Non-injectivity Approach to Causal Inference

Marek Kaluba, Jonas Peters …

Wednesday 21st June, 2017

Abstract

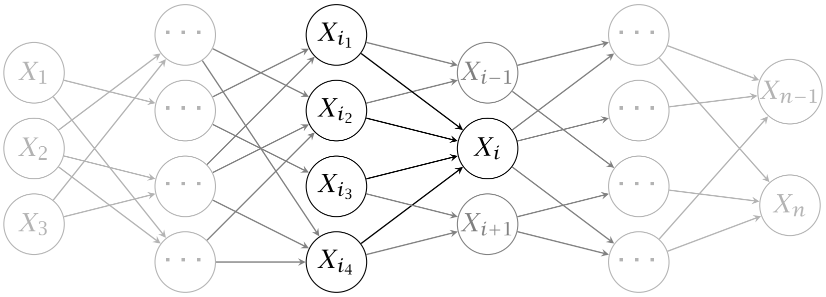
We propose a new method for causal inference based on homological description of the shape constructed from a sample from the joint distribution.

1 Introduction

The fundamental difference between correlation and causation is hard to identify given only a finite sample. Although performing an intervention (controlled randomised experiment) can explain the difference, usually it is expensive, time and labour consuming. Therefore there is a need for context agnostic methods that will allow to draw conclusions on the causal link.

missing refs

In the literature there can be found several algorithms for detection of the underlying causal link using statistical analysis of the data **??**. Usually it is assumed that the underlying causal structure takes a form of a directed acyclic graph (DAG), where arrows are interpreted as dependencies. The value of a node (variable) X_i depends on all nodes (parents) X_{i_k} such that there exists a directed edge from X_{i_k} to X_i.



Such DAGs are classified up to Markov equivalence which, however, does not include the directional information on the dependence between variables. Also addition of noise variables to the system which jointly independently influence all the nodes is a common assumption which brings the model closer to applications, where noise is a prevailing.

As DAG formalism provides only qualitative description of the system of dependencies, one may try to asses the systems quantitative properties in the form of partially ordered set of equations (variable assignments). These equations may be used to model *functional* relationships between variables. The structural equation model (SEM) is a poset of equations and encodes more information than the DAG. Usually SEM is denoted as

$$\{X_i = f_i(\mathbf{pa}(X_i), N_i)\}_i$$

where i runs over all possible nodes, $\mathbf{pa}(X_i) = \{X_{i_k}\}$ denote the set of parents of the vertex X_i in the DAG and N_i are jointly independent noise variables with possibly different distributions for each i.

The standard *functional* approach is to limit the function class of possible functions orientating the edges. Then a hypothesis on the direction can be statistically tested and, if certain confidence level is obtained, a decision on the direction of the arrow is drawn. The addition of the noise variables N_i to each node is natural from the point of view of applications (e.g. measurement noise) and significantly helps in the step.

ref

Examples of such methods include additive noise model (ANM), in which one assumes that nodes are sums of functions of their parents

$$\left\{ x_i = \sum_{j=1}^{i_k} f_j(x_{i_j}) + n_i \right\}$$

ref

and linear non-gaussian additive model (LINGAM) which assumes that all the functions f_k are linear and n_i is not drawn from from a Gaussian distribution. Other methods rely on the postulate of independence of cause and mechanism and use independence of residuals , information contained in the residuals or other quantities that may be explained by the asymmetry between causes and effects.

ref

ref

In this paper we propose a not-standard way of estimating the causal graph structure without relying on regressing the functional dependence. We use the Delaunay or, in higher dimensions, the Vietoris-Rips simplicial complex to approximate the *graph* of the function. Then we create different filtrations by projecting the complex on different axes and obtain persistence homology diagrams of each. We combine these diagrams into a single confidence score which is used to infer the orientation of the arrows.

