

Clustering by the local intrinsic dimension: the hidden structure of real-world data

Michele Allegra^{a,b,*}, Elena Facco^a, Alessandro Laio^{a,c}, and Antonietta Mira^{d,e}

^aScuola Internazionale Superiore di Studi Avanzati, Trieste, Italy

^bInstitut de Neurosciences de la Timone UMR 7289, Aix Marseille Université, CNRS, 13005 Marseille, France

^cInternational Centre for Theoretical Physics, Trieste, Italy

^dUniversità della Svizzera italiana, Lugano, Switzerland

^eUniversità dell'Insubria, Como, Italy

*correspondence: micheleallegra85@gmail.com

Abstract

It is well known that a small number of variables is often sufficient to effectively describe high-dimensional data. This number is called the intrinsic dimension (ID) of the data. What is not so commonly known is that the ID can vary within the same dataset. This fact has been highlighted in technical discussions, but seldom exploited to gain practical insight in the data structure. Here we develop a simple and robust approach to cluster regions with the same local ID in a given data landscape. Surprisingly, we find that many real-world data sets contain regions with widely heterogeneous dimensions. These regions host points differing in core properties: folded vs unfolded configurations in a protein molecular dynamics trajectory, active vs non-active regions in brain imaging data, and firms with different financial risk in company balance sheets. Our results show that a simple topological feature, the local ID, is sufficient to uncover a rich structure in high-dimensional data landscapes.

Introduction

From string theory to science fiction, the idea that we might be glued onto a low-dimensional surface embedded in a space of large dimensionality has tickled the speculations of scientists and writers alike. When it comes to multidimensional data, however, such situation is quite common rather than a wild speculation: data often concentrate on hypersurfaces of low *intrinsic dimension* (ID). Estimating the ID of a dataset is a routine task in data analysis: it yields important information on the global structure of a dataset, and is a necessary preliminary step in several analysis pipelines.

Common approaches for dimensionality reduction, ID estimation and manifold learning assume that the ID is constant in the dataset. This assumption is implicit in projection-based estimators, such as Principal Component Analysis (PCA) and its variants [22], Locally Linear Embedding [29], and Isomap [33]; and it also underlies geometric ID estimators [18, 23, 30], which infer the ID from the distribution of the distances between points.

The hypothesis of a constant ID complies with simplicity and intuition, but is not necessarily valid. In fact, many authors have considered the possibility of ID variations within a dataset [3–5, 7, 10, 11, 14–16, 19, 21, 25, 32, 34, 35], often proposing to cluster the data according to this feature. However the dominant opinion in the community is still that a variable ID is a peculiarity, or a technical detail, rather than a common feature to take into account before performing a data analysis. This perception is at least in part due to the restrictive assumptions which are at the basis of many of the methods developed in this field. Refs. [4, 5, 19, 21, 25] use local ID estimators that implicitly or explicitly assume a uniform density, while refs. [7, 10, 15] jointly estimate the density and the ID from the scaling on the neighbor distances, by approaches which work well only if the density varies slowly and is approximately constant in large neighborhoods of each point. Ref. [32] requires the previous knowledge on the number of the clusters and of their IDs. Refs. [11, 16, 34, 35] all require that the manifolds on which the data lay are hyperplanes, or topologically isomorphic to hyperplanes. These assumptions (locally constant density, and linearity in a suitable set of coordinates) are quite strong in the case of real-world data. Moreover, many of the above approaches [4, 5, 11, 16, 21, 25, 32, 34, 35] work explicitly with the coordinates of the data, while in many applications one only knows the distances between pairs of data points. To our knowledge, only refs. [3, 14] do not make any assumption about the density, as they derive a parametric form of the distance distribution using extreme-value theory, which in principle is valid independently of the form of the underlying density. However, they assume that the low tail of the distance distribution is well approximated by its asymptotic form, an equally non-trivial assumption.

In this work we propose a manner to perform clustering based on the local ID which overcomes all the aforementioned limitations. Building on TWO-NN [12], a recently proposed ID estimator which is insensitive to density variations and uses only the distances between points, we develop a Bayesian framework which allows identifying, by Gibbs sampling, the regions in the data landscape in which the ID can be considered constant. Our approach works even if the data are embedded on highly curved and twisted manifolds, if the manifolds are topologically complex and not isomorphic to hyperplanes and if the probability density from which the data are harvested is non-uniform. Moreover, it is specifically designed to use only the distance between the data points, and not their coordinates. These features, as we will show, make our approach robust and computationally efficient.

Applying our approach to data of various origin, we show that ID variations of a factor two or more between different regions are not a peculiarity. These variations often reveal fundamental properties of the data: for example,

unfolded states in a molecular dynamics trajectory of a protein fall on a manifold of a lower dimension than the one hosting the folded states. Identifying regions of different dimensionality in a dataset can thus be a way to perform an unsupervised classification of the data. This type of clustering is based on very different premises from common approaches: instead of grouping together data according to their density of large-scale organization, we perform classification based on a geometrical property, defined on the local scale: the intrinsic dimension, which identifies the number of linearly independent directions along which neighbouring data are spread.

Methods

A Bayesian approach for discriminating manifolds with different ID

We start from the recently proposed TWO-NN estimator [12], which infers the IDs from the statistics of the distances of the first two neighbors of each point. Let the data $x \doteq (x_1, x_2, \dots, x_N)$, with N the number of points, be sampled from a density $\rho(x)$ defined on a manifold with unknown intrinsic dimension d , such that ρ is approximately constant in the region defined by the second neighbor of each point. If r_{i1} and r_{i2} are the distances of the first and second neighbor of x_i , then $\mu_i \doteq \frac{r_{i2}}{r_{i1}}$ follows the Pareto distribution $f(\mu_i|d) = d\mu_i^{-(d+1)}$. This readily allows the estimation of d from $\mu_i, i = 1, 2, \dots, N$. Assuming that the $\boldsymbol{\mu} \doteq (\mu_1, \mu_2, \dots, \mu_N)$ are independent, we can write the global likelihood of $\boldsymbol{\mu}$ as

$$P(\boldsymbol{\mu}|d) = d^N \prod_{i=1}^N \mu_i^{-(d+1)} = d^N e^{-(d+1)V}, \quad (1)$$

where $V \doteq \sum_{i=1}^N \log(\mu_i)$. From Eq. (1), and upon specifying a suitable prior on d , a Bayesian estimate of d is immediately obtained. TWO-NN can be extended to yield a heterogeneous-dimensional model with an arbitrarily high number of components. Let x be sampled from a density $\rho(x)$ with support on the union of K manifolds with varying dimensions. This multi-manifold framework is common with many previous works investigating heterogeneous dimension in a dataset [5, 7, 10, 11, 15, 17, 19, 32, 34–36]. Formally, let $\rho(x) = \sum_{k=1}^K p_k \rho_k(x)$ where each $\rho_k(x)$ has support on a manifold of dimension d_k and $\mathbf{p} \doteq (p_1, p_2, \dots, p_K)$ are the a priori probabilities that a point belongs to the manifolds $1, \dots, K$. We shall first assume that K is known, and later show how it can be estimated from the data. The distribution of the μ_i is simply a mixture of Pareto distributions:

$$P(\mu_i|\mathbf{d}, \mathbf{p}) \doteq \sum_{k=1}^K p_k d_k \mu_i^{-(d_k+1)}. \quad (2)$$

Following the customary approach [28] we introduce latent variables $\mathbf{z} \doteq (z_1, z_2, \dots, z_K)$ where $z_i = k$ indicates that point i belongs to manifold k . We have $P(\mu_i|\mathbf{d}, \mathbf{p}, \mathbf{z}) =$

$P(\mu_i|z_i, \mathbf{d})P_{pr}(z_i|\mathbf{p})$ with $P(\mu_i|z_i, \mathbf{d}) = d_{z_i}\mu_i^{-(d_{z_i}+1)}$, $P_{pr}(z_i|\mathbf{p}) = p_{z_i}$. This yields the posterior

$$P_{post}(\mathbf{z}, \mathbf{d}, \mathbf{p}|\boldsymbol{\mu}) \propto P(\boldsymbol{\mu}|\mathbf{z}, \mathbf{d})P_{pr}(\mathbf{z}|\mathbf{p})P_{pr}(\mathbf{d})P_{pr}(\mathbf{p}). \quad (3)$$

We use independent Gamma priors on \mathbf{d} , $d_k \sim \text{Gamma}(a_k, b_k)$ and a joint Dirichlet prior on $\mathbf{p} \sim \text{Dir}(c_1, \dots, c_K)$. We fix $a_k, b_k, c_k = 1$, corresponding to a maximally non-informative prior on the \mathbf{p} and an expectation of generally low \mathbf{d} . If one has different a priori expectation on \mathbf{d} and \mathbf{p} , other choices of prior may be more convenient.

The posterior (3) does not have an analytically simple form, but it can be sampled by standard Gibbs sampling [8], allowing for the joint estimation of $\mathbf{d}, \mathbf{p}, \mathbf{z}$. Model [3] has, however, a serious limitation: Pareto distributions with (even largely) different values of d overlap to a great extent. Therefore, the method can not be expected to correctly estimate the z_i : a given value μ_i may be compatible with several manifold memberships. This issue can be addressed by correcting an unrealistic feature of model [3], namely, the independence of the z_i . We assume that the neighborhood of a point is more likely to contain points from the same manifold than from different manifolds. This requirement can be enforced with an additional term in the likelihood that penalizes local inhomogeneity (see also Ref. [19]). Consider the q -neighbor matrix $\mathcal{N}^{(q)}$ with nonzero entries $\mathcal{N}_{ij}^{(q)}$ only if $j \neq i$ is among the first q neighbors of i . Let ζ be the probability to sample the neighbor of a point from the same manifold, and $1 - \zeta$ the probability to sample it from a different manifold, with $\zeta > 0.5$. Define n_i^{in} as the number of neighbors of i with the same manifold membership ($n_i^{in} \equiv \sum_j \mathcal{N}_{ij}^{(q)} \delta_{z_i z_j}$). Then we introduce a probability distribution for $\mathcal{N}^{(q)}$ as:

$$P(\mathcal{N}^{(q)}|\mathbf{z}) = \prod_{i=1}^N \frac{\zeta^{n_i^{in}} (1 - \zeta)^{q - n_i^{in}}}{\mathcal{Z}(\zeta, N_{z_i})} \quad (4)$$

where \mathcal{Z}_q is a normalization factor that depends also on the sizes of the manifolds (see SI for its explicit expression). This term favors homogeneity within a q -neighborhood of each point. With this addition, the model now reads:

$$P_{post}(\mathbf{z}, \mathbf{d}, \mathbf{p}|\boldsymbol{\mu}, \mathcal{N}^{(q)}) \propto P(\boldsymbol{\mu}|\mathbf{z}, \mathbf{d})P(\mathcal{N}^{(q)}|\mathbf{z})P_{pr}(\mathbf{z}|\mathbf{p})P_{pr}(\mathbf{d})P_{pr}(\mathbf{p}) \quad (5)$$

The posterior (5) is sampled with Gibbs sampling starting from a random configuration of the parameters. The parameters \mathbf{d} and \mathbf{p} can be estimated by their posterior averages. As for the \mathbf{z} , we estimate the value of $\pi_{ik} \equiv P_{post}(z_i = k)$. Point i can be safely assigned to manifold k if $\pi_{ik} > 0.8$, otherwise we will consider its assignment to be uncertain.

We name our method Hidalgo (Heterogeneous Intrinsic Dimension Algorithm). Hidalgo has three free parameters: the number of manifolds, K ; the local homogeneity range, denoted by q ; the local homogeneity level, denoted by ζ . As is common for mixture models, the value of K can be estimated by model selection. In particular, we can compare the models with K and $K+1$ for

increasing K , starting from $K = 1$ and stopping when there is no longer significant improvement, as measured with the BICm criterion [27], that makes use only of the posterior samples. Instead, q and ζ are fixed based on preliminary tests conducted on artificial data sets (see SI).

