Journal of Siberian Federal University. Mathematics & Physics 2018, 11(2), 001–012

УДК 57:015 + 573.2

New Method to Determine Topology of Low-Dimension Manifold Approximating Multidimensional Data Sets

Michael G. Sadovsky*
Institute of computational modelling SB RAS, Siberian federal university
Akademgorodok, Krasnoyarsk, 660036 Russia

Anatoly N. Ostylovsky†
Siberian Federal University, Institute of mathematics & fundamental informatics

Received 10.06.2017, received in revised form 30.07.2017, accepted 05.10.2017

New method is proposed to identify topology of a low-dimensional manifold approximating multidimensional datasets. The method is based on the implementation of the compliment for the discrete set of data. Some essential properties and constraints of the method are discussed.

Keywords: order, complexity, clusterization, complement, surface genius

Introduction

Rapid growth of capabilities of hardware and power of software to compute and maintain extended data sets meets a reciprocal growth of available data capacity thus resulting in implementation of basically new approaches to treat multidimensional datasets. Indeed, a search and exploration of a structuredness, or partial order, or patterns in huge multidimensional data sets goes ahead; classical way consisted in formulation of a hypothesis accompanied with further verification of that latter is not currently the only way of a study. A search for structures and patterns being apparent in tremendous datasets may help in further formulation of various hypotheses and suggestions that later could be verified.

Instead, the up-to-date approach pursues to seek for structuredness, order, patterns and other peculiarities that might be gathered into "unexpectedness". That latter is an issue that differs from that one expected from the combination of "smaller" parts, or other "bricks" used to develop an entity [4–6,9]. Clustering techniques become the key issue here, changing the methods based on distribution parameters estimation, etc. Thus, an up-to-date approach to treat the multidimensional data is to model them and approximate with manifolds of lower dimension, with due accuracy. Basically speaking, this approach had taken the start in principle component analysis (PCA), where the linear vector space was the approximating manifold. The point is that PCA is the linear procedure, thus failing to treat properly strongly non-linearly distributed data. So, currently the basic idea is to approximate (and model) the multidimensional data sets with general (non-linear) manifolds.

A sounding progress has been achieved in this direction, both in theory [1–3], and in specific applications [9–11]. Meanwhile, the topology of an approximating manifold becomes the essential constraint here, since the key idea of the approximation is to maintain the topology of the

*msad@icm.krasn.ru
†hinayana@g-service.ru

manifold used to model (or approximate) the data set. The conservation rule does not make a problem itself, theoretically. Indeed, there is (almost) no problem to fit a manifold of the given topology to a dataset; the point is that one knows nothing about the topology of this manifold, in advance. There might be two ways here: to use a manifold with the given topology, and to try to learn the details of the topology of an approximating manifold.

Below we show the problem of the topology impact on the approximation manifold choice. We shall use two-dimensional illustrations, keeping in mind their incompleteness, constraints and fallacy. The simplest example of a data configuration posing a problem for conventional methods of clustering is shown in Fig. 1. Yet, even simpler configuration (that is a two-dimensional torus) makes a problem: one fails to identify reliably any cluster pattern within such data set, if a genus-one manifold is not used to approximate the dataset. Obviously, being a "flatland inhabitant", one fails to make a clue towards the stricture of the torus shown in Fig. 1. A growth of the data set dimension just makes the problem worse.

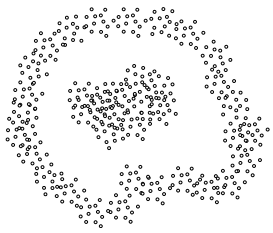


Figure 1: An example of a "tough" topology.

One must avoid a mispresentation of the problem of dimension reduction, and the cast of a dataset elabouration, for determination of a topology of an approximating manifold; the method provided below has nothing to do with the data set dimension reduction, at least, immediately. On the contrary, the method aims to held a researcher to fit the best initial manifold, to model (or approximate) the data.

1. The method

Basically, the idea of the method is rather apparent: to make a kind of cast of the dataset, and then study its structure. More rigorously, the idea is to change a study of the original set for the complement $\overline{\mathfrak{M}}$ of that former. Here a problem arises towards the definition of the complement, and the method essentially addresses the question. Let $\mathfrak{M}$ be the set of multidimensional data points $\mathfrak{m}_j \in \mathfrak{M}$ with index j enlisting the points at the dataset, so that $|\mathfrak{M}| = M$ and $\forall j\ \mathfrak{m}_j \in \mathbb{R}^n$; $\mathfrak{m}_j = (x_1^j, x_2^j, x_3^j, \ldots, x_{n-1}^j, x_n^j)^\intercal$. Here $|\cdot|$ is the capacity of a set.

At the first step, suppose the point $\mathfrak{m}_j \in \mathfrak{M}$ are located in space rather densely, and one is always able to figure out an ellipsoid, or n-dimensional cube, or any other quite simple body Ω gathering the greatest majority of the points, so that they do not diffuse outside of the border slowly (see Fig. 3(a)). To make a cast, one needs to know an average density of the points determined over the ellipsoid; otherwise, one needs to know M figure, and the volume of Ω. So, let the density be equal to d. Suppose, then, that all the points $\mathfrak{m}_j \in \mathfrak{M}$ are colored in blue. Disperse then randomly and independently the points from another set $\mathfrak{L}$, $\mathfrak{l}_j \in \mathfrak{L}$, colored in red (see Fig. 3(b)), supposing that $|\mathfrak{M}| \sim |\mathfrak{L}|$.

At the third step, to reveal the gaps and breaches, one must eliminate all the red points located closely enough to blue ones. The proximity of the points $\mathfrak{l}_j \in \mathfrak{L}$ and $\mathfrak{m}_i \in \mathfrak{M}$ could be determined in a number of ways, see Sec. 1., and the result of $\overline{\mathfrak{M}}$ development may differ, for various definitions of a used metrics. As soon, as the added (red) points $\mathfrak{l}_j \in \mathfrak{L}$ are eliminated, for some $j \in J$, one must eliminate the original set $\mathfrak{M}$ (blue points). Here J is the set of indices of the points $\mathfrak{l}_j \in \mathfrak{L}$ that are located closely enough to the points $\mathfrak{m}_i \in \mathfrak{M}$. Thus, the rest points colored in red (those belonging to $\mathfrak{L}$) represent the complement $\overline{\mathfrak{M}}$ to $\mathfrak{M}$; $\overline{\mathfrak{M}} \subset \mathfrak{L}$. Finally, one should study the complement with a number of convenient methods and advanced techniques [7];

Figs. 3 and 2 illustrate the method.

1. Decision rule to remove the points

A simple idea to eliminate the added (red) points $\mathfrak{l}_j \in \mathfrak{L}$ located proximately enough to the points $\mathfrak{m}_j \in \mathfrak{M}$ to figure out a topology of $\mathfrak{M}$ sounds apparently, but may cause some calamities. A proximity of the added (red) points to the original ones is not defined in an unambiguous and obvious way. The point is that the local density of the point $\mathfrak{m}_j \in \mathfrak{M}$ may depend on a space location, thus making a problem of a "stop" of the selection process of the added (red) points to be eliminated. To choose these points, one may follow two opposing ways. Let now discuss three more issues in the method implementation; these are

a) the method to identify the red (added) points $\mathfrak{l}_j \in \mathfrak{L}$ located closely enough to the original (blue) ones to be eliminated (Sec. 1.);
b) the problem of a "fuzzy" pattern of an original set $\mathfrak{M}$, and
c) the choice of parameters of the distribution of the added points set $\mathfrak{l}_j \in \mathfrak{L}$.