1.1 Motivation

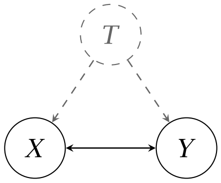
Suppose that a sample drawn form the joint distribution following density $p(X_1, \ldots, X_k)$ of the random variables $\{X_i\}_{i=1}^k$ is given. The task is to recover a DAG G consistent with $p(X_1, \ldots, X_k)$, i.e. a graph such that repeated sampling from each node X_i converges to the initial joint distribution. We make use of the usual assumptions in causal inference, i.e. the underlying causal stucture in the form of a DAG, jointly independent noise and continuity of variables. Moreover we introduce the following **unimodality assumption** as it is fundamental for the method described.

> Each probability density p_{X_i} and $p(X_i|\mathbf{pa}(X_i))$ has one maximum, i.e. X_i and the "visible noise" are uni-modally distributed random variables.

In the discussion of the assumption we will restrict to a simple case of two variables X and Y and the task of inferring the causal relationship between them.

Bivariate case The motivation for this method is based on the following simple observation. Assume for now that p_X is uniform over an interval and X and Y are 1-dimensional. Suppose that Y is a stochastic function of X, and the *trend* function is regular in some sense (e.g. smooth). Under the unimodality assumption, a sufficiently dense sample from the joint distribution would have the *large scale geometry* (e.g. its shape observed from very far away) of the graph of the trend function. This is especially visible in the case of data generated by LINGAM or ANM models: a scatter plot of sampled points $\{(x, f(x) + n)\}$ will approximate the graph of the function f, with error determined by the noise.

In particular, if we can learn that the shape of the empirical density $\{(x_i, y_i)\}$ is close to the graph of a non-injective function $y_i = f(x_i)$, this excludes the causal direction $Y \to X$. We stress again that the unimodality assumption is crucial here: the simple case of binomial noise (or a continuous version of it) invalidates the whole reasoning. Moreover, if we assume that there is always a causal relationship between X and Y this allows us to conclude that the remaining possibility one is the true causal structure. Note that this assumption is limiting and even may be misleading, since a true causal connection between X and Y can be a non-direct one, i.e., there may be an unobserved confounding variable(s) Z, driving patterns of X an Y simultaneously.



To asses the non-injectivity of a function without regressing the function itself we develop *Topological Injectivity Test* (TIT) using computational geometry and algebraic topology tool. We will use simplicial complexes to approximate the graph of a function and introduce filtrations and compute the complex's persistence homology to produce a non-injectivity score.

Multivariate case In high-dimensional setting we provide a similar framework. Given $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_k)$ we can embedd the empirical distribution into $\mathbb{R}^{n+k}$ and ponder the question of non-injectivity. Note however that in this setting it is possible that some *directions* are injective, whereas others are not.

Example. Consider set of points in $\mathbb{R}^4$ generated according to SEM

$$\begin{cases} X_1 &= U(-1,1), \\ X_2 &= N(0,1), \\ X_3 &= X_2^3 + N(0,1), \\ X_4 &= X_1^2 X_2 + N(0,1). \end{cases}$$

We may evaluate “non-injectivity” by looking at different projections. It is clear that when looking at triples (X_1, X_2, X_3) we will not be able to apply the argument above, however triple (X_1, X_2, X_4) reveals that there is a cause of X_4 in the *set* $\{X_1, X_2\}$. Further projections on (X_2, X_4) and (X_1, X_4) reveal that it is X_1 responsible for the non-injvectivness of X_4.

The paper is organised as follows. In the next section we provide a very quick informal introduction to computational algebraic topology. We recall basic concepts and provide related examples to ease the effort of the reader. A still informal, yet more rigorous overview of the topic can be found in Appendix A. Next we prove basic facts about expected behaviour of topology of simplicial complexes which serve as a backbone for further experiments with data in Section 4. We use both simulated and real world data from . Other existing methods are quoted and comparisons are made in Section 4.3.

pain?

@@@REF@@@ Cause-Effect-Pairs, CE-Benchmarks

2 Algebraic topology

2.1 Informal exposition

We will only treat triangulated shapes, i.e. objects given in the form of a triangular mesh in $\mathbb{R}^d$. Usually it is enough to specify a set $S \subset \mathbb{R}^d$ of vertices of the mesh and a list of maximal simplices (set of subsets of S). E.g., to specify a triangulated *surface* in $\mathbb{R}^3$ it is enough to specify vertices of the mesh and list of triples (which correspond to triangles) of vertices. The union of all vertices, edges and triangles we call a **triangulation**. Triangulations and graphs (or their generalisation – simplicial complexes, see Appendix A) will be objects of primary interest throughout the paper (we denote them by $\mathcal{X}$).

Commonly used in computational geometry is the Delaunay triangulation on a set of points in the plane. It may be constructed by first finding the Voronoi tessellation corresponding to the points, and placing an edge between points whose Voronoi cells share an edge, see Figure 1 for an example. These triangulations exists in any dimension, however they are computationally feasible only for $d = 2, 3$.

Given such regularly structured objects one can try to assign some well behaved *fingerprint*, or *signature* in the terms of an algebraic object. Homology of a triangulated shape $\mathcal{X}$ is a set of descriptors meant to capture qualitative information about topology of the object. To keep close to the geometric intuition we note that 0-homology encodes information on *connected components* of $\mathcal{X}$. Similarly if $\mathcal{X}$ is a graph, 1-homology of $\mathcal{X}$ is spanned by *linearly independent cycles* of $\mathcal{X}$. The void in the 2-sphere (approximated e.g. as triangulated icosahedron) results in non-zero 2-homology and similarly higher homology captures higher-dimensional features. These descriptors do not depend on the particular choice of triangulation (or even tessellation into non-triangular objects) as long as certain regularity conditions are satisfied.

While homology gives some information about the topology of an object, this is usually very hard to interpret without additional information. Typically one resolves to look at *persistent homology* which contains information on the evolution of a

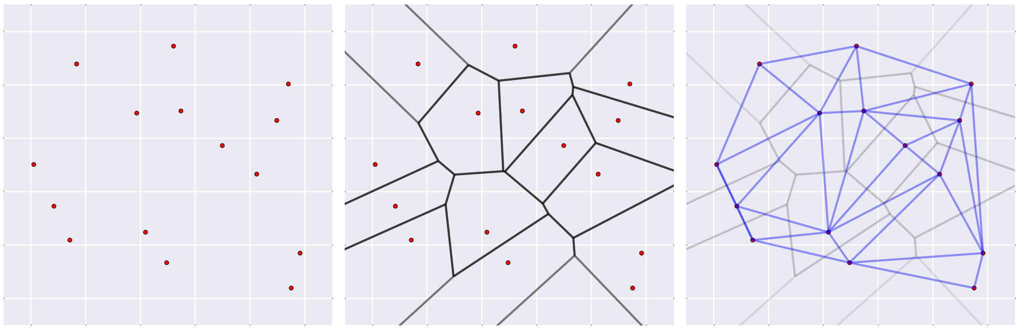


Figure 1: Voronoi tesselation of the plane and the associated Delaunay Triangulation.

geometric shape. Imagine a situation where the geometric shape undergoes a change over time. We would like to capture information about geometrical features that persists for a long periods of time as this brings some robustness. We might be interested in the time needed for two components to merge, or the time required to close a cycle. The underlying intuitive notion of time-based evolution is made rigorous by the concept of filtration.

A filtration on $\mathcal{X}$, is a stratification of $\mathcal{X}$ into increasing sequence of sets with well defined "steps":

$$\emptyset = \mathcal{X}_{-1} \subseteq \mathcal{X}_0 \subseteq \mathcal{X}_1 \subseteq \cdots \subseteq \mathcal{X}_n \subseteq \cdots \subseteq \mathcal{X}_\infty = \mathcal{X}.$$

A convenient way to describe a filtration is by specifying a non-negative, continuous function $f\colon \mathcal{X} \to \mathbb{R}$ and setting

$$\mathcal{X}_t = \{x \in \mathcal{X}\colon f(x) \leqslant t\}.$$

For example, the length of edges can be considered as a filtering function on a graph. This filtration arises in the construction of ε-neighbouring graphs for an increasing sequence of thresholds. Instead of choosing an arbitrary threshold for the distance it might be more useful to observe how geometry of the graph changes as we incorporate more and more edges. *Persistent homology* is the rigorous way to capture these changes. We postpone the definition till the next paragraph, and analyse the following example.

Example. Consider a sample $\{X_i = Z_i + N\}$, where Z_i is drawn from uniform distribution on a circle in $\mathbb{R}^2$ and N follows a bivariate normal distribution (see Figure 2, top-left). We build the full Delaunay traingulation $\mathcal{X}$ on the set $\{X_i\}$ and filter edges (and triangles) by distance. The filtration comes from limiting the length of edges we accept. That is we have a function $f\colon \mathcal{X} \to \mathbb{R}$ which assigns to every vertex 0, to every edge its length and to every triangle the maximal length of its sides, see Figure 2 for a few stages of the filtration. A notable feature of the filtration is that the centre of the circle is surrounded by an empty cycle very early and persists for a relatively long time, until it is finally filled by triangles. While reconstruction of the original density

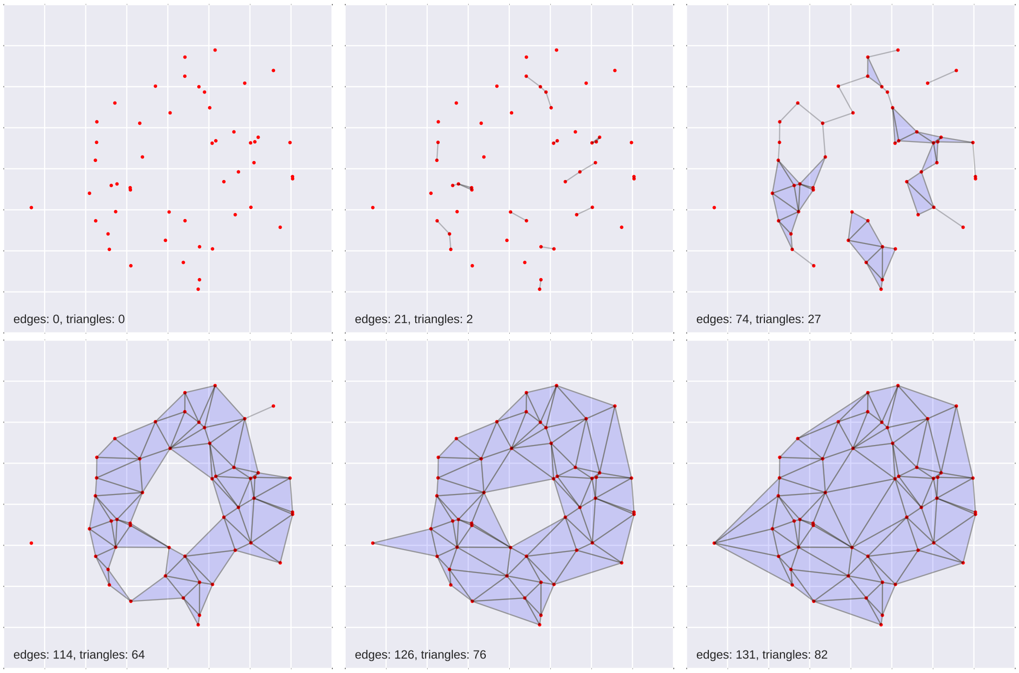

edges: 0, triangles: 0
edges: 21, triangles: 2
edges: 74, triangles: 27
edges: 114, triangles: 64
edges: 126, triangles: 76
edges: 131, triangles: 82

Figure 2: The length-filtered Delaunay triangulation on $n = 50$ of points sampled uniformly from a circle of radius $R = 1000$ with the bivariate Gaussian noise ($\mu = (0,0)$, $\sigma = (300,300)$). For $d = 2,3$ there exist fast algorithms generating the triangulation. In higher dimensions, however, it becomes computationally expensive (complexity scales as $O(n^{\lceil d/2 \rceil} + n \log n)$ [**?** , Corollary 17.3.2].

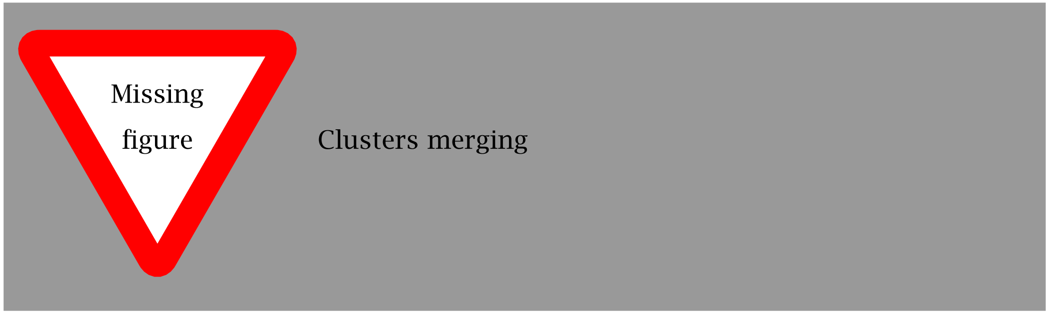
would be impossible based purely on this homological information, the existence of the cycle indicates that the underlying density is of "circular shape", i.e. that there is a low density region surrounded by high density one(s).

Persistence The changes in topology of a space are best spoken of using the language of filtering functions which we describe in vusual terms in Section 2.2. Below, we provide some motivation behind the idea of persistence.

Suppose that we have a space $\mathcal{X}$ filtered by the sub-level sets

$$\mathcal{X}_t = \{x \in \mathcal{X} \colon f(x) \leqslant t\}.$$

A 0-dimensional feature emerges at time t_b (*birth time*) when a new cluster of points is added to X_{t_b} that was not present for $t < t_b$. Similarly when the cluster merges (i.e. becomes connected by an edge) with an older cluster (because the shape has grown) we say that the 0-feature has died and record the minimal time t_d (*death time*) with this property. Obviously $t_d \geqslant t_b$ and we add point (t_b, t_d) to the 0-diagram of $\mathcal{X}$. 0-dimensional features of $\mathcal{X}$ at time t are (in this case) the connected components of sub-level set. The longer the feature lives, the more prominent it seems to us, and the further is the corresponding point from the diagonal. If the feature persists to the last stage of the filtration we say that it is of infinite life. Note that the 0-th persistence diagram will contain an infinite life point for every connected component of the final shape $\mathcal{X}_\infty = \mathcal{X}$.



Clusters merging

Formally speaking, a k-persistence diagram, denoted by $D_k(\mathcal{X}, f)$ is a multiset of pairs $\{(t_b, t_d) \in \mathbb{R}^2 \colon t_b \leqslant t_d\}$ recording the evolution of k-dimensional features. For technical reasons we also include points on the diagonal with infinite multiplicity. These diagrams provide a natural way of comparing two filtered triangulations (or simplicial complexes). We have a natural notion of distance between the diagrams and, in certain cases, robustness to small alterations (see Theorem 5).

Definition 1 (Bottleneck distance)**.** Let D and D' be two persistence diagrams. The bottleneck distance (or matching distance) between them is defined as

$$d_\infty(D, D') = \inf_f \sup_{p \in D} \|p - f(p)\|_\infty,$$

where supremum is taken over all bijections $f \colon D \to D'$ and $\| \cdot \|_\infty$ denotes the standard supremum norm. Note that since we included all points along the diagonal with infinite multiplicity, matching points to the diagonal is still a bijection.

2.2 Theory behind the method

Suppose that $F = (f_1, \ldots, f_\ell) \colon M \to \mathbb{R}^\ell$ is a smooth (at least C^2) function defined on a manifold (possibly with boundary) $M \subset \mathbb{R}^k$. Set $d = k + \ell$ and consider the graph of F, i.e. the manifold defined as

$$G(F) = \left\{ (x, F(x)) \in \mathbb{R}^d \colon x \in M \right\},$$

filtered by the projection on the j-th coordinate ($j > k$), that is the filtering function $\pi_j \circ F = f_j$. It will follow that if f_j is an injective function, then all points in a persistence diagram (of any dimension) of $G(F)$ (filtered by f_j) belong to the diagonal.

In practice we usually do not have access to the whole manifold $G(F)$, hence we will approximate it by a simplicial complex of identical homology properties (is homotopy equivalent). Will show that one can reconstruct the approximating complex and its presistence diagrams from a finite, noisy sample drawn from $G(F)$. Therefore, later on,

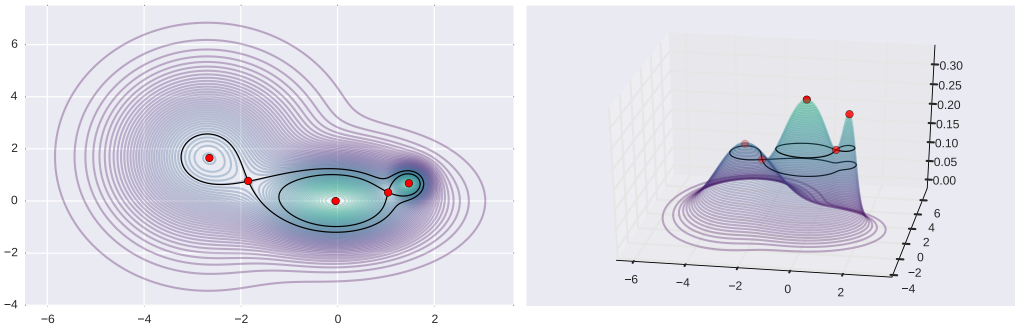


Figure 3: A graph of the probability density of a mixture of Gaussians. In this case $M = \mathbb{R}^2$, $k = 1$ and $F(x, y) = p(x, y)$. The filtering function f is defined as the projection to the z-axis, i.e. $f(x, y, p(x, y)) = p(x, y)$. The sub-level sets are given as $X_t = \{(x, y, p(x, y)) \colon p(x, y) \leqslant t\}$. Critical points of f are points where tangent plane to the surface is parallel to the XY-plane. As we increase t the qualitative changes of the sub-level sets X_t are captured by presistent homology.

we can conclude that if a persistence diagram of the approximating complex contains points *far from the diagonal* this can be attributed to non-injectivity of the function f_j rather than the noise or the manifold M itself.

Watch out: homology of M may be non-trivial!

But I believe it doesn't matter as we filter by function which collapses all cycles in M

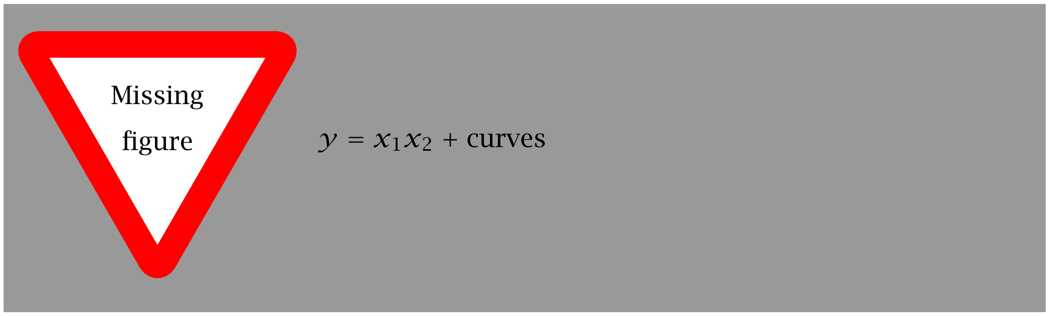
Morse functions and filtrations In the case where M is a manifold, the geometric features that took part in the definition of persistence homology can be characterised using notion of Morse functions. An example is presented in Figure 3.

We say that a C^2 function on a manifold $f \colon M \to \mathbb{R}$ is *Morse* if the gradient $\mathrm{D}f$ has only isolated zeros (so called critical points of f) and at those points the Hessian matrix $\mathrm{D}^2 f$ is non-degenerate. We note that Morse functions are dense in the set of all functions on M in C^m-topology for all $m \geq 2$. It is a standard fact that critical points of a Morse function correspond bijectively to changes of the topology of the sub-level sets and those in turn are reflected by changes in homology.

Recall that we would like to find the "non-injective parts" of $G(F)$ to exclude possible cause-effect relations. In the bivariate case we have to find out if the relation between X and Y follows an injective function $F \colon \mathbb{R} \to \mathbb{R}$. Obviously it requires only on a part of the domain of Y to have multiple corresponding values of X, to claim that X is not the value of a continuous function of Y. Note that we use the continuity of X and Y as well as the unimodality assumption here. This obvious statement, however, does not generalise so easily to multivariate setting.

x_i denotes i-th coordinate of a vector. Is it an acceptable notation??

Consider an example of $M = \{(x_1, x_2)\} \subset \mathbb{R}^2$ and $y = F(x_1, x_2) = x_1 x_2$. When restricting to $x_1 = a$, $F|_{x_1 = a} = a x_2$ is injective for any a. An analogous statement holds also for restriction to $x_2 = a$. Nevertheless, inspection of y's along line $x_1 + x_2 = 0$ reveals that y is a function of both x_1 and x_2. See also Section 1.1 for a different example.



$y = x_1 x_2$ + curves

In the general case we consider a manifold $M = \{(x_1, \dots, x_k)\} \subset \mathbb{R}^k$ and

$$(y_1, \dots, y_\ell) = (f_1(x_1, \dots, x_k), \dots, f_\ell(x_1, \dots, x_k)) = F(x_1, \dots, x_k) \in \mathbb{R}^\ell.$$

Motivated by the example above we ask if there exists an injective functional combination γ of x_i's such that $y_j = f_j(\gamma(x_1, \dots, x_k))$ is not injective. By injective functional combination of x_i's we mean a smooth, injective function $\gamma\colon (-1,1) \to M \subset \mathbb{R}^k$. In the example above $\gamma\colon (-1,1) \to \mathbb{R}^2$ is given as $\gamma(t) = (t, -t)$. We will say that $f\colon G(F) \to \mathbb{R}$ is injective with respect to $\gamma\colon (-1,1) \to M$ if $f \circ F(\gamma)\colon (-1,1) \to \mathbb{R}$ is injective.

Note that knowledge about *some* of the non-injectivities (eg. when restricted to a subset of variables or injective functional combination of them) is enough to identify a *set* of variables containing potential causes. The following lemma states that the existence of non-trivial points in a persistence diagram of $G(F)$ implies non-injectivity of some of f_j with the respect to some combination of its arguments. We believe that the result is not new, yet we were unable to find a specific reference hence we provide a proof.

Lemma 1. *Fix j and suppose that f_j is injective with the respect to any injective curve $\gamma\colon (-1,1) \to M$. Moreover assume that $\pi_j\colon G(F) \to \mathbb{R}$ for some j is a Morse function. Then $\pi_j(x, F(x)) = f_j(x)$ has no critical points and hence all persistence diagrams in any dimension of $G(F)$ filtered by π_j are empty (contain no points away from the diagonal).*

Proof. Suppose the contrary that f_j is injective with respect to all injective γ's, but there is a point (t_b, t_d), $t_b < t_d$ in m-th persistence diagram of $(G(F), \pi_j)$. This means that a cycle in homology of $G(F)_{t_b}$ is present that did not exist for $t < t_b$. By Morse theory this means that there is a critical point $y_b = (x_b, f(x_b))$ of index m in $\pi_j^{-1}(t_b)$. Again Morse theory tells us that there exists a local chart (in particualr a continuous bijection) $\psi\colon \mathbb{R}^k \to M$ around x_b such that $\pi_j(x, F(x)) = f_j(x)$ can be expressed as

$$\pi_j\Big(\psi(z_1, \dots, z_k), F\left(\psi(z_1, \dots, z_k)\right)\Big) = f_j(x_b) - \sum_{i=1}^{m} z_i^2 + \sum_{i=m+1}^{d} z_i^2.$$

If we set $\gamma(t) = (t, 0, \dots, 0) \in \mathbb{R}^k$ for $t \in (-1,1)$, then the composition $f_j \circ \psi \circ \gamma(t) = f_j(x_b) - t^2$ is clearly non-injective. As the curve $\psi \circ \gamma$ is injective this contradicts the initial assumption and finishes the proof. □

Even if $\pi_j(x, F(x)) = f_j(x)$ does not satisfy Morse conditions, one can find a Morse function $\hat{f}_j\colon G(F) \to \mathbb{R}$ such that $\hat{f}_j$ and f_j are close in the C^k-norm. Note that we can estimate the persistence diagram of f_j by using $\hat{f}_j$ instead, with precision related to ε by Theorem 5.

Lemma 2. *Adjust f_j to be a Morse function $\hat{f}_j$ such that $\|\hat{f}_j - f_j\|_{C^m} < \varepsilon$. If F is injective with the respect to any (injective) curve γ, all points in persistence diagrams of $G(F)$ filtered by $\hat{f}_j$ have lifespan shorter than 2ε.*

Proof. Suppose that an m-th persistence diagram of the function $\hat{f}_j\colon G(F) \to \mathbb{R}$ contains a non-trivial point (t_b, t_d). Let y_b and y_d be the critical points responsible for the point. By Morse theory there exists a flowline $\gamma\colon [-1, 1] \to G(f)$ of negative gradient flow (of $\hat{f}_j$) connecting y_d and y_b. Moreover, while $\hat{f}_j$ is strictly decreasing when restricted to the image of γ, $f_j \circ \gamma$ is increasing. Therefore

there seem to be unexplained assumptions

$$
\begin{aligned}
0 < t_d - t_b = \hat{f}_j(x_d) - \hat{f}_j(x_b) \leqslant \\
\hat{f}_j(x_b) - f_j(x_d) + (f_j(x_b) - \hat{f}_j(x_d)) + (f_j(x_d) - f_j(x_b)) \leqslant \\
2\varepsilon + \underbrace{f_j(\gamma(-1)) - f_j(\gamma(1))}_{<0} \leqslant 2\varepsilon
\end{aligned}
$$

as required. □

Approximation of manifolds by simplicial complexes It was shown by de Rham in [] that the homology of a manifold can be approximated by the homology of the Čech complex (the nerve) of a sufficiently fine, well behaved[1] cover of the manifold. Suppose that a sample $S \subset R^n$ of points sampled with noise from $G(F)$ is given. Of course points in S will not necessarily belong to M, but their expected distance is bounded by the noise. We investigate under which conditions we can use Vietoris-Rips complex build on S to approximate the homology of $G(F)$.

@@@REF: grab the reference from JANKO LATSCHEV, Vietoris-Rips complexes of metric spaces near a closed Riemannian manifold @@@

Definition 2 (ε-approximation)**.** A finite set $K \subset \mathbb{R}^n$ is a uniform ε-approximation of M if the Hausdorff distance $d_H(K, M)$ is smaller than $\varepsilon < r(M)$, where $r(M)$ is reach of M.

For a proper definition of reach (using distance to medial axis) please refer to . Intuitively reach of M can be described as the maximal radius r such that $(r - \varepsilon)$-thickening of M in the normal direction has the same topology as M itself.

@@@REF: Chazal, Lieutier: Smooth manifold reconstruction from noisy and non-uniform approximation...

Theorem 3. *Fix $\varepsilon < 1/8$. Let M be a manifold and let K be a uniform ε-approximation of M. For a radius $\alpha \in [\frac{7}{2}\varepsilon r(M), 1 - \frac{9}{2}\varepsilon r(M)]$, the union of balls $\bigcup_{x \in K} B(x, \alpha)$ deformation retracts on M, hence the Čech complex of $\bigcup_{x \in K} B(x, \alpha)$ is homotopy equivalent to M.*

@@@REF: Chazal, Lieutier: Smooth manