Homological assessment of data representations

Serguei Barannikov, Ilya Trofimov, Ekaterina Trimbach, Jun Wang, Evgeny Burnaev

Homological Assessment of Data Representations

Serguei Barannikov[a,b], Ilya Trofimov[a], Ekaterina Trimbach[c], Jun Wang[a], and Evgeny Burnaev[a]

[a] Skolkovo Institute of Science and Technology, Moscow, Russia
[b]CNRS, IMJ, Paris University, France
[c]Moscow Institute of Physics and Technology, Moscow, Russia

ABSTRACT

In this paper* we discuss the concept of the Cross-Barcode(P,Q) introduced and studied in the recent work [1]. In particular, we describe the emergence of this concept from the combinatorics of matrices of the pairwise distances between the two data representations. We also illustrate the applications of the Cross-Barcode(P,Q) to the evaluation of disentanglement in data representations. Experiments are carried out with the dSprites dataset from computer vision.

Keywords: Data representations, data manifolds, persistent homology, disentangled representations.

1. INTRODUCTION

The success of machine learning methods relies to a great extent upon the choice of the features, i.e. the data representations. That is why it is important to develop methods that help to control, visualize and quantify differences in the data representations.

One point of view on the data representations is through the geometric concept of the manifold, based in particular on the manifold hypothesis, which asserts that real-world data are concentrated in a neighborhood of a significantly lower-dimensional manifold, sitting inside a higher dimensional space $\mathbb{R}^D$ [2, 3, 4].

There has been a number of interesting applications of geometrically inspired methods in machine learning [2, 5, 6, 7, 8, 9, 10]. Most of the time these methods start from some construction associated with a given data representation.

In this paper, we discuss the recently introduced concept of Cross-Barcode [1] that associates a robust set of geometrically explainable features in the settings of comparison of two data point clouds. We give an application of this concept to evaluation of disentanglement in data representations.

2. CROSS-BARCODE FOR COMPARING TWO DATA POINT CLOUDS

Let P,Q be two data point clouds. A new algebro-topological tool named Cross-Barcode capturing multiscale discrepancies between two representations P and Q was proposed recently in Ref. [1]. For reader convenience we recall briefly the definition of the Cross-Barcode in this section and refer to loc.cit. for explanations, definitions and proofs.

The Cross-Barcode$_*(P,Q)$ describes specific topological features that distinguish the data representation P from the data representation Q. The Cross-Barcode$_*(P,Q)$ is the set of disjoint intervals, each interval corresponding to a single topological feature. The interval records the "birth" and the "death" scales of the corresponding topological feature.

The definition of the Cross-Barcode is based on the notion of the filtered simplicial complex. The simplicial complex is a combinatorial concept that can be understood as a higher-dimensional generalization of the concept of the graph. Simplicial complex S is a collection of $k-$simplices, $k \geq 0$. The geometric form of the simplex is the higher dimensional generalisation of the geometric forms of the segments ($1-$simplices), triangles ($2-$simplices), solid tetraedres ($3-$simplices) etc. A $k-$simplex is described by the set of its vertices. Each $k-$simplex in a simplicial complex is included together with all lower dimensional simplices described by subsets of the simplex vertices. The vertices set of each $k-$simplex in our case is an arbitrary $(k+1)-$elements subset of the set $P \cup Q$. For an arbitrary simplicial complex, its part consisting of all $0-$ and $1-$simplices is a graph.

Email for correspondence: s.barannikov@skoltech.ru
*This work was supported by Ministry of Science and Higher Education grant No. 075-10-2021-068

Let $C_k(S)$ denotes the vector space whose basis elements are the $k-$simplices of the simplicial complex S. The data representation can be thought of as constructed from such simple $(k+1)-$elements subsets, and their adjacency (inclusions of $k-$element subsets into $(k+1)-$element subsets) is described by the following boundary linear operator on $C_k(S)$. The boundary linear operator $\partial_k : C_k(S) \to C_{k-1}(S)$ is defined on $\sigma = \{x_0, \ldots, x_k\}$ as

$$\partial_k \sigma = \sum_{j=0}^{k} (-1)^j \{x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k\}. \tag{1}$$

The basic and most robust invariants of the simplicial complexes are homology groups. The $k-$th **homology** group $H_k(S)$ is defined as the vector space $\ker \partial_k / \operatorname{im} \partial_{k+1}$. The elements $c \in \ker \partial_k$ are called cycles. The elements of $H_k(S)$ represent $k-$dimensional topological features in S. A basis in $H_k(S)$ represent a set of the basic topological features.