Validation of the method on artificial data

We first tested Hidalgo on artificial data for which the true manifold partition of the data is known. We start from the simple case of two manifolds with different ID, d_1 and d_2 . We consider several examples, fixing the lower dimension d_1 to 4 and varying the higher dimension d_2 from 5 to 9. On either manifold, $N = 1000$ points are sampled from a multivariate Gaussian with unitary variance. The two manifolds are embedded in a space with dimension corresponding to the higher dimension d_2 , with their centers at a distance of 0.5 standard deviations, so they are partly overlapping. In Fig. 1a-b we illustrate the results obtained in the case of fixed $\zeta = 0.5$, equivalent to the absence of any statistical constraint on neighborhood uniformity. The estimate of the two dimensions is shown together with the mutual information (MI) between the estimated assignment of points and the true one. As expected, without a constraint on the assignment of neighbors, the method is not able to correctly separate the points and thus to estimate the dimensions of the two manifolds, even in the case of quite different ID. As soon as we take $\zeta > 0.5$, results improve. A detailed analysis on the influence of the hyperparameters q and ζ is reported in SI. On the basis of such analysis, we identify the following as good parameter ranges: $q \in \{3, 4, 5\}$ and $\zeta \in [0.7, 0.85]$. In Fig. 1c-d we repeat the same tests as in 1a-b but with $q = 3$ and $\zeta = 0.8$. Now the MI between the estimated and ground truth assignment is almost 1 in all cases and correspondingly the estimation of d_1 and d_2 is accurate. To verify whether our approach is able to discriminate between more than two manifolds ($K > 2$), we consider a more challenging dataset consisting of five Gaussians with unitary variance in dimensions 1, 2, 4, 5, 9 respectively. Some of the Gaussians have similar IDs, as in the case of dimensions 1 and 2, or 4 and 5; moreover they can be very close to each other, for instance the centers of those in dimensions 4 and 5 are only half a variance far from each other, and they are crossed by the Gaussian in dimension 1. To analyze such dataset we again choose the hyperparameters $q = 3$ and $\zeta = 0.8$. We do not fix the number of manifolds K to its ground truth value $K = 5$, but we try to let the method estimate K without relying on a priori information. We perform the analysis with different values of $K = 1, \dots, 6$ and compute an estimate of the maximum likelihood \mathcal{L} for each K . Results are shown in Fig. 1e. We see that \mathcal{L} increases up to $K = 5$, and then decreases, from which we infer that the optimal number of manifolds is $K = 5$. In Fig. 1f we illustrate the final assignment of points to the respective manifolds together with the estimated dimensions, upon setting the number of manifolds to $K = 5$. The separation of the manifolds is very good. Only a few points of the manifold with dimension 1 are incorrectly assigned to the one with dimension 2 and vice versa. The values of normalized mutual information between the ground truth and our classification is 0.89.

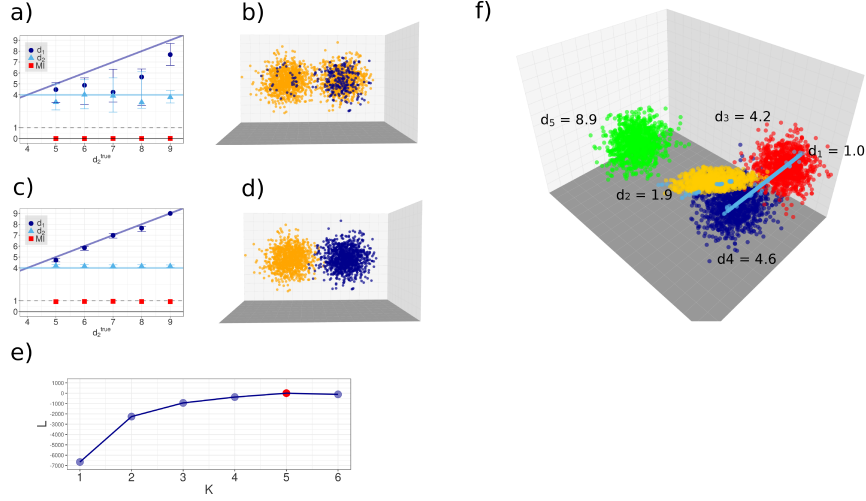


Figure 1: **Results on artificial data sets.** We considered sets of points drawn from mixtures of multivariate Gaussians in different dimensions. In all cases, we performed 10^5 iterations of the Gibbs sampling, and repeated the sampling $M = 10$ times starting from different random configurations of the parameters. We kept the sampling with highest maximum likelihood. **Panels a)-b)**: Points drawn from two Gaussians in different dimensions. The lower dimension is fixed at $d_1 = 4$, the higher varies from $d_2 = 5$ to $d_2 = 9$. $N = 1000$ points are sampled from each manifold. We show results obtained with $\zeta = 0.5$, namely, without enforcing neighborhood uniformity (here $q = 1$, but since $\zeta = 0.5$ the value of q is irrelevant). In panel a) we plot the estimated dimensions of the manifolds and the MI between our classification and the ground truth. In panel b) we show the assignment of points to the low-dimensional (blue) and high-dimensional (orange) manifold (points are projected onto the first 3 coordinates). **Panels c)-d)**: The same setting as in panels a)-b), but now we enforce neighborhood uniformity, using $\zeta = 0.8$ and $q = 3$. Points are now correctly assigned to the manifolds whose ID is properly estimated. **Panels e)-f)**: Points drawn from five Gaussians in dimensions $d_1 = 1, d_2 = 2, d_3 = 4, d_4 = 5, d_5 = 9$. $N = 1000$ points are sampled from each manifold. Some pairs of manifolds are intersecting, as their centers are one standard deviation apart. The analysis is performed with $\zeta = 0.8$, $q = 3$ and with different values of K . In panel e) we show the average log-likelihood \mathcal{L} as a function of K . The maximum \mathcal{L} corresponds to the ground truth value $K = 5$. In panel f) we show the assignment of points to the five manifolds in different colors (points were projected onto the first 3 coordinates).

Results

ID variability in a protein folding trajectory

As a first real application of Hidalgo, we address ID estimation for a dynamical system. Asymptotically, dynamical systems are usually restrained to a low-dimensional manifold in phase space, called an attractor. Much effort has been devoted to characterizing the ID of such attractor [18]. However, in the presence of multiple metastable states an appropriate description of the visited phase space may require the use of multiple IDs. Here, we consider the dynamics of the villin headpiece (PDB entry: 2F4K). Due to its small size and fast folding kinetics, this small protein is a prototypical system for molecular dynamics simulations. Our analysis is based on the longest available simulated trajectory of the system from Ref. [24]. During the simulated 125 μs , the protein performs approximately 10 transitions between the folded and the unfolded state. We expect to find different dimensions in the folded and unfolded state, since these two states are metastable, and they would be considered as different attractors in the language of dynamical systems. Moreover, they are characterized by different chemical and physical features: the folded state is compact and dry in its core, while the unfolded state is swollen, with most of the residues interacting with a large number of water molecules. We extract the value of the 32 key observables (the backbone dihedral angles) for all the $N = 31,000$ states in the trajectory and apply Hidalgo to this data of extrinsic dimension $D = 32$. We obtain a vector of estimated intrinsic dimensions \mathbf{d} and an assignment of each point i to one of the K manifolds. We find four manifolds, three low-dimensional ones ($d_1 = 11.8$, $d_2 = 12.9$, $d_3 = 13.2$) and a high-dimensional one ($d_4 = 22.7$). Note that two spatially separated regions with approximately the same dimension (in this case, d_2 and d_3) are recognized as distinct manifolds by our approach. To test whether this partition into manifolds is related to the separation between the folded and the unfolded state we relate the partition to the fraction of native contacts Q , which can be straightforwardly estimated on each configuration of the system. Q is close to one only if the configuration is folded, while it approaches zero when the protein is unfolded. In Fig. 2a we plot the probability distribution of Q restricted to the four manifolds. We find that the vast majority of the folded configurations ($Q > 0.8$) are assigned to the high-dimensional manifold. Conversely, the unfolded configurations ($Q < 0.7$) are most of the times assigned to one of the low-dimensional manifolds. This implies that a configuration belonging to the low dimensional manifolds is almost surely unfolded. Thus, we can essentially identify the folded state using the intrinsic dimension, a purely topological observable unaware of any chemical detail.

ID variability in time-series from brain imaging

In the next example, we analyze a set of time-series from functional resonance imaging (fMRI) of the brain, representing the BOLD (blood oxygen-level dependent) signal of each voxel, which captures the activity a small part of the

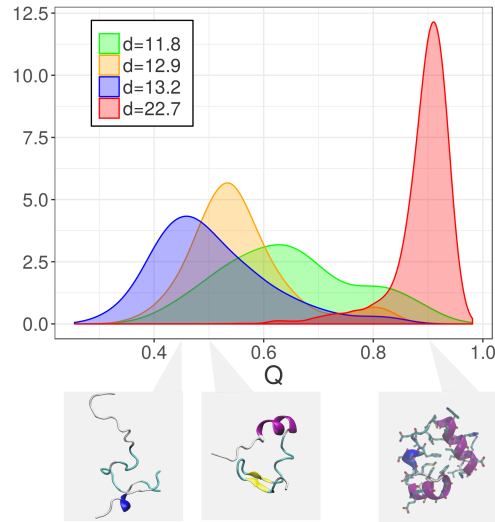


Figure 2: **Analysis of a protein folding trajectory.** We considered $N \sim 31,000$ configurations of a protein undergoing successive folding/unfolding cycles. For each configuration, we extracted the value of the $D = 32$ backbone dihedral angles. Applying Hidalgo to these data, we found four manifolds, of intrinsic dimensions 11.8, 12.9, 13.2 and 22.7. For each configuration, we also computed the fraction of native contacts, Q , which measures to which degree the configuration is folded. The figure shows the probability distribution of Q in each manifold. Nearly all the folded configurations belong to the high-dimensional manifold: the analysis essentially identifies the folded configurations as a region of high intrinsic dimension. Results were obtained with $q = 3$ and $\zeta = 0.8$. The distance between each pair of configurations was computed by the Euclidean metric with periodic boundary conditions on the vectors of the dihedral angles.

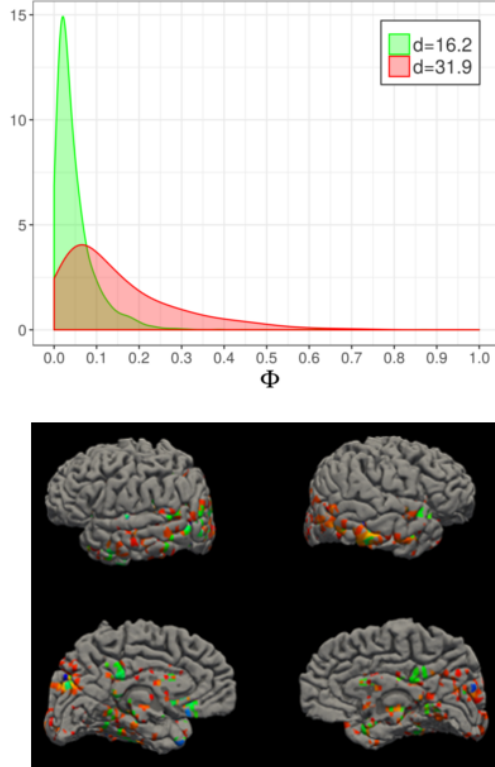


Figure 3: **Analysis of neuroimaging data.** We considered the BOLD time series of $N \sim 30,000$ voxels in an fMRI experiment with $D = 202$ scans. Hidalgo found two manifolds, a low-dimensional one ($d = 16.2$) and a high-dimensional one ($d = 31.9$). For each voxel, we computed the clustering frequency Φ , which measures the participation of each voxel to coherent activation patterns and is a proxy for voxel involvement in the task [2]. (Top) we show the probability distribution of Φ in the two manifolds. Strongly activated voxels ($\Phi > 0.2$) are consistently assigned to the high-dimensional manifold. (Bottom) we report a rendering of the cortical surface (left: left hemisphere; right: right hemisphere). Voxels with high clustering frequency ($\Phi > 0.2$) are shown in blue, voxels assigned to the high-dimensional manifold in red, and voxels satisfying both criteria in green. Almost all voxels with high clustering frequency are assigned to the high-dimensional manifold, and are concentrated in the occipital, temporal and parietal cortex. Results were obtained with $q = 3$ and $\zeta = 0.8$. The distance between two time series was computed by a Euclidean metric, after standard pre-processing steps [1]

brain [20]. fMRI time-series are often projected on a lower dimension through linear projection techniques like PCA [26], a step that assumes a uniform ID. However, the gross features of the signal (e.g., power spectrum and entropy) are often highly variable in different parts of the brain, and also non-uniformities in the ID may well be present. Here, we consider a single-subject fMRI recording containing $D = 202$ images collected while a subject was performing a visuo-motor task [2, 31]. From the images we extracted the $N = 29851$ time series corresponding to the BOLD signals of each voxel. Applying our Hidalgo, we find two manifolds with very different dimensions $d_1 = 16.2$, $d_2 = 31.9$. Again, we relate the identified manifolds to a completely independent quantity, the clustering frequency Φ introduced in [1, 2], which measures the temporal coherence of the signal of a voxel with the signals of other voxels in the brain. Voxels with non-negligible clustering frequency ($\Phi > 0.2$) are likely to belong to brain areas involved in the cognitive task at hand. In Fig. 3 we show the probability distribution of Φ restricted to the two manifolds. We find that the “task-related” voxels ($\Phi > 0.2$) almost invariably belong to the manifold with high dimensionality. These voxels appear concentrated in the occipital, parietal and temporal cortex (Fig. 3b), and belong to a task-relevant network of coherent activity [2]. This result finds a natural and appealing interpretation: the subset of “relevant” voxels give rise to patterns that are not only coherent, but also characterized by a larger ID than the remainder of the voxels. On the contrary, the incoherent voxels exhibit a lower ID, hence a reduced variability, which is consistent with the fact that the corresponding time series are dominated by low-dimensional noise. Again, this feature emerges from the global topology of the data, revealed by our ID analysis, without exploiting any knowledge of the task that the subject is performing.

