloud) [1].

Different approaches have been used to formalise the concept of ID. We refer the reader to other works for a list and comparisons of the different definitions and a discussion of their properties [1], [2]. A compact presentation of the currently used definitions is also available in Section II.

Despite the great diversity of currently used approaches for defining the intrinsic data dimensionality, most of them assume that there exists a relatively low-dimensional variety embedded into the high-dimensional space around which the data cloud is organised. Moreover, it is frequently assumed that the nature of this variety is a manifold, and that the data point cloud represents an i.i.d. sample from the manifold with some simple model of noise. In practice, the ID of the manifold is assumed to be not only much smaller than the number of variables defining the data space but also to be small in absolute number. Thus, any practically useful non-linear data manifold should not have more than three or four intrinsic degrees of freedom.

Theoretically, the manifold concept does not have to be universal in the case of real-life datasets. Even if a low-dimensional variety exists, it can be more complex than

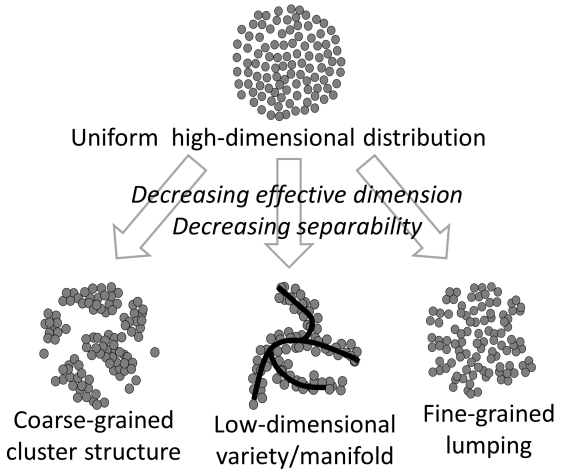

Fig. 1. Stereotypical scenarios of real-life high-dimensional data point cloud organisation, affecting the effective dimensionality of the data. (Top) Data point cloud can be modeled as a close-to-uniform multidimensional distribution (and benefit from "blessing of dimensionality" in high dimensions). (Bottom-left) Data point cloud can be organized into a relatively small number of well-defined clusters. (Bottom-middle) Data point cloud can be located close to a low-dimensional variety (or manifold, in simple cases). (Bottom-right) Data point cloud can be characterised by a fine-grained lumping (heterogeneity) which can not be well represented as being located close to a manifold of low dimension. In cases represented in the bottom panels the data cloud is characterised by lower ID than the full (ambient) dimension of the data space. The figure is drawn from personal communication with A.N.Gorban and I.Tyukin who coined the term "fine-grained lumping".

* These Authors contributed equally to this work. † To whom correspondence should be addressed. This project was supported by the Ministry of education and science of Russia (Project No. 14.Y26.31.0022)

978-1-7281-2009-6/$31.00 ©2019 IEEE

a simple manifold: for example, it can contain branching points or be of variable local intrinsic dimension. Principal trees, graphs and principal cubic complexes (direct product of principal graphs as factors) have been suggested as a constructive approach to deal with such complex cases [3], [4]. In the case of existence of a well-defined cluster structure in the data cloud, the underlying variety can be thought of as discontinuous (e.g., the model of principal forest [5]).

Conversely, a typical mental image of a "genuinely high-dimensional" data point cloud is a uniformly sampled n-dimensional sphere or a n-hypercube, where $n \gg 1$ (at least several tens). Interestingly, in this model, the data point cloud can enjoy the "blessing of dimensionality", which results in almost any two data vectors being almost orthogonal and almost any data point being linearly separable from the rest of the data point cloud [6]. The separability properties can be used, for example, to provide simple non-destructive (not requiring retraining) correctors for the legacy AI systems [7]–[9]. In this sense, truly high-dimensional data distributions are characterised by surprising "simplicity" as opposite to low-dimensional varieties which can possess rather complex non-linear branching or looping structure.

However, real life datasets can be characterised by properties which are difficult to fit into such simple paradigms. In particular, real-life datasets are expected to significantly violate the i.i.d. sampling assumption. The data inhomogeneity can manifest by the existence of micro-clusters which are not globally organized into a low-dimensional structure [7]. These micro-clusters might be undetectable by standard clustering algorithms because of their small size, fuzziness and instability. Existence of such *fine-grained lumping* in the data (Figure 1) can be also thought of as a decrease in ID. For example, it leads to destroying the separability and measure concentration properties, making the data more similar to lower-dimensional (but uniformly distributed) data point clouds.

A single data point cloud can combine regions with several structure types described above in different regions of the data space, and hence be characterised by variable ID. In this context, two important questions are: 1) what parts of the point cloud can be reasonably approximated by a low-dimensional object (e.g., principal curves or trees) and which can not? 2) for those parts which can not be described by a locally small intrinsic dimension, can we estimate how close we are to the "blessing of dimensionality" scenario and can we profit from it, or not? In this work, we suggest an approach for answering these questions.

In recent works by A. Gorban, I. Tyukin and colleagues, the authors proved a series of stochastic separation theorems that can be used to define the properties of high-dimensional data distributions in an efficient and scalable fashion. The authors exploited a convenient framework of Fisher linear discriminants [6]–[8]. We show that such framework can be adapted to construct computationally efficient estimators of local dimensionality tackling different data organization types (Figure 1). This analysis will then be applied to biological data to show the value of dimensionality analysis in deriving actionable information from data.

Throughout the text, R^n will denote the Euclidean n-dimensional linear real vector space, $\mathbf{x} = (x_1, \ldots, x_n)$ the elements of R^n, $(\mathbf{x}, \mathbf{y}) = \sum_{k=1\ldots n} x_k y_k$ the inner product of $\mathbf{x}$ and $\mathbf{y}$, and $||\mathbf{x}|| = \sqrt{(\mathbf{x}, \mathbf{x})}$ the standard Euclidean norm in R^n. $S^{n-1} \subset R^n$ will denote the unit n-dimensional sphere and $|Y|$ or N the number of points in a finite set Y.