First of all, there are two opposing, to some extent, approaches to define the proximity of the points $\mathfrak{l}_j \in \mathfrak{L}$ to the points $\mathfrak{m}_i \in \mathfrak{M}$: the former is absolutely local, and the latter is based on the overall distribution of point $\mathfrak{m}_i \in \mathfrak{M}$. Locality in $\mathfrak{l}_j \in \mathfrak{L}$ determination means that the nearest point $\mathfrak{m}_i \in \mathfrak{M}$ (the blue one) point is used to do it. Practically, it means that one must cover all point $\mathfrak{m}_i \in \mathfrak{M}$ with balls of the given radius ε, and eliminate all those $\mathfrak{l}_j \in \mathfrak{L}$ (red ones) that fall into the ball

$$\mathfrak{l}_j \in B_\varepsilon(\mathfrak{m}_i)\,.$$

Here $B_\varepsilon(\mathfrak{m}_i)$ is a ball of the radius ε centered at $\mathfrak{m}_i$. So, the set $\widetilde{\mathfrak{L}}$ of eliminated points is defined then as

$$\widetilde{\mathfrak{L}} = \bigcup_{\mathfrak{m}_i \in \mathfrak{M}} \mathfrak{l}_j \in B_\varepsilon(\mathfrak{m}_i)\,. \tag{1}$$

The method (1) is absolutely insensitive to a pattern of whole distribution of points $\mathfrak{m} \in \mathfrak{M}$. Meanwhile, there are no any reasons to eliminate the impact of the whole set $\mathfrak{M}$ on the choice of excluded points belonging to $\mathfrak{L}$: one may want to include the effect of an "environment" on the selection process of the points $\mathfrak{l}_j \in \mathfrak{L}$ to be eliminated. Typical way to do it consists in implementation of some field (called also "glue" here) to take into account the impact of all experimental (blue) points on the choice of the excluded red ones.

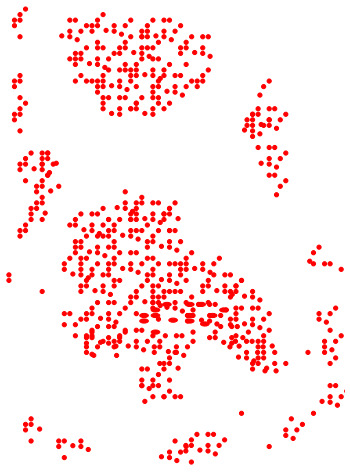
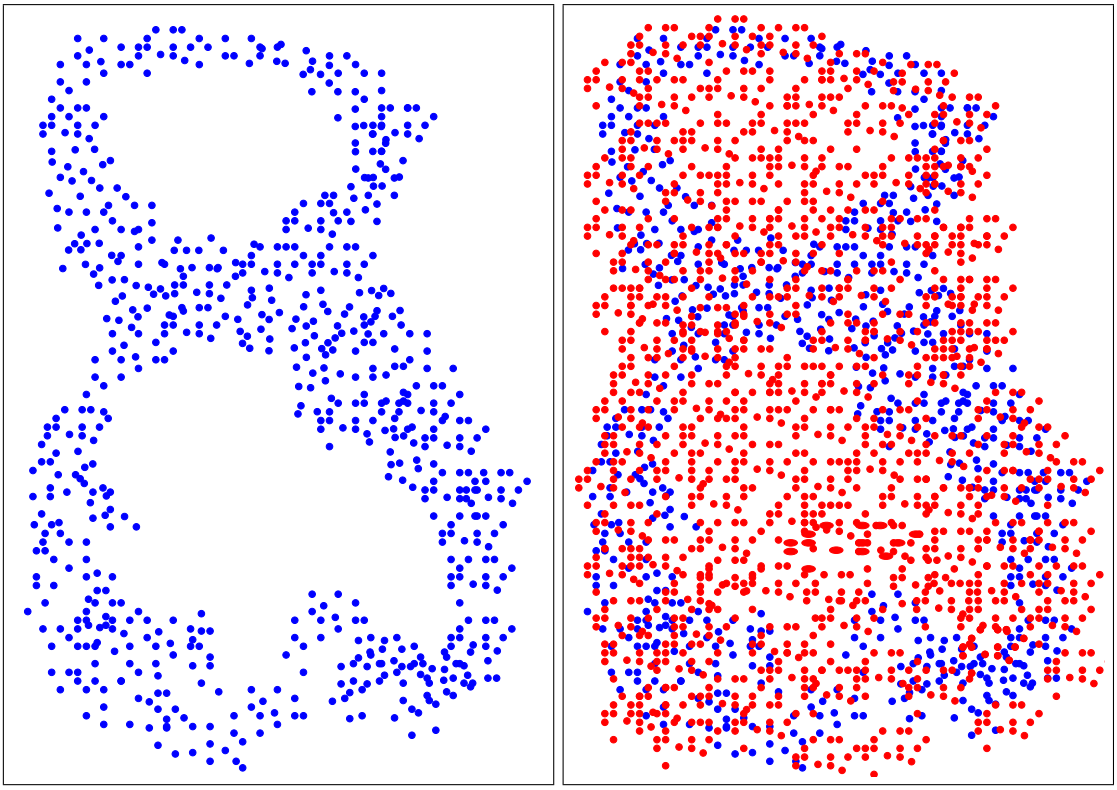


Figure 2: The complement $\widehat{\mathfrak{L}}$.

Another option consists in the similar procedure, while it takes start from the point $\mathfrak{l}_k \in \mathfrak{L}$: considering each point $\mathfrak{l}_k \in \mathfrak{L}$ as a center, cover them with the balls of the radius r, and remove all the centers of the balls containing points from $\mathfrak{M}$. These two procedure yield the same set $\mathfrak{L}^*$ of the points to be removed. Indeed, consider the set $\mathfrak{L}_1^*$ of the points falling inside a ball of the radius r with the center at a point $\mathfrak{m}_0 \in \mathfrak{M}$. Since the radius r is the same, for both procedures, then all the points from $\mathfrak{L}_1^*$ considered as centers would contain the point $\mathfrak{m}_0$ inside the balls centered at the point from $\mathfrak{L}_1^*$. Similar conclusion holds true, if one changes a point $\mathfrak{m}_0$ for $\mathfrak{l}_0$, and the set $\mathfrak{L}_1^*$ for the similarly defined set $\mathfrak{M}_1^*$.



(a) Original set $\mathfrak{M}$. (b) Superposition of the sets $\mathfrak{M}$ and $\mathfrak{L}$.

Figure 3: Illustration of the method to develop the complement $\widehat{\mathfrak{L}}$.

In other words, a relation *to occupy the same ball* is symmetric, for two points x and y, if $\rho(x,y) \leqslant r$, where r is the ball radius.

As it has been said above, the first approach is absolutely local: there is no matter how other points (blue or red ones) are located in space, if one needs to make a decision towards the elimination of a given $\mathfrak{l}_j \in \mathfrak{L}$ (red) point. The second approach resembles, to some extent, a mean-field approach. The idea standing behind this approach is to develop a "glue" that adheres the points to be removed. To do that, supply each point $\mathfrak{m}_j \in \mathfrak{M}$ with a bell-shaped function $f(r)$, and make a sum of all the functions

$$\mathcal{F}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n) = \sum_{\mathfrak{m}_j \in \mathfrak{M}} f_{\mathfrak{m}_j}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n) \tag{2}$$

to get an averaged "potential field". The function $f_{\mathfrak{m}_j}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n)$ in (2) is defined over Ω (or even $\mathbb{R}^n$). Of course, the set of $\mathfrak{l}_j \in \mathfrak{L}$ to be eliminated strongly depends on the type of the function $f_{\mathfrak{m}_j}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n)$. Probably, the function $f(x_1, x_2, x_3, \ldots, x_{n-1}, x_n)$ is not an n-dimensional function, but a single-dimensional one, depending on the radius r measured from the center point $\mathfrak{m}_j$ only:

$$r = \sqrt{\sum_{i=1}^{n} \Big(z_i - x_i\Big)^2}, \tag{3}$$

where x_i is i-th coordinate of a point $\mathbf{x}$, and z_i is the i-th coordinate of the point $\mathfrak{m}_j$.

Function $f(r)$ may be chosen in a number of ways; it must be integrable (probably, with its square) in Ω. Another stipulation is monotonicity of the function; thus, a typical function has a bell-shaped form. Practically, one might want to implement the following functions, in (2):

1. Gaussian function
$$f(r) = \exp\left\{-\frac{r^2}{\beta^2}\right\};$$
this is the classical function to develop a mean-field like approximation, and the motivation to use it comes from probability theory (the law of large numbers). Here parameter β is adjustable one, changing the typical width of a bell surrounding a center.
2. Exponential function
$$f(r) = \exp\left\{-\frac{r}{\beta}\right\}$$
with β having the same meaning.
3. Resonance curve function
$$f(r) = \frac{1}{\beta^2 r^2 + 1}$$
(to be more exact,
$$f(r) = \frac{1}{\beta^{n-1} r^{n-1} + 1},$$
for n-dimensional case) with β having the same meaning.

Of course, there could be other functions meeting the constraints mentioned above.

As soon, as the function $\mathcal{F}(x_1, x_2, x_3, \ldots, x_{n-1}, x_n)$ is developed, one should choose the cut-off (or glue-off) level γ, and finally remove from $\mathfrak{L}$ all the point $\mathfrak{l}^* \in \mathfrak{L}$ so that

$$\mathcal{F}(x_1^{\mathfrak{l}^*}, x_2^{\mathfrak{l}^*}, x_3^{\mathfrak{l}^*}, \ldots, x_{n-1}^{\mathfrak{l}^*}, x_n^{\mathfrak{l}^*}) > \gamma, \tag{4}$$

thus yielding the complement $\widehat{\mathfrak{L}}$.

We started from the case where the set $\mathfrak{M}$ could be almost unambiguously identified (as embedded into Ω), so that no problem takes place with the definition of the set $\mathfrak{L}$. A configuration of $\mathfrak{M}$ meeting this supposition might be met in a number of situations, nonetheless, there could be alternative patterns with fuzzy "border"; here we quote this word, since no one has clear, concise, self-consistent and productive definition of the border, for discrete sets.

Less evident is the situation, if $\mathfrak{M}$ looks like a gradually dispersal set, as one goes outside from the center of that former. Thus, some difficulties in determination of Ω might take place. Here few options could be implemented, to overcome the problem. Firstly, one can follow a standard technique of image filtration [12]. A glance at Fig. 2 allows to see that the effect of contouring is present even for rather compact sets $\mathfrak{M}$, so the filtration should be applied at any chance.

Another option is to develop the area Ω artificially, say, building up the body Ω due to an ellipse of scattering implementation: counting all the eigenvectors of the covariance matrix for the set $\mathfrak{M}$, one can then build up the corresponding (n-dimensional) ellipsoid. By scaling that latter, one can fit the best subset $\mathfrak{M}^* \subset \mathfrak{M}$ completely falling inside the ellipsoid; thus, that latter might represent Ω.

As soon, as the complement is developed, one may treat it with standard and custom techniques, to find out, say, its cluster structure. As one can see from Figs. 3 and 2, the approximating manifold, at this example, is to be a part of plane with two holes inside. In other words, this is must be a manifold of genius two type. The occurrence of two holes could easily be detected with K-means technique applied to the complement $\widehat{\mathfrak{L}}$.

2. Discussion and Conclusion

A new method to discover the features of a low-dimension manifold to approximate multidimensional data is proposed. The method is based on the development of specially constructed the complement of an original dataset, through the implementation of special finite discrete set of randomly dispersed points covering the same area in a space, as the original dataset does.

Namely, the method aims to address the question towards the topology peculiarities of the approximating manifold, since one has no other way to be assigned to the manifold. Apparently, whether an approximation with (say) elastic map would be of a proper quality, strongly depends on the well done choice of a starting manifold to do the approximation.

The proposed technique to reveal the features of topology of an approximating manifold may face some problems resulted from the dimensionality, and dataset structure. For sufficiently high dimensions, one may meet various topological features that may not be visible, for two-dimensional case. In fact, there is no guarantee that the complement $\widehat{\mathfrak{L}}$ will always be simpler, from topological point of view, than the original set $\mathfrak{M}$. Yet, an absolute universality was not the ultimate value here. Still, there are some other ways to improve a situation, if a direct implementation of the technique described above fails to figure out the topology features. A change from Euclidean metrics for Mahalanobis metrics may bring a success here, as well, as an implementation of other metrics into the analysis.

An idea to figure out a border of a complement $\widehat{\mathfrak{L}}$ of $\mathfrak{M}$ seems to be rather close to the ideas of the clusters identification through the local density methods (see, e. g., [13, 14]). Yet, they are not equivalent; meanwhile, further investigations are to be done in order to reveal the closer relations of these two approaches.

References

[1] Leskovec J., Rajaraman A., Ullman J. D. Mining of massive datasets. (2014) — Cambridge Univ. Press, 495 p.

[2] Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya A. Y., Foufou S., Bouras A. (2014) A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Trans. on emerging topics in computing, **2**(3): 267 – 279.

[3] Dongkuan Xu, Yingjie Tian (2015) A Comprehensive Survey of Clustering Algorithms. Ann. Data. Sci., **2**(2): 165 – 193.

[4] Gorban A. N., Kégl B., Wünsch D. III, Zinovyev A. (Eds.), Principal Manifolds for Data Visualisation and Dimension Reduction, LNCSE **58**, Springer, Berlin – Heidelberg – New York, 2007.

[5] Gorban A. N., Roose D. (Eds.) Coping with Complexity: Model Reduction and Data Analysis (Lecture Notes in Computational Science and Engineering, Springer, 2010.

[6] Gorban A. N., Tyukin I. Yu., Prokhorov D. V., Sofeikov K. I. (2016) Approximation with random bases: Pro et Contra. Information Sciences, **364 – 365**: 129 – 145.

[7] Sadovsky M. G., Ostylovsky A. N. (2017) New Clusterization Method Based on Graph Connectivity Search. Journal of Siberian Federal Univ., ser. Math. & Phys, **10**: in press

[8] Comaniciu D., Meer P. (2002) Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Machine Intell., **24**: 603 – 619.

[9] Mirkes E. M., Alexandrakis I., Slater K., Tuli R., Gorban A. N. (2014) Computational diagnosis and risk evaluation for canine lymphoma. Computers in Biology and Medicine, **53**: 279 – 290.

[10] Girvan M., Newman M. E. J. (2002) Community structure in social and biological networks. PNAS, **99**(12): 7821 – 7826.

[11] Akinduko A. A. Mirkes E. M., Gorban A. N. SOM: Stochastic initialization versus principal components. (2016) Information Sciences, **364–365**: 213 – 221.

[12] Lecarme O., Delvare K. (2013) The Book of GIMP: A Complete Guide to Nearly Everything. No Starch Press, 676 pp.

[13] Ester M., Kriegel H.-P., Sander J., Xiaowei X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD-96 Proceedings. (Eds. E. Simoudis, J. Han, and U. Fayyad). Pp.226 – 231.

[14] Sander J., Ester M., Kriegel H.-P., Xiaowei X. (1998) Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Data Mining and Knowledge Discovery, V. **2**(2): 169–194.