ID variability in financial data

Our final example is in the realm of economics. We considered firms in the well-known Compustat database ($N = 8309$). For each firm, we consider $D = 31$ balance sheet variables from the fiscal year 2016 (for details, see Table 1 in the SI). Applying Hidalgo we find four manifolds of dimensions $d_1 = 5.4$, $d_2 = 6.3$, $d_3 = 7.0$ and $d_4 = 9.1$. To understand this result, we try to relate our classification with common indexes showing the type and financial stability of a firm. We start by relating our classification to the Fama-French classification [13], which assigns each firm to one of twelve categories depending on the firm’s trade. In Fig. 4a we separately consider firms belonging to the different Fama-French classes, and compute the fraction of firms assigned to each of the four manifolds identified by Hidalgo. The two classifications are not independent, since the fractions for different Fama-French classes are highly non-uniform. More precisely, the mutual information (MI) between the two classifications is 0.19, rejecting hypothesis of statistical independence (p -value $< 10^{-5}$). In particular, firms in the utilities and energy sector show a preference for low dimensions (d_1 and d_2), while firms purchasing products (nondurables, durables, manufacturing chemicals, equipment, wholesale) are concentrated in the manifold with highest

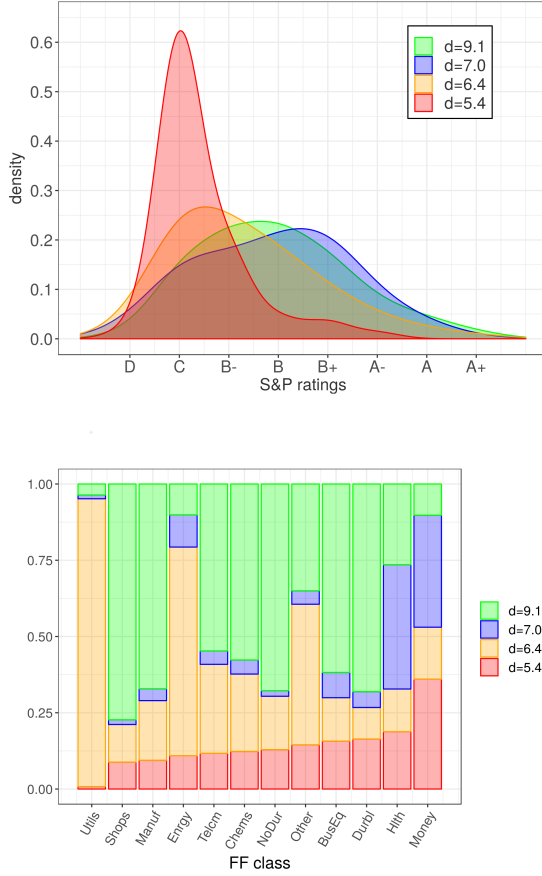


Figure 4: **Analysis of financial data.** We considered $N \sim 8,000$ firms selected from the COMPUSTAT database. For each firm, we considered a set of $D = 31$ variables from their yearly balance sheets. Hidalgo finds four manifolds of intrinsic dimensions 5.4, 6.4, 7.0 and 9.1. (Top) we show the fraction of firms assigned to the four manifolds for each type of firm, according to the Fama-French classification. The four manifolds contain unequal proportions of manifolds belonging to different classes, implying that some classes of firms are preferentially assigned to manifolds of high vs low dimension. (Bottom) we show the probability distribution of the S&P ratings of the firms assigned to each manifold. Firms with low ratings preferentially belong to low-dimensional manifolds. Results were obtained with $q = 3$ and $\zeta = 0.8$. To correct for firm size, we divided the variable vector of each firm by its norm, and then applied standard Euclidean metric on the normalized vectors.

dimension d_4 . The manifold with intrinsic dimension d_3 mostly includes firms in the financial and health care sectors. Different dimensions are not only related to the classification of the firm, but also to their financial robustness. We consider the S&P quality ratings (also from Compustat) for the firms assigned to each manifold. In Fig. 4b we show the distribution of ratings for the different manifolds. These distributions appear to be different. In particular, companies belonging to manifolds of lower dimensions exhibit worse ratings. We suggest a simple interpretation for this phenomenon: a low ID may imply a more rigid balance structure, which may entail a higher sensitivity to market shocks which, in turn, may trigger domino effects in contagion processes. This result shows that a close proxy of the S&P rating can be derived using only topological properties of the data landscape, without any in-depth financial analysis. For example, no information on the commercial relationship between the firms or on the nature of their business is used.

Discussion

The increasing availability of large amount of data has considerably expanded the opportunities and challenges for unsupervised data analysis. Often data come in the form of a completely uncharted “point cloud” for which no model is at hand. A primary goal of the analyst is to uncover some structure within the data. For this purpose, a typical approach is dimensionality reduction, whereby the data are simplified by projecting them onto a low-dimensional space. Another typical approach is clustering, by which the data are classified into several classes grouping together similar elements. In this work we show that these two approaches are deeply entangled. The appropriate *intrinsic dimension (ID)* of the space onto which one should project the data is not constant everywhere. Instead, its variations within a dataset can be used to cluster the data into different categories.

The idea that ID may vary in the same data is not new. In fact, many works have discussed the possibility of a variable ID and developed methods to estimate multiple IDs [3–5, 7, 10, 11, 14–16, 19, 21, 25, 32, 34, 35]. However, these works have received little attention outside the boundaries of specialized literature, probably because they make rather strong assumptions on the structure of the data (for example, a uniform density or the availability of coordinates for the data points).

In this work, we developed a method to cluster the data into regions (manifolds) with different local ID. Our method, named Hidalgo (Heterogeneous Intrinsic Dimension Algorithm), builds on previous contributions but is designed with the specific goal of overcoming technical limitations of other available approaches, and *make local ID-based clustering a general purpose tool*. Our scheme uses only the distances between the data points, and not their coordinates, which significantly enlarges its scope of applicability. Moreover, the scheme uses only the distances between a point and its q nearest neighbours, with $q \leq 5$. We thus circumvent the notoriously difficult problem of defining a globally meaningful

metric [33], only needing a consistent metric on a small local scale. Hidalgo assumes that the data lie on several manifolds with different ID and posits a simple model of the first q nearest-neighbor distances, with unknown parameters corresponding to the number of manifolds, their IDs and sizes, and the assignment of points to the manifolds. All these parameters are estimated via a Bayesian approach resting on Gibbs sampling of the joint posterior distribution for the parameters. By virtue of the linear structure of Gibbs sampling, Hidalgo is computationally efficient and scalable. Moreover, an a priori estimation of the parameters is not required (if available, it can be incorporated in the prior to improve the estimation).

We applied Hidalgo to datasets of diverse origin (a molecular dynamics simulation, a set of time series from brain imaging, a dataset of firm balance sheets). In all cases, we observed large variations of the ID. This finding suggests that a highly non-uniform ID is not an oddity, but a rather common feature. Strikingly, in the cases we analyzed, regions of different dimension were found to host data points differing in important properties. Thus, the ID-based clustering devised in this work is able to retrieve a meaningful structure in the data, leading to a classification of points into fundamentally heterogeneous classes. This classification is enabled by a simple topological property, the local ID, confirming the potential of topological properties for unsupervised data analysis [6, 37],

Not only do our results establish a new clustering criterion. They also suggest a caveat with respect to common practices of dimensionality reduction, which assume a uniform ID. In case of significant variations, a global dimensionality reduction scheme may become inaccurate. In principle, the partition in manifolds obtained with Hidalgo may be the starting point for using standard dimensionality reduction schemes. For example, one can imagine to apply PCA [22] or Isomap [33], or sketchmap [9] separately to each manifold. However, we point out that a feasible scheme to achieve this goal does not come as an immediate byproduct of our method. Once a manifold with given ID is identified, it is highly non trivial to provide a suitable parametrization thereof, especially because the manifolds may be highly nonlinear, and even topologically non-trivial. How to suitably integrate our approach with a dimensionality reduction scheme remains a topic for further research.

Obviously, Hidalgo has some limitations. Some are intrinsic to the way the data are modeled: Hidalgo is not suitable to cover cases in which the ID is a continuously varying parameter [3], or in which sparsity is so strong that points cannot be assumed to be sampled from a continuous distribution. Others are technical issues related with the estimation procedure, and, at least in principle, susceptible of improvement in refined versions of the algorithm. Currently, a major issue consists in the presence of free parameters in the model, especially the number of manifolds K . The likelihood-based method we currently employ to fix K is not fully consistent with the Bayesian approach, as it requires to estimate the model for different values of K . A convenient alternative would be to use a model with an infinite number of components, i.e., a Dirichlet process, which would automatically select the right number of components within a single Gibbs sampling.

Acknowledgements

We thank Giovanni Barone Adesi and Julia Reynolds (Institute of Finance, USI, Lugano, Switzerland) for helping us with the Compustat dataset and offering precious suggestions for its analysis. We thank Giulia Sormani (SISSA, Trieste, Italy) for suggesting, and providing us with the protein dynamics data. We thank Alex Rodriguez (SISSA, Trieste, Italy) for helping with the analysis of protein data. M.A. thanks the SISSA community at large for the substantial moral and intellectual support received, which has been critical for completion of this work.

The authors declare that they have no competing financial interests.

References

- [1] Michele Allegra, Shima Seyed-Allaei, Fabrizio Pizzagalli, Fahimeh Baftizadeh, Marta Maieron, Carlo Reverberi, Alessandro Laio, and Daniele Amati. fmri single trial discovery of spatio-temporal brain activity patterns. *Human brain mapping*, 38(3):1421–1437, 2017.
- [2] Michele Allegra, Shima Seyed Allaei, Nicolas W Shuck, Daniele Amati, Alessandro Laio, and Carlo Reverberi. Brain network dynamics during spontaneous strategy shifts and incremental task optimization. *bioRxiv*, 2018.
- [3] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stephane Girard, Michael E Houle, Ken-Ichi Kawarabayashi, and Michael Nett. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 29–38. ACM, 2015.
- [4] Daniel Barbará and Ping Chen. Using the fractal dimension to cluster datasets. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 260–264. ACM, 2000.
- [5] Paola Campadelli, Elena Casiraghi, Claudio Ceruti, Gabriele Lombardi, and Alessandro Rozza. Local intrinsic dimensionality based features for clustering. In *International Conference on Image Analysis and Processing*, pages 41–50. Springer, 2013.
- [6] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.
- [7] Kevin M Carter, Raviv Raich, and Alfred O Hero III. On local intrinsic dimension estimation and its applications. *IEEE Transactions on Signal Processing*, 58(2):650–663, 2010.
- [8] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

- [9] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, 2011.
- [10] Jose A Costa, Abhishek Girotra, and AO Hero. Estimating local intrinsic dimension with k-nearest neighbor graphs. In *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pages 417–422. IEEE, 2005.
- [11] Ehsan Elhamifar and Rene Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.
- [12] Elena Facco, Maria d’Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.
- [13] Eugene F Fama and Kenneth R French. Industry costs of equity. *Journal of financial economics*, 43(2):153–193, 1997.
- [14] Davide Faranda, Gabriele Messori, and Pascal Yiou. Dynamical proxies of north atlantic predictability and extremes. *Scientific reports*, 7:41278, 2017.
- [15] Aristides Gionis, Alexander Hinneburg, Spiros Papadimitriou, and Panayiotis Tsaparas. Dimension induced clustering. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 51–60. ACM, 2005.
- [16] Alvina Goh and René Vidal. Segmenting motions of different types by unsupervised manifold clustering. In *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*, pages 1–6. IEEE, 2007.
- [17] Andrew Goldberg, Xiaojin Zhu, Aarti Singh, Zhiting Xu, and Robert Nowak. Multi-manifold semi-supervised learning. In *Artificial Intelligence and Statistics*, pages 169–176, 2009.
- [18] Peter Grassberger and Itamar Procaccia. Measuring the strangeness of strange attractors. In *The Theory of Chaotic Attractors*, pages 170–189. Springer, 2004.
- [19] Gloria Haro, Gregory Randall, and Guillermo Sapiro. Translated poisson mixture model for stratification learning. *International Journal of Computer Vision*, 80(3):358–374, 2008.
- [20] Scott A Huettel, Allen W Song, Gregory McCarthy, et al. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.
- [21] Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):1–1, 2015.

- [22] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.
- [23] Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2005.
- [24] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.
- [25] Philippos Mordohai and Gérard G Medioni. Unsupervised dimensionality estimation and manifold learning in high-dimensional spaces by tensor voting. In *IJCAI*, pages 798–803, 2005.
- [26] Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.
- [27] Adrian E Raftery, Michael A Newton, Jaya M Satagopan, and Pavel N Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *biostatistics*, 2006.
- [28] Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.
- [29] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.
- [30] Alessandro Rozza, Gabriele Lombardi, Claudio Ceruti, Elena Casiraghi, and Paola Campadelli. Novel high intrinsic dimensionality estimators. *Machine learning*, 89(1-2):37–65, 2012.
- [31] Nicolas W Schuck, Robert Gaschler, Dorit Wenke, Jakob Heinzle, Peter A Frensch, John-Dylan Haynes, and Carlo Reverberi. Medial prefrontal cortex predicts internally driven strategy shifts. *Neuron*, 86(1):331–340, 2015.
- [32] Richard Souvenir and Robert Pless. Manifold clustering. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 648–653. IEEE, 2005.
- [33] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.
- [34] René Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.
- [35] Yong Wang, Yuan Jiang, Yi Wu, and Zhi-Hua Zhou. Multi-manifold clustering. In *Pacific Rim International Conference on Artificial Intelligence*, pages 280–291. Springer, 2010.

- [36] Rui Xiao, Qijun Zhao, David Zhang, and Pengfei Shi. Data classification on multiple manifolds. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3898–3901. IEEE, 2010.
- [37] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

Supplementary Materials

Enforcing neighborhood uniformity. In our model, we wish to obtain well separated manifolds. We do not wish to impose this condition in the form of a rigid constraint, since in real cases regions with different ID are not completely separated, but only as a “soft constraint”, privileging configurations of \mathbf{z} such that the first neighbors of each point are preferentially assigned to the same manifold. In a Bayesian framework, this means that given that j is among the first neighbors of i , the probability that $z_i = z_j$ is increased. Consider the *neighbor matrix* $\mathcal{N}_{ij}^{(q)}$ defined as:

$$\mathcal{N}_{ij}^{(q)} = \begin{cases} 1 & \text{if } j \neq i \text{ is among the first } q \text{ neighbors of } i \\ 0 & \text{otherwise, including } i = j \end{cases} \quad (6)$$

Intuitively, we would like to impose

$$P_{post}(z_i = z_j | \mathcal{N}_{ij}^{(q)} = 1, \boldsymbol{\mu}, \mathbf{p}) > P_{post}(z_i = z_j | \mathcal{N}_{ij}^{(q)} = 0, \boldsymbol{\mu}, \mathbf{p}). \quad (7)$$

However, Eq. 7 is a relation between posterior probabilities, hence it cannot be directly embedded in the likelihood. What we can specify in the likelihood is the probability of observing the data $\mathcal{N}_{ij}^{(q)}$, given an assignment \mathbf{z} of the points. The way to enforce Eq. 7 is assuming that *the first neighbors of each point are preferentially points of the same manifold*. Consider the i -th row of the neighbor matrix, $\boldsymbol{\mathcal{N}}_i^{(q)} \equiv \{\mathcal{N}_{ij}^{(q)}, j = 1, \dots, N\}$. $\boldsymbol{\mathcal{N}}_i^{(q)}$ is a vector containing q ones and $N - q$ zeros. Without any assumption, all configurations of q zeros and $N - q$ ones are equally likely. Instead, we assume that neighbors are preferentially points from the same manifold. Formally, we assume that neighbors are selected from the points of the same manifold with probability ζ and from a different manifold with probability $1 - \zeta$, with $\zeta > 1/2$. Correspondingly, we introduce a new term in the likelihood:

$$\mathcal{L}(\boldsymbol{\mathcal{N}}_i^{(q)} | \mathbf{z}) = \frac{\zeta^{n_i^{in}(\mathbf{z})} (1 - \zeta)^{q - n_i^{in}(\mathbf{z})}}{\mathcal{Z}(\zeta, N_{z_i})}, \quad (8)$$

where

$$n_i^{in}(\mathbf{z}) = \sum_j \mathcal{N}_{ij}^{(q)} \mathbb{I}_{z_j = z_i} \quad (9)$$

is the number of neighbors of i sampled from the same manifold, and

$$q - n_i^{in}(\mathbf{z}) = \sum_j \mathcal{N}_{ij}^{(q)} \mathbb{I}_{z_j \neq z_i} \quad (10)$$

is the number of neighbors of i sampled from a different manifold. Function \mathcal{Z} is a normalization factor that depends on ζ :

$$\mathcal{Z}(\zeta, N_{z_i}) = \sum_{\{\boldsymbol{\mathcal{N}}_i^{(q)}\}} \zeta^{n_i^{in}(\mathbf{z})} (1 - \zeta)^{q - n_i^{in}(\mathbf{z})}. \quad (11)$$

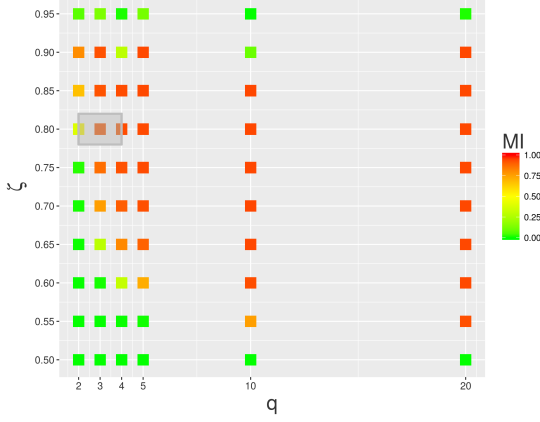


Figure 5: **Choice of parameters**

and can be expressed in a compact way as

$$\mathcal{Z}(\zeta, N_{z_i}) = (1 - \zeta)^q \binom{N - N_{z_i}}{q} {}_2F_1(-q, 1 - N_{z_i}, N - N_{z_i} - q, \frac{\zeta}{1 - \zeta}), \quad (12)$$

where ${}_2F_1(a, b, c, x)$ is the hypergeometric function. The derivation of this expression and the details about the likelihood term in (8) are presented below. By considering all points i , we obtain the global likelihood

$$\mathcal{L}(\mathcal{N}^{(q)} | \mathbf{z}, \zeta) = \prod_i \mathcal{L}(\mathcal{N}_i^{(q)} | \mathbf{z}, \zeta) = \prod_k \frac{\zeta^{n_k^{in}} (1 - \zeta)^{qN_k - n_k^{in}}}{\mathcal{Z}(\zeta, N_k)^{N_k}} \quad (13)$$

where

$$n_k^{in} = \sum_{ij} \mathcal{N}_{ij}^{(q)} \mathbb{I}_{z_i=k} \mathbb{I}_{z_j=k} \quad (14)$$

is the total number of “internal” neighbors of points from manifold k and

$$n_k^{out} = \sum_{ij} \mathcal{N}_{ij}^{(q)} \mathbb{I}_{z_i=k} (1 - \mathbb{I}_{z_j=k}) = qN_k - n_k^{in} \quad (15)$$

is the total number of “external” neighbors of points from k . Note that since \mathcal{Z} depends on i only through the hidden variables \mathbf{z} we are able to split the product into K components.

With this additional term in the likelihood, we obtain

$$\frac{P_{post}(z_i = z_j | \mathcal{N}_{ij}^{(q)} = 1, \boldsymbol{\mu}, \mathbf{p})}{P_{post}(z_i = z_j | \boldsymbol{\mu}, \mathbf{p})} = \frac{\zeta}{1 - \zeta} > 1/2$$

Derivation of the neighborhood uniformity term. With reference to

$\mathcal{N}_i^{(q)}$, without any assumption, all configurations containing q zeros and $N - q$ ones are equally likely. It is easy to compute the number of such configurations. The problem is analogous to the problem of selecting q balls from a box containing $N - 1$ balls: we have to choose q neighbors among $N - 1$ points, point i being excluded. The number of possible choices is $\binom{N-1}{q}$. Hence, all configurations of $\mathcal{N}_i^{(q)}$ being equally likely we would have

$$\mathcal{L}(\mathcal{N}_i^{(q)}|\mathbf{z}) = \binom{N-1}{q}^{-1}, \quad \forall i.$$

Instead, we assume that the neighbors of a point are preferentially points from the same manifold. Formally, we assume that neighbors are selected with probability ζ among the N_{z_i} points assigned to the same manifold of i , and with probability $1 - \zeta$ among the $N - N_{z_i}$ points assigned to a different manifold. Here $\zeta > 1/2$, so that configurations with neighbors assigned to the same manifold are more likely. Now, the problem is analogous to the problem where we have to select q balls from two boxes, a black box containing N_b balls and a white one containing N_w balls. Before selecting each ball, we choose the box, the black one with probability ζ and the white one with probability $1 - \zeta$. Clearly, the probability of a choice with n_b black and $q - n_b$ white balls is then proportional to $\zeta^{n_b}(1 - \zeta)^{q - n_b}$. For a given n_b , the number of possible choices of balls is

$$\binom{N_b}{n_b} \binom{N_w}{q - n_b}$$

One can easily verify that $\sum_{n_b=0}^q \binom{N_b}{n_b} \binom{N_w}{q - n_b} = \binom{N_b + N_w}{q}$. The probability of a given choice is then

$$\frac{\zeta^{n_b}(1 - \zeta)^{q - n_b}}{\mathcal{Z}}$$

where $\mathcal{Z} = \sum_{n_b=0}^q \binom{N_b}{n_b} \binom{N_w}{q - n_b} \zeta^{n_b}(1 - \zeta)^{q - n_b}$. By using the formula (Abramowitz and Stegun, 15.4.1)

$${}_2F_1(-m, b, c, z) = \sum_{n=0}^m (-1)^n \binom{m}{n} \frac{(b)_n}{(c)_n} z^n$$

where $(a)_n = a(a + 1) \dots (a + n - 1)$ is the Pochhammer symbol and doing some simple algebra, \mathcal{Z} can be compactly expressed as

$$\mathcal{Z} = (1 - \zeta)^q \binom{N_w}{q} {}_2F_1(-q, -N_b, N_w - q, \frac{\zeta}{1 - \zeta}).$$

Substituting N_b with $N_{z_i} - 1$ (the number of points assigned to the same manifold as i , excluding i), N_w with $N - N_{z_i}$ (the number of points assigned to a different manifold), and n_b with n_i^{in} , we obtain the likelihood of a given configuration of $\mathcal{N}_i^{(q)}$ as

$$\mathcal{L}(\mathcal{N}_i^{(q)}|\mathbf{z}, \zeta) = \frac{\zeta^{n_i^{in}(\mathbf{z})}(1 - \zeta)^{q - n_i^{in}(\mathbf{z})}}{(1 - \zeta)^q \binom{N - N_{z_i}}{q} {}_2F_1(-q, 1 - N_{z_i}, N - N_{z_i} - q, \frac{\zeta}{1 - \zeta})}.$$

Choice of the free parameters. In order to find a good configuration for the parameters (q, ζ) , we perform tests with several values of $q \in \{2, 3, \dots, 20\}$ and $\zeta \in [0.5, 1)$. We focus on the most challenging case, the one of two Gaussians in dimensions 4 and 5. The crucial figure of merit to assess the performance of the method is the mutual information between the estimated and the true assignment of \mathbf{z} , which measures the quality of the assignment of points to manifolds. Indeed, once the manifolds are correctly separated, the problem is essentially reduced to a dimension estimation within the single manifolds (which is successfully solved by TWO-NN). In supplementary Fig. 5 we show the MI as a function of (q, ζ) for a Gibbs sampling with 10^5 iterations. For most values of q , the MI first increases, then decreases with ζ . This can be expected on the basis of the following considerations. When ζ is close to 0.5, as we discussed above, the method cannot discriminate different manifolds. When ζ is increased, the posterior distribution starts to prefer configurations that approximately satisfy the neighborhood homogeneity constraint. For sufficiently high ζ , the posterior distribution is sharply peaked at the configuration that optimally satisfies this constraint; correspondingly, if the Gibbs sampler is able to explore the parameter space exhaustively, it will eventually find this peaked region and remain trapped there. Hence, the MI achieves average values close to 1. However, for ζ close to 1 the posterior distribution is very likely to also have pronounced local maxima and, depending on the initial configuration, the sampler may remain trapped in one of them. Hence, one can observe a drop in the MI. In general these sampling issues can be worsened when q is increased since the local maxima become more and more pronounced. In principle, this problem may be dealt with by resorting to well established enhanced sampling techniques. For simplicity, in the present work we prefer to verify that there is a region of the parameter space where the results appear stable, and restrict to this regions for subsequent analyses.

#	Variable	#	Variable	#	Variable
1	Acquisitions	12	Liabilities - Total	23	Interest and Related Expense - Total
2	Assets - Total	13	Net Income (Loss)	24	Goodwill
3	Capital Expenditures	14	Operating Income Before Depreciation	25	Intangible Assets - Total
4	Cash	15	Property Plant and Equipment - Total (Net)	26	Pretax Income
5	Common Shares Outstanding	16	Purchase of Common and Preferred Stock	27	Pretax Income - Foreign
6	Common/Ordinary Shareholders	17	Sales/Turnover (Net)	28	Investment and Advances - Equity
7	Debt in Current Liabilities - Total	18	Stockholders Equity - Parent	29	Investment and Advances - Other
8	Long-Term Debt - Total	19	Income Taxes Paid	30	Increase in Investments
9	Cash Dividends (Cash Flow)	20	Research and Development Expense	31	Sale of Investments
10	Earnings Before Interest and Taxes	21	Price Close - Annual - Fiscal		
11	Employees	22	Preferred/Preference Stock (Capital) - Total		

Table 1: Compustat variables used