II. Defining and measuring intrinsic dimensionality

Despite being used in machine learning research, the term intrinsic dimensionality lacks a unique consensus definition [1]. One of its first use traces back to the context of signal analysis, referring to the minimum number of parameters required by a signal generator to closely approximate each signal in a collection [10]. Other authors, e.g. [11] define the ID of a dataset to be m if it lies entirely within an m-dimensional manifold embedded in R^n with none or little information loss. By shifting the attention from a finite set of points to a generating process, other authors say that the data generating process Y_i has ID m if Y_i can be written as $Y_i = X_i + \epsilon_i$, where X_i is sampled according to a probability measure with a smooth density and with a support on a smooth $m-$dimensional manifold M, ϵ_i is a noise component which is small on a scale where M is well approximated by a m-dimensional subspace [12]. These definitions are grounded in the so-called *manifold hypothesis*, i.e. that data is sampled from an underlying m-dimensional manifold. Following this hypothesis, the goal of ID estimation is to recover m.

While these definitions are very important to better comprehend the problem at hand, they do not provide a way to directly estimate ID. Over the years, researchers devised different estimators, which can be roughly classified by their mode of operation (see [1] for the details of these categories).

Topological methods explicitly seek to estimate the topological dimension (e.g. as defined by the covering dimension) of a manifold. However, they are unsuitable for most practical applications [1], [13], [14]. *Fractal methods* are grounded in the theory of fractal geometry and have been developed by adapting the ideas originally used to study strange attractors' dimensionality. *Projective methods* use different approaches (such as multi-dimensional scaling (MDS) or principal component analysis (PCA)) that perform a mapping of the points into a relatively low-dimensional subspace, by minimising some cost function, which should not exceed certain threshold (e.g. the reconstruction error in ISOMAP) [15]. *Graph-based methods* exploit scaling properties of graphs, such as the length of the geodesic minimum spanning tree [16]. Finally, the *Nearest neighbours* category includes those methods that work at the local level, and rely on properties of distributions of distances or angles. It is worth noting that some recent estimators in this category have been expressly designed to exploit properties of concentration of measure [12], [17]–[19].

One of the most popular dimensionality estimator uses the notion the fractal correlation dimension [20], which is based on the fact - also exploited by many other estimators - that

the number of points contained in a ball of growing radius r will scale exponentially with the dimension of the underlying n-manifold. This counting process is performed by computing the correlation sum :

$$C(r) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i<j} \mathcal{H}(r - ||\mathbf{x}(i) - \mathbf{x}(j)||)$$

with $\mathcal{H}$ the Heaviside step function $\mathcal{H}(x) = \{0, \textit{ if } x < 0; 1, \textit{ if } x \geq 0\}$. The dimension m is then

$$m = \lim_{r\to 0} \frac{\log C(r)}{\log r}$$

In practice, m is approximated by fitting a linear slope of a series of estimates of increasing r and C in logarithmic coordinates.

As outlined before [2], an ideal estimator should be robust to noise, high dimensionality and multiscaling, as well as accurate and computationally tractable. Moreover, it should provide a range of values for the input data in which it operates properly. As of today, no single estimator meets all these criteria and using an ensemble of estimators is generally recommended.

Many dimensionality estimators provide a single value for the whole dataset and thus belong to the category of global estimators. However, datasets can be composed of complex structures with zones of varying dimensionality. In such a case, the dataset should be explored using local estimators, which estimate ID for each point by looking at its neighbourhood. The neighbourhood is typically defined by taking a ball, with a predetermined fixed radius, centered in the reference points or by considering the k closest neighbours. Such approaches allow repurposing global estimators as local estimators. Notably, it is also possible to partition the data into contiguous areas and compute the dimensionality in each of them. However, this may lead to unwanted border effects.

The idea behind local ID estimation is to operate at a scale where the manifold can be approximated by its tangent space [2]. The data contained in each neighbourhood is thus usually assumed to be uniformly distributed over an m-dimensional ball [12], [17]–[19]. In practice, ID proves sensitive to scale and finding an adequate neighbourhood size can be difficult, as it requires a trade-off between opposite requirements [1], [21]. Ideally, the neighbourhood should be big relative to the scale of the noise, and contain enough points for the chosen method to work properly. At the same time, it should be small enough to be well approximated by a flat and uniform tangent space.

III. Estimating intrinsic data dimension based on separability properties

In the present work, we will follow the framework and notations on estimating the dimensionality of a data point cloud based on the description provided in the works by A.Gorban, I.Tyukin and their colleagues [7].

We remind the reader that a point $\mathbf{x} \in R^n$ is linearly separable from a finite set $Y \subset R^n$ if there exists a linear functional l such that $l(\mathbf{x}) > l(\mathbf{y})$ for all $y \in Y$. If for any point $\mathbf{x}$ there exists a linear functional separating it from all other data points, then such a data point cloud is called *linearly separable* or 1-convex. The separating functional l may be computed using the linear Support Vector Machine (SVM) algorithms, the Rosenblatt perceptron algorithm, or other comparable methods. However, these computations may be rather costly for large-scale estimates. Hence, it has been suggested to use the simplest non-iterative estimate of the linear functional by Fisher's linear discriminant which is computationally inexpensive, after a well-established standardised pre-processing described below [7].

Let us assume that a dataset X is normalized in the following (standardised) way:

1) centering
2) projecting onto the linear subspace spanned by first k principal components, where k may be relatively large
3) whitening (i.e., applying a linear transformation after which the covariance matrix becomes the identity matrix)
4) normalising each vector to the unit length, which corresponds to the projection onto a unit sphere.

The 4$^{\text{th}}$ transformation (projecting on the sphere) is optional for the general framework previously defined, but it is necessary for comparing the data distribution with a unity sphere. Choosing the number of principal components to retain in the 2$^{\text{nd}}$ step of the normalisation has the objective of avoiding excessively small eigenvalues of the covariance matrix (strong collinearity in the data). An effective way to estimate k, is by selecting the largest k (in their natural ranking) such that the corresponding eigenvalue λ_k is not smaller that λ_1/C, where C is a predefined threshold. Under most circumstances, $C = 10$ (i.e., the selected eigenvalue is 10 times smaller than the largest one) will result in the most popular linear estimators to work robustly.

After such normalization of X, it is said that a point $\mathbf{x} \in X$ is Fisher-linearly separable from the cloud of points Y with parameter α, if

$$(\mathbf{x}, \mathbf{y}) \leq \alpha(\mathbf{x}, \mathbf{x}) \tag{1}$$

for all $\mathbf{y} \in Y$, where $\alpha \in [0, 1)$. If equation (1) is valid for each point $\mathbf{x} \in X$ such that Y is the set of points $\mathbf{y} \neq \mathbf{x}$ then we call the dataset X Fisher-separable with parameter α. In order to quantify deviation from perfect separability, let us introduce $p_\alpha(\mathbf{x})$, the probability that a point $\mathbf{x}$ is not separable from a random point y. Let us denote $\bar{p}_\alpha(\mathbf{X})$ the mean value of the distribution of $p_\alpha(\mathbf{x})$ over all data points.

Following [7], for the equidistribution on the unit sphere $S^{n-1} \in R^n$, p_α does not depend on the data point thanks to the distribution symmetry. It can be estimated as follows :[1].

[1]In [22], this formula, derived for large n, has $\alpha\sqrt{2\pi(n-1)}$ in the denominator. We empirically verified (see Numerical examples section) that changing the denominator to $\alpha\sqrt{2\pi n}$ makes this formula applicable for low dimensions, and the two expressions are very close for large n, since $n/(n-1) \to 1$. In [7] this formula contains a misprint (personal communication with the authors of [7]).

$$p_\alpha = \bar{p}_\alpha = \frac{(1-\alpha^2)^{\frac{n-1}{2}}}{\alpha\sqrt{2\pi n}} \quad (2)$$

Therefore, the distribution of p_α for a uniform sampling from an $n-$sphere is a delta function centered in $\bar{p}_\alpha$. The effective dimension of a data set can be evaluated by comparing $\bar{p}_\alpha$ for this data set to the value of $\bar{p}_\alpha$ for the equidistributions on a ball, a sphere, or the Gaussian distribution. Comparison to the sphere is convenient thanks to having an explicit formula (2). In order to use this formula, one should project data points on a unit sphere. If $\bar{p}_\alpha$ can be empirically estimated for a given α, then the effective dimension can be estimated by solving (2) with respect to n:

$$n_\alpha = \frac{W(\frac{-\ln(1-\alpha^2)}{2\pi\bar{p}_\alpha^2\alpha^2(1-\alpha^2)})}{-\ln(1-\alpha^2)} \quad (3)$$

where $W(x)$ is the real-valued branch of the Lambert function [23]. As a reminder, the Lambert function solves equation $v = we^w$ with respect to w, i.e. $w = W(v)$. By substituting $w = -\ln(1-\alpha^2)n$, the formula 2 can be re-written as $we^w = -\frac{\ln(1-\alpha^2)}{2\pi\bar{p}_\alpha^2\alpha^2(1-\alpha^2)}$ from which it follows 3. The self-contained description of the algorithm for computing n_α is provided below (Algorithm 1).

Based on the above definitions, the fine-grained lumping of the data point cloud can be identified by two interesting features: the histogram of empirical p_α distribution (probabilities of individual point non-separability) and the profile of intrinsic dimensions n_α (3) for a range of α values (e.g., $\alpha \in [0.5, 1.0)$).

Algorithm 1 Computing data point cloud effective dimension from Fisher-separability with parameter α

For a given data matrix X
Center the data by columns$X \leftarrow X - \bar{X}$
Apply PCA: $[U, S] = PCA(X)$,
where U*are projections onto principal vectors,*
and S *are explained variances*
Select the number of components:
$k = max\{i : S(1)/S(i) < C\}$
For columns of U, u_i, *apply data whitening:*
$u_i \leftarrow u_i/\sigma(u_i), i = 1...k$
Project the data vectors, rows of U, u_j, *onto a unit sphere:*
$u_j \leftarrow u_j/||u_j||, j = 1...N$
Compute the Gram matrix $G = UU^T$
Normalize the Gram matrix by the diagonal elements:
$G_{ji} \leftarrow G_{ji}/G_{ii}$
Set to zero diagonal elements of G: $G_{ii} = 0$
For each row of G, compute the number of elements exceeding α*:* $v_j = \#G_{ji} > \alpha, i, j = 1...N$
Compute empirical unseparability probability distribution:
$p_\alpha^j = v_j/(N-1)$
Compute empirical mean of p_α *:*$\bar{p}_\alpha = \frac{1}{N}\sum_{i=1...N} p_\alpha^j$
Compute intrinsic dimension n_α *from the formula (3)*

IV. Numerical results

A. Benchmark data

We first checked that the method correctly determines the dimension of uniformly sampled n-dimensional spheres (Figure 2). The ability to correctly estimate the dimension in this case depends on the accuracy of estimating the mean empirical unseparability probability for α sufficiently close to 1 which requires a certain number of data points.