Filtration on a $(k+1)-$element subset of $P \cup Q$ is the diameter of this subset without taking into account the distances between $Q-$points, i.e. this is the maximal distance between pairs of $P-$to$-P$ and the $P-$to$-Q$ points in the $(k+1)-$element subset σ:

$$d_P(\sigma) = \max_{(x,p), x \in \sigma, p \in \sigma \cap P} d(p,x) \tag{2}$$

It is the real number assigned to each simplex of the simplicial complex. In particular the simplices with $d_P(\sigma) \leq \alpha$ form the simplicial subcomplex S_α. And these collections of simplices are nested: for $\alpha_1 < \alpha_2$ all simplices of S_{α_1} are also in S_{α_2}.

The inclusions $S_\alpha \subseteq S_\beta$ induce naturally the maps on the homology groups $H_k(S_\alpha) \to H_k(S_\beta)$. The evolution of the cycles through the nested family of simplicial complexes S_α is described by the barcodes. The persistent homology principal theorem [11] states that for each dimension there exists a choice of a set of basic topological features across all S_α so that each feature appears in $H_k(S_\alpha)$ at specific time $\alpha = b_j$ and disappears at specific time $\alpha = d_j$. The $i-$th Cross-Barcode$_i(P,Q)$ is the record of these times represented as the collection of segments $[b_j, d_j]$.

If the two point clouds coincide, i.e. $P = Q$, then Cross-Barcode$_i(P,Q) = \varnothing$ for all i. Therefore the set of basic topological feature from Cross-Barcode$_*(P,Q)$ is the summary of multiscale discrepancies between the data representations P and Q [1].

An example of the Cross-Barcode$_i(P,Q)$, with $i = 0, 1$ is shown on Fig.1.

2.1 Cross-Barcode$_0(P,Q)$ tracks the scale dependence of $P-$clusters that are remote from Q

The Cross-Barcode$_0(P,Q)$ captures the scale dependence of the $P-$clusters as they join with each other and with $Q-$clusters with increase of the scale α. At the beginning at $\alpha = 0$ each $P-$ cluster corresponds to a single $P-$point and they are all disjoint from Q, assuming that $P \cap Q = \varnothing$. With increase of the threshold α we start adding the edges of length less or equal to α that connect two $P-$points or a $P-$point and a $Q-$point. So that some initial clusters merge between them or with the set of $Q-$points.

The 0th Cross-barcode controls how the $P-$clusters that are disjoint from $Q-$clusters evolve with the increase of the threeshold α.

Let $b_0(S_\alpha)$ be number of bars $[0,d)$ in the Cross-Barcode$_0(P,Q)$ with $\alpha < d$. Then $b_0(S_\alpha)$ is the number of the $P-$clusters that are disjoint from the set Q. Where the points are in the same cluster if the distance between them is less or equal to the threshold α. In other words, the number of bars with "death" bigger than α is the number of $0-$dimensional topological features distinguishing at the threshold α the points P from Q.

It should be stressed that it is impossible in general to know the "death" scale of the cluster born at $\alpha = 0$ at the given point p by looking at the distances from this point p to the different points $p' \in P$ and $q \in Q$ only. This "death" scale depends on the *global* clustering structure of the clouds P and Q.

From another point of view, the zero-dimensional Cross-Barcode$_0(P,Q)$ can be understood as follows. It is natural to start analysing the closeness of the point cloud P to the point cloud Q by looking at the matrix of the $P-Q$ pairwise distances. Namely, if there are many points p_i from P such that their distance to their closest point from Q is relatively big then this implies that the representations P and Q are not close. However one should distinguish the situations when all these remote from Q points p_i are close to each other, and then their remoteness from Q represents the same topological feature,

or when the remote from Q points p_i form several clusters, so that there are several topological features distinguishing P from Q and each such remote from Q cluster represents a separate topological feature. The long bars in the zero dimensional Cross-Barcode$_0(P,Q)$ record precisely the lifespans of these remote from Q clusters of $P-$points .

2.2 Cross-Barcode$_{i\geq 1}(P,Q)$ tracks the scale dependence of the $i-$dimensional topological features distinguishing P from Q

It also happens more often in practice that it is not possible to distinguish a separate cluster of points in P which are all remote from Q. Rather, inside the same cluster of points in P there are some points which are close to Q and other points which are far from Q. This situation is captured and quantified by the higher dimensional topological features distinguishing the representation P from Q. The "birth" and "death" scales of these topological features are recorded as bars in the $i-$dimensional Cross-Barcode$_{i\geq 1}(P,Q)$.

Intuitively, these $i-$dimensional topological features represent an $i-$dimensional subsurface of $P-$points whose boundary is close to $Q-$cloud, but whose interior $P-$points are remote from Q.

2.3 Simplest configurations with a nontrivial bar in Cross-Barcode$_i(P,Q)$

The simplest configuration with a nontrivial bar in Cross-Barcode$_0(P,Q)$ consists of just two points: a $P-$point and a $Q-$point at a non-zero distance $d(p,q)$. The Cross-Barcode$_0(P,Q)$ for such configuration has the unique bar $[0,d(p,q))$.

The simplest configuration of points with a nontrivial bar in Cross-Barcode$_1(P,Q)$ consists of four points: a pair of $P-$points p_1,p_2 and a pair of their closest $Q-$points q_1,q_2 such that the lengths of each of the diagonals $[p_1,q_2]$, $[p_2,q_1]$ are bigger than the lengths of the three segments $[p_1,p_2]$, $[p_1,q_1]$, $[p_2,q_2]$ of the chain connecting q_1 and q_2. The Cross-Barcode$_1(P,Q)$ for such configuration has the unique bar $[b,d)$ with $b=\max\{d(p_1,q_1),d(p_1,p_2),d(p_2,q_2)\}$ and $d=\min\{d(p_1,q_2),d(p_2,q_1)\}$. The nontrivial path connecting points q_1 and q_2 is born at the scale b and this path plus the $Q-$segment q_1q_2 becomes a boundary of a sum of two triangles at the scale d. In a sense the length of lifespan of this feature is the penalty for the following incoherence in the approximation of P cloud by Q cloud when the distance from, say, the point q_1, which approximates the point p_1, to the second $P-$point p_2 is bigger than the distance from p_1 to p_2.

Similarly the simplest configuration with a nontrivial bar in Cross-Barcode$_i(P,Q)$ consists of $2i+2$ points. For $i=2$ for example it consists of two points p_1,p_2 from P, plus two close to p_1 points q_1,q_1', and plus two close to p_2 points q_2,q_2'. Let then $d(p_2,q_1)<d(p_2,q_1')$, i.e. q_1 denotes the point that is closer than q_1' to p_2. And similarly q_2 is closer than q_2' to p_1. Then the "birth" scale of the $2-$dimensional feature corresponds to the maximal length of segments in the set of triangles formed by the two smaller diagonals, the segment p_1p_2, and the four small segments q_1p_1, $q_1'p_1$, q_2p_2, $q_2'p_2$:

$$b=\max\{d(p_1,q_2),d(p_2,q_1),d(p_1,p_2),d(q_1,p_1),d(q_1',p_1),d(q_2,p_2),d(q_2',p_2)\}. \tag{3}$$

The "death" scale of this $2-$dimensional feature corresponds then to the smallest of the two bigger diagonals

$$d=\min\{d(p_1,q_2'),d(p_2,q_1')\}, \tag{4}$$

so that adding this segment permits to represent the set of the previous triangles as the boundary of a union of three $3-$simplices.

3. COMPARISON WITH MINIMUM MATCHING DISTANCE (MMD)

The Minimum Matching Distance (MMD) is the measure of fidelity of the P samples cloud with respect to the Q data cloud. To calculate the MMD one matches every point p of $P-$cloud to its closest point from the $Q-$cloud, i.e. the $Q-$point with the minimum distance from p, and then reports the average of these minimum distances.

Proposition. The average length of intervals from Cross-Barcode$_0(P,Q)$ is bounded from above by the MMD and coincides with MMD in the limit when the distances within the $P-$cloud are much bigger than the distances from P to Q cloud.

Proof. To prove this one can multiply the distances entering the matrix of distances within the $P-$cloud by some sufficiently big constant, and keep the matrix of distances from P to Q the same, so that for every point of P its closest another $P-$point is further than its closest $Q-$point. Then in the process of the above evolution of clusters of $P-$points each initial cluster, corresponding to the given point p, does not interact with other $P-$clusters but joins a $Q-$point at the scale $d(p,Q)$, which

is the minimum distance from p to $Q-$points. Then the average of these "death" scales over all $P-$points coincides with the MMD. When we consider now the matrix of initial distances within $P-$cloud, then some $P-$clusters may have their "death" scale decreased, so that in general the average length of intervals in the Cross-Barcode$_0(P,Q)$ does not exceed the MMD.

4. CROSS-BARCODE$_*(P,Q)$ CAN BE USED TO EVALUATE THE DISENTANGLEMENT OF LATENT DIRECTIONS IN DATA REPRESENTATIONS

We study how the Cross-Barcode can evaluate the disentanglement of data representations. By definition, the latent representation is disentangled, if it has factorized space bijective to interpretable factors of variation. We study the manifold M by slicing it into submanifolds $M_{i,v}$ conditioned by some factor $x_i = v$. We show that the directions x_i corresponding to interpretable factors are salient. The Cross-Barcode$_1(P,Q_{disent})$, where $P \subseteq M$, $Q_{disent} \subseteq M_{i,v}$, has smaller segments, and smaller overall number of segments, when compared to Cross-Barcode$_1(P,Q_{rand})$, where $Q_{rand} \subseteq M$ is the random point subcloud from M of the same cardinality, $|Q_{disent}| = |Q_{random}|$. Thus the slice of the disentangled direction captures better the topology of the whole cloud, as the Cross-Barcode$_1(P,Q)$ that measures the difference between them is smaller in the disentangled case. Below the results of the calculation carried out for the dSprites, disentanglement testing Sprites dataset [12].

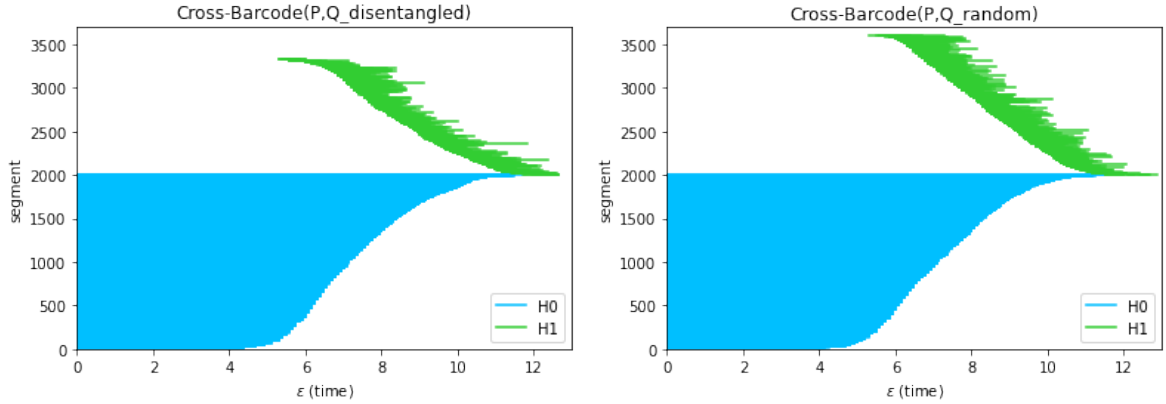

Figure 1: The slice of the disentangled direction captures better the topology of the whole cloud, as the Cross-Barcode$_1(P,Q)$ that measures the difference between them is smaller in the disentangled case. The total length of intervals in Cross-Barcode$_1(P,Q)$ in the disentangled case(left): 530.2 ± 3.4 with the number of intervals: 1329.7 ± 5.1, in the random case(right): 641.6 ± 3.3 with the number of intervals: 1583.1 ± 5.7

5. CONCLUSIONS

We have discussed the concept of the Cross-Barcode [1] that allows to control, visualize and quantify the differences in two data representations. We have explained this concept from the point of view of combinatorics of the set of simplices formed by the points in data representations. The discussed methodology is applicable in many different settings and permits in particular to construct a set of robust features that tracks the discrepancies in data representations. The application for the evaluation of disentanglement of latent directions in data representations is presented.

REFERENCES

[1] Barannikov, S., Trofimov, I., Sotnikov, G., Trimbach, E., Korotin, A., Filippov, A., and Burnaev, E., "Manifold topology divergence: a framework for comparing data manifolds," in [*35th Conference on Neural Information Processing Systems (NeurIPS 2021)*], (2021).

[2] Belkin, M. and Niyogi, P., "Laplacian eigenmaps and spectral techniques for embedding and clustering," in [*Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*], 585–591 (2001).

[3] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y., [*Deep learning*], vol. 1, MIT press Cambridge (2016).

[4] Bengio, Y., Courville, A., and Vincent, P., "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence* **35**(8), 1798–1828 (2013).

[5] Vincent, P. and Bengio, Y., "Manifold parzen windows," in [*Proceedings of the 15th International Conference on Neural Information Processing Systems*], 849–856 (2002).

[6] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P., "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Processing Magazine* **34**(4), 18–42 (2017).

[7] Moor, M., Horn, M., Rieck, B., and Borgwardt, K., "Topological autoencoders," in [*International Conference on Machine Learning*], 7045–7054, PMLR (2020).

[8] Bunrit, S., Kerdprasop, N., and Kerdprasop, K., "Improving the representation of cnn based features by autoencoder for a task of construction material image classification," *Journal of Advances in Information Technology Vol* **11**(4) (2020).

[9] Kim, K., Kim, J., Zaheer, M., Kim, J., Chazal, F., and Wasserman, L., "PLLay: Efficient topological layer based on persistence landscapes," in [*34th Conference on Neural Information Processing Systems (NeurIPS 2020)*], (2020).

[10] Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P., "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478* (2021).

[11] Barannikov, S., "Framed Morse complexes and its invariants," *Adv. Soviet Math.* **21**, 93–115 (1994).

[12] Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A., "DSprites: Disentanglement testing Sprites dataset." https://github.com/deepmind/dsprites-dataset/ (2017).