The performance of ID estimation methods is usually assessed on synthetic data consisting of samples generated from n-dimensional manifolds linearly or non-linearly embedded into a higher dimensional space. The results are then evaluated according to the *mean percentage error*, defined as:

$$Mean\%error = \frac{100}{\#\{M_i\}} \sum_{i=1}^{\#\{M_i\}} \frac{|\hat{n}_{M_i} - n_{M_i}|}{n_{M_i}}$$

where $\hat{n}_{M_i}$ is the estimated ID and n_{M_i} the true ID of the dataset M_i [24]. Different datasets have been used for this purpose. Here, we use the benchmark library made available by Hein and Audibert [25], which is standard across publications as a core benchmark battery. It consists in 13 uniformly sampled manifolds, to which we added isotropic gaussian noise with standard deviation $\sigma = .05$. We also used the ISOMAP Faces dataset [15], which is composed of images from a sculpture's face generated with three degrees of freedom (horizontal pose, vertical pose, lighting direction). Results are shown for different estimators (Table I), including those recently published and not covered in the existing reviews [1], [12], [26].

We find that Fisher separability is an accurate estimator of ID across the manifold library. Notably, it is one of the few methods performing well in high dimension. Indeed, methods exploiting concentration of measure (*FisherS*, *DANCo*, *ESS*) manage to give a close estimate for $\mathbf{M_{10d}}$, a 70-cube, while all other methods largely underestimate the dimension. We observed that the performance of the Fisher separability method was close with *DANCo* and *ESS*. However, *FisherS* estimated well small effective dimensions in addition to large ones. Both *ESS* (implemented in R) and *FisherS* (implemented in Python 3) are faster than *DANCo* (implemented in MATLAB), which scales worse with respect to increasing dimension (respectively 0.5s, 1.9s, 25.6s on $\mathbf{M_{10d}}$, over an average of 7 runs).

Additionally, we generated three versions of a dataset with random clusters, to illustrate the idea of using the separability probability distribution to characterise non-homogeneities in the data cloud. These datasets consist in a mixture of samples from a uniform distribution $U(0, 1)$ and from uniformly sampled balls centered at random points (Figure 3).

The complete analysis containing more detailed results is available as an interactive Python 3 notebook, including the necessary code to test additional methods and manifolds. The notebook can interface methods in various languages and thus be a useful basis to perform future benchmark tests.

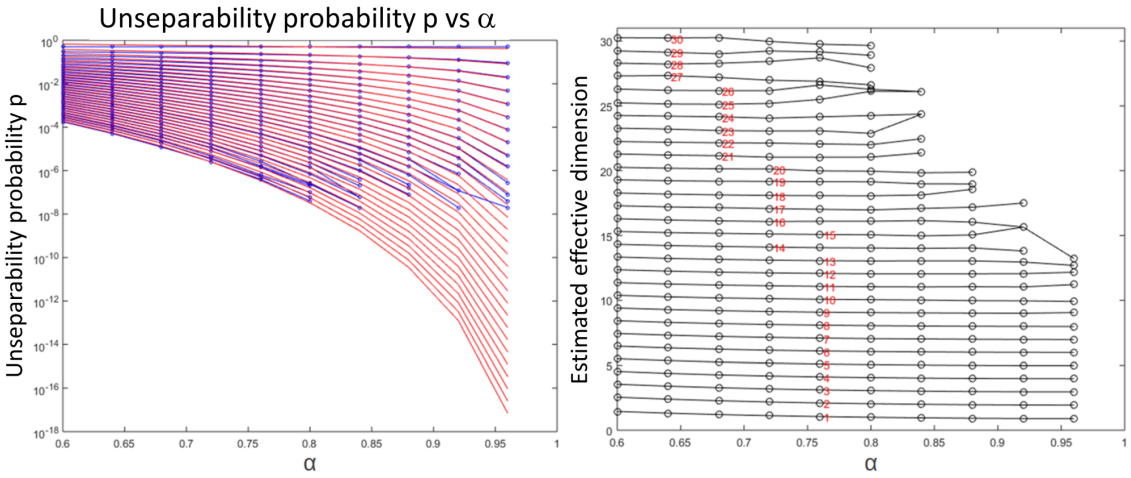

Fig. 2. Estimating effective dimensions of uniformly sampled unit spheres of various dimensions (from 1 to 30). (Left) Comparison of theoretical (red lines) and empirical (blue lines with markers) estimates of the average unseparability probability $\bar{p}_\alpha$. (Right) Estimated effective dimension of n-dimensional spheres ($n = 1..30$), as a function of α. A single ad-hoc estimate of n is indicated by a number, as such n_α for which $\alpha = 0.8\alpha_{max}$, where α_{max} is the maximum value of α for which the empirical $\bar{p}_\alpha > 0$.

TABLE I

Predicted ID for synthetic datasets evaluated globally, with added multidimensional isotropic Gaussian noise (standard deviation $\sigma = .05$), and the ISOMAP Faces dataset. Cardinality: Number of points of the dataset, N: embedding dimension, n: intrinsic dimension. *FisherS*: Fisher Separability (The number in parentheses indicates the number of components retained by PCA preprocessing for the separability-based method), *CD*: Correlation Dimension [20], *GMSTL*: Geodesic Minimum Spanning Tree Length [16], *DANCo*: Dimensionality from Angle and Norm Concentration, *LBMLE*: Levina-Bickel Maximum Likelihood Estimation [27], *ESS*: Expected Simplex Skewness, *FanPCA*: PCA based on [28], *TwoNN*: Two Nearest Neighbors [26]

	Cardinality	N	n	FisherS	CD	GMSTL	DANCo	LBMLE	ESS	FanPCA	TwoNN
$\mathbf{M_{13}}$	2500	13	1	1.67 (3)	1.64	3.73	4	3.74	3.16	2	5.50
$\mathbf{M_5}$	2500	3	2	2.57 (3)	2.14	2.47	3	2.66	2.74	1	2.73
$\mathbf{M_7}$	2500	3	2	2.94 (3)	2	2.24	2	2.39	2.93	2	2.67
$\mathbf{M_{11}}$	2500	3	2	1.96 (2)	2.33	2.21	2	2.49	2.34	1	2.69
Faces	698	4096	3	3.12 (28)	0.78	1.64	4	4.31	7.49	8	3.49
$\mathbf{M_2}$	2500	5	3	2.66 (3)	3.60	4.61	4	4.42	2.66	2	4.69
$\mathbf{M_3}$	2500	6	4	2.87 (4)	3.16	3.36	4	4.40	3.11	2	4.36
$\mathbf{M_4}$	2500	8	4	5.78 (8)	3.90	4.33	4	4.38	7.79	5	3.96
$\mathbf{M_6}$	2500	36	6	8.50 (12)	5.99	6.62	7	7.05	11.98	9	6.27
$\mathbf{M_1}$	2500	11	10	11.03 (11)	8.96	9.02	11	9.88	10.81	7	9.43
$\mathbf{M_{10a}}$	2500	11	10	9.46 (10)	7.86	9.50	10	8.90	10.31	7	8.57
$\mathbf{M_8}$	2500	72	12	17.41 (24)	10.97	13.04	17	14.74	24.11	18	13.15
$\mathbf{M_{10b}}$	2500	18	17	15.94 (17)	11.88	13.15	16	13.89	17.35	13	13.59
$\mathbf{M_{12}}$	2500	20	20	19.83 (20)	10.62	16.05	20	17.07	19.90	11	16.94
$\mathbf{M_9}$	2500	20	20	19.07 (20)	13.51	14.26	19	15.73	20.26	11	15.68
$\mathbf{M_{10c}}$	2500	25	24	22.62 (24)	15.15	21.94	23	18.24	24.42	17	17.36
$\mathbf{M_{10d}}$	2500	71	70	68.74 (70)	29.89	36.62	71	38.92	71.95	43	39.18
Mean%error				28.82	32.45	36.35	43.04	43.83	66.78	67.56	74.91

B. Cancer somatic mutation data: an example of fine-grained lumping

Cancer is a complex disease, that is largely caused by the accumulation of somatic mutation during the lifetime of cells in the organism body. Large-scale genomic profiling provides information on which genes are mutated in the cells composing a tumor at the moment of cancer diagnosis and there is a hope that this information can help driving therapeutic decisions. However, application of standard machine learning methods for this kind of data is difficult because of their extreme sparsity and non-homogeneity of mutation profiles [29]. A mutation matrix (genes vs tumor) in its simplest form is a binary matrix marking non-sense or missense mutation of a certain gene in a chohort of tumors. Because there exists very small overlap between mutation profiles in any two tumors, the data cloud representing a mutation matrix is usually thought to be high-dimensional and suffering from the curse of dimensionality.

We obtained the mutation matrix for 945 breast cancer tumors from The Genome Cancer Atlas (TCGA) as it is provided in [29]. After filtering genes having less than 5 mutations in all tumors, we were left with 2932 genes. For each tumor, we divided its binary mutation profile by the total number of mutations in this tumor, in order to compensate for

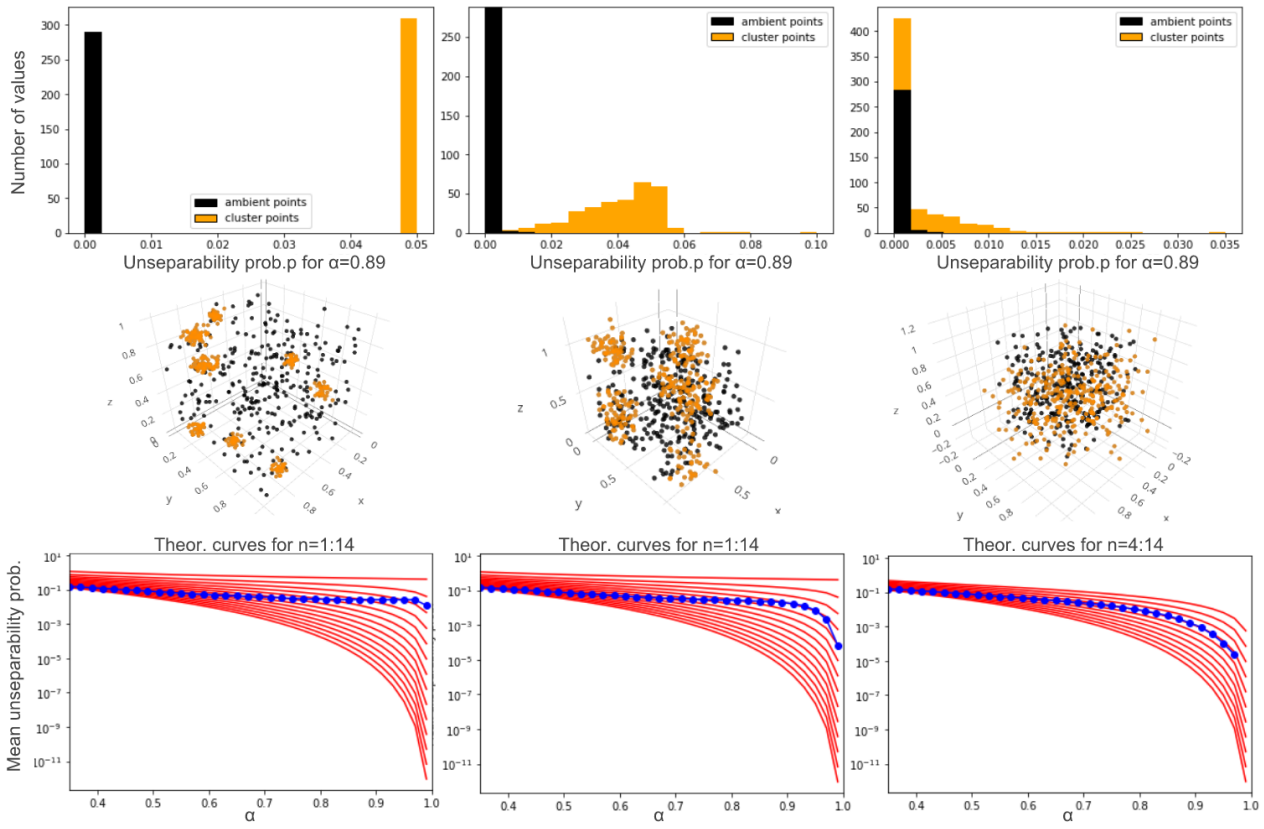

Fig. 3. Illustration of the presented method on datasets consisting of samples from a uniform distribution in the unity 10-dimensional cube, and of 10 clusters formed by choosing random center points to sample uniform 10-balls with radiuses (from left to right) 0.1, 0.3, and 0.6. First row: histogram of unseparability probability $\bar{p}_\alpha$. Second row: 3D scatter plot of the datasets. Third row: The empirical mean unseparability probability as a function of parameter α value (blue) shown on top of the theoretical curves (red) as the clusters become fuzzier due to increasing radius.

large differences in total mutational load between tumors. We analyzed the data point cloud where each point corresponded to a gene, and studied its separability properties using Algorithm 1. The criterion used in the Algorithm 1 for determining the number of principal components selected 34 dimensions, indicating relatively large dimension of the linear manifold approximating the mutation data. Despite this, the separability analysis showed that the separability properties of this data cloud is close to the uniformly sampled 7-dimensional sphere (Figure 4,A,C).

We observed that the p_α probability distributions were overall close to the delta function (Figure 4B), indicating good separability properties of the data cloud. However, there was a non-negligible fraction of data points which could not be separated from the rest of the data cloud even for relatively small $\alpha = 0.88$. We further visualized the data point cloud by applying t-distributed stochastic neighbour embedding (t-SNE) [30], which showed existence of small clusters where the points are less separable, embedded into the sparse cloud of separable points.

C. Highlighting the variable complexity of single cell datasets through separability analysis

Single cell transcriptomics allows the simultaneous measurement of thousands of genes across tens of thousands of cells, resulting in potentially very complex biological big data that can be used to identify cell types or even reconstruct the dynamics of biological differentiation [5], [31].

In a recent work, this technology has been used to explore the different cell types contained in an adult organism of the regenerative planarian Schmidtea mediterranea [32]. Using these data, the authors has been able to identify (via computational analysis) 51 different cell types and the transcriptional changes associated with the commitment of different stem cells (*neoblasts*) into various subpopulations.

Given the complex nature of the data, we decided to use Fisher separability to highlight potential biological properties. After a standard preprocessing pipeline, which included selection of the overdispersed genes and log-transformation of gene expression, the datasets contained 21612 cells characterised by 4515 genes. After an initial filtering that retained 7 PCs, our analysis estimates a global ID close to 4 (5A). By looking at the unseparability probability per cell, we can further appreciate how separability varies across different parts of the dataset (5C). To further explore this aspect we looked at the

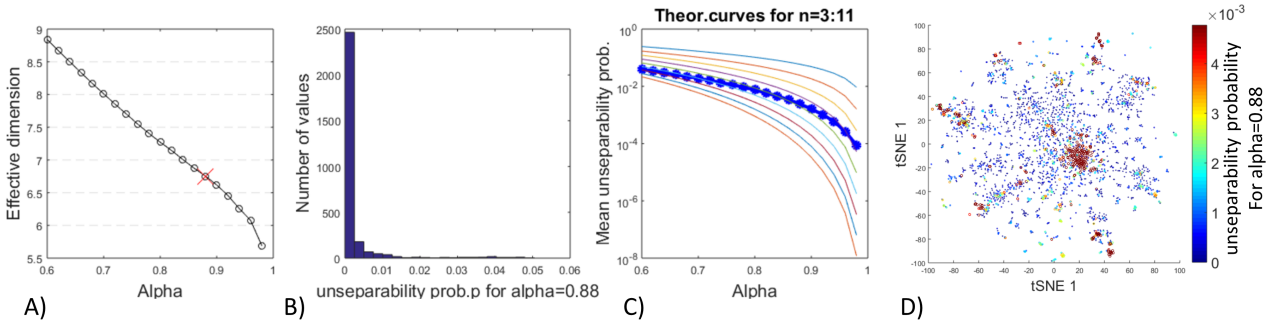

Fig. 4. Analysis of breast cancer somatic mutation data. The initial dataset represents a binary matrix (genes vs tumors) with ones marking any deleterious mutation in a gene found in a tumor. A) Plot showing a range of estimated effective dimensions for a range of α. An single ad-hoc estimate is shown by cross. B) Distribution of unseparability probability distribution for a particular value of $\alpha = 0.88$. C) Empirical estimates for the mean unseparability probability for several α values, shown on top of the theoretical curves for n-dimensional uniformly sampled spheres (starting from n=3). D) tSNE visualization of the dataset (each point corresponds to a gene). Colors show the estimated empirical unseparabity probability p_α for a given data point.

distribution of unseparability probability per cell type.

Interestingly, different populations show a tendency to have different ranges of unseparability probabilities (5E-F) which cannot be explained by the population size (5D). The presence of a multi-peak distribution (5E) indicates the presence of multiple *dimensionality scales* and suggests the presence of micro/meso-clusters embedded into a more uniform manifold.

Remarkably, neurons tend to have a larger unseparability probability, an indication of locally compact distribution and hence of a potentially structured heterogeneity, while epidermal cells are on the other end of the spectrum. The different neoblast populations display a comparable range of unseparability probability, which sits somewhere in the middle, a potential indication of a controlled heterogeneity.

V. Implementation

We provide MATLAB and Python 3 implementations of ID estimation based on data point cloud Fisher-separability at https://github.com/auranic/FisherSeparabilityAnalysis together with benchmarking code.

VI. Conclusion

In this paper, we have exploited the framework of linear Fisher separability in order to estimate the intrinsic dimension of both synthetic and real-life biological datasets. The suggested approach does not assume the presence of a low-dimensional variety around which the data point cloud is organized. According to this framework, deviations from uniformity of data sampling lead to a decrease in the intrinsic dimensionality. Despite this general assumption, the approach demonstrated a surprisingly good performance even for estimating the dimensionality of datasets representing noisy samples from embedded manifolds. The advantages of the method manifest in its efficiency across a wide range of dimensions, robustness to noise, and ability to quantify the presence of fine-grained lumping in the data.

Structures found in the data point clouds resulting from applications of modern biotechnologies might reflect details of molecular mechanisms shaping life. Indeed, computational biology approaches have been capable to gain new insights from mining large-scale molecular datasets and to provide new information that continuously improve our understanding of life and suggest new therapeutic avenues to treat diseases such as cancer. In this paper, we demonstrated how the suggested approach can be used in exploring the structure of two data types that are generally considered to be hard to analyse (mutation and single cell RNA-Seq data) and concluded that separability analysis can provide insights into the organization of their data point clouds.

References

[1] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza, "Intrinsic Dimension Estimation: Relevant Techniques and a Benchmark Framework," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–21, 2015.

[2] F. Camastra and A. Staiano, "Intrinsic dimension estimation: Advances and open problems," *Information Sciences*, vol. 328, pp. 26–41, jan 2016.

[3] A. Gorban, N. Sumner, and A. Zinovyev, "Topological grammars for data approximation," *Applied Mathematics Letters*, vol. 20, no. 4, pp. 382 – 386, 2007.

[4] A. N. Gorban and A. Zinovyev, "Principal graphs and manifolds," *In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, eds. Olivas E.S., Guererro J.D.M., Sober M.M., Benedito J.R.M., Lopes A.J.S.*, 2009.

[5] L. Albergante, E. M. Mirkes, H. Chen, A. Martin, L. Faure, E. Barillot, L. Pinello, A. N. Gorban, and A. Zinovyev, "Robust and scalable learning of data manifolds with complex topologies via ElPiGraph," *CoRR*, vol. abs/1804.07580, 2018. [Online]. Available: http://arxiv.org/abs/1804.07580

[6] A. N. Gorban and I. Y. Tyukin, "Blessing of dimensionality: mathematical foundations of the statistical physics of data," *Philos Trans A Math Phys Eng Sci*, vol. 376, no. 2118, Apr 2018.

[7] A. Gorban, A. Golubkov, B. Grechuk, E. Mirkes, and I. Tyukin, "Correction of AI systems by linear discriminants: Probabilistic foundations," *Information Sciences*, vol. 466, pp. 303 – 322, 2018.

[8] A. N. Gorban and I. Y. Tyukin, "Stochastic separation theorems," *Neural Netw*, vol. 94, pp. 255–259, Oct 2017.

[9] I. Y. Tyukin, A. N. Gorban, K. I. Sofeykov, and I. Romanenko, "Knowledge Transfer Between Artificial Intelligence Systems," *Front Neurorobot*, vol. 12, p. 49, 2018.

[10] R. Bennett, "The intrinsic dimensionality of signal collections," *IEEE Transactions on Information Theory*, vol. 15, no. 5, pp. 517–525, September 1969.

[11] K. Fukunaga, *Intrinsic dimensionality extraction*, ser. in: P.R. Krishnaiah, L.N. Kanal (Eds.), Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics, Vol. 2, North-Holland,

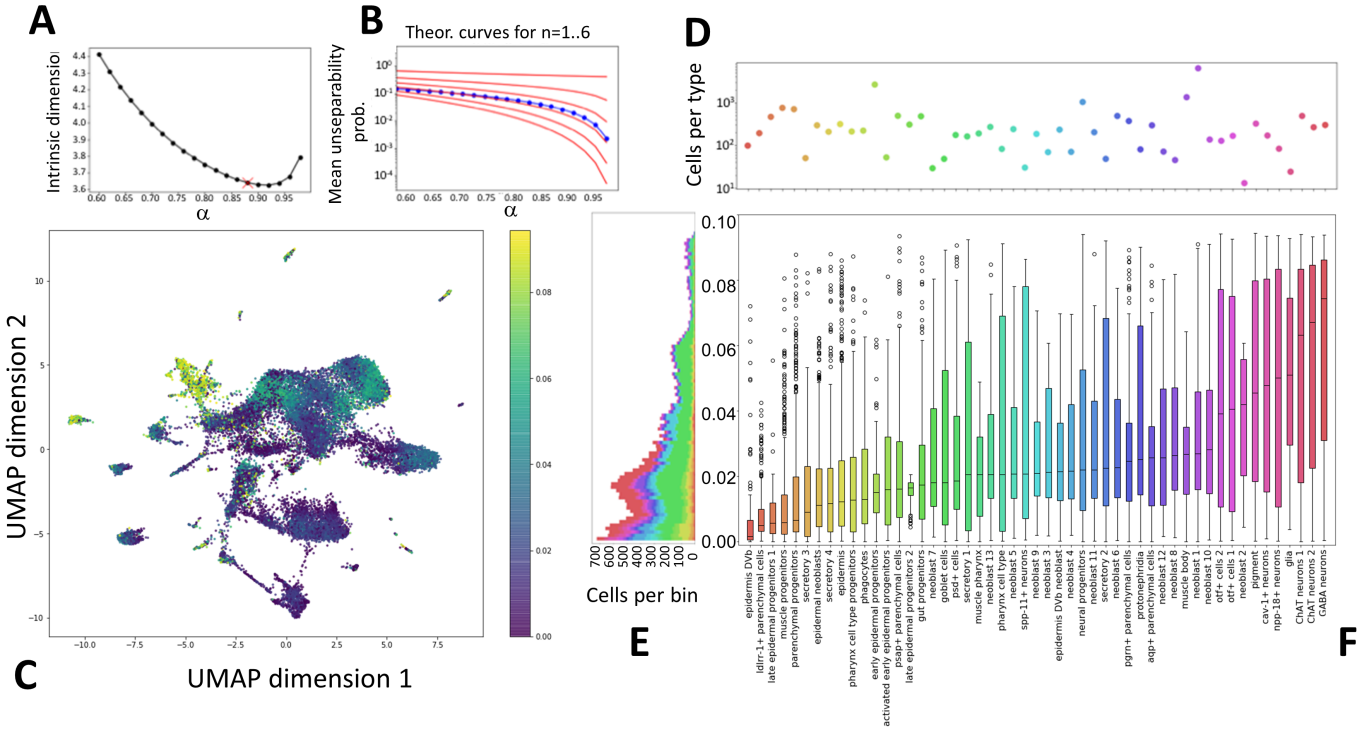

Fig. 5. ID of the planarian transcriptome with data from [32]. **(A-B)** ID and mean unseparability probability as a function of α with the same graphic conventions used before. **(C)** UMAP projection in 2D of the cells of the datasets color-coded according to their unseparability probaiblity. **(D)** Number of cells of different types. Note the logarithmic scale on the y axis. **(E)** Histogram of unseparability probability for $\alpha = 0.88$. **(F)** Boxplot of unseparability probability for $\alpha = 0.88$ by cell type. Note that in panels D-E the different colours indicate the different cell types identified in the original publication.

Amsterdam, 1982, pp. 347362, 1982. [Online]. Available: https://doi.org/10.1016/S0169-7161(82)02018-5

[12] K. Johnsson, "Structures in high-dimensional data: Intrinsic dimension and cluster analysis," Ph.D. dissertation, Faculty of Engineering, LTH, 8 2016. [Online]. Available: https://portal.research.lu.se/ws/files/10994514/Kerstin_Johnsson_PhD_thesis.pdf

[13] P. Mordohai and G. Medioni, "Dimensionality estimation, manifold learning and function approximation using tensor voting," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 411–450, 2010.

[14] C.-G. Li, J. Guo, and B. Xiao, "Intrinsic dimensionality estimation within neighborhood convex hull," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 01, pp. 31–44, 2009.

[15] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[16] J. A. Costa and A. O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, August 2004.

[17] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli, "DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration," *Pattern Recognition*, vol. 47, no. 8, pp. 2569–2581, aug 2014.

[18] M. Daz, A. J. Quiroz, and M. Velasco, "Local angles and dimension estimation from data on manifolds," 2018.

[19] D. R. Wissel, "Intrinsic dimension estimation using simplex volumes," Ph.D. dissertation, 2018. [Online]. Available: http://hss.ulb.uni-bonn.de/2018/4951/4951.htm

[20] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D: Nonlinear Phenomena*, vol. 9, no. 1-2, pp. 189–208, oct 1983.

[21] A. V. Little, Y.-M. Jung, and M. Maggioni, "Multiscale Estimation of Intrinsic Dimensionality of Data Sets," Tech. Rep., 2009. [Online]. Available: www.aaai.org

[22] A. N. Gorban, V. A. Makarov, and I. Y. Tyukin, "The unreasonable effectiveness of small neural ensembles in high-dimensional brain," *Physics of Life Reviews*, Oct 2018. [Online]. Available: http://dx.doi.org/10.1016/j.plrev.2018.09.005

[23] E. W. Weisstein. Lambert W-function. [Online]. Available: http://mathworld.wolfram.com/LambertW-Function.html

[24] J. V. Lindheim, "On intrinsic dimension estimation and minimal diffusion maps embeddings of point clouds," Master's thesis, Freien Universitt Berlin, 2018. [Online]. Available: http://www.zib.de/ext-data/manifold-learning/thesis.pdf

[25] M. Hein and J.-Y. Audibert, "Intrinsic dimensionality estimation of submanifolds in R^d," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 289–296.

[26] E. Facco, M. D'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," *Scientific Reports*, vol. 7, no. 1, p. 12140, dec 2017.

[27] E. Levina and P. J. Bickel, "Maximum Likelihood estimation of intrinsic dimension," pp. 777–784, 2004.

[28] M. Fan, N. Gu, H. Qiao, and B. Zhang, "Intrinsic dimension estimation of data by principal component analysis," Tech. Rep., 2010.

[29] M. Le Morvan, A. Zinovyev, and J. P. Vert, "NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis," *PLoS Comput. Biol.*, vol. 13, no. 6, p. e1005573, Jun 2017.

[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[31] H. Chen, L. Albergante, J. Y. Hsu, C. A. Lareau, G. Lo Bosco, J. Guan, S. Zhou, A. N. Gorban, D. E. Bauer, M. J. Aryee, D. M. Langenau, A. Zinovyev, J. D. Buenrostro, G.-C. Yuan, and L. Pinello, "Single-cell trajectories reconstruction, exploration and mapping of omics data with stream," *Nature Communications*, 2019. [Online]. Available: https://www.biorxiv.org/content/early/2018/04/18/302554.1

[32] M. Plass, J. Solana, F. A. Wolf, S. Ayoub, A. Misios, P. Glažar, B. Obermayer, F. J. Theis, C. Kocks, and N. Rajewsky, "Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics," *Science*, vol. 360, no. 6391, 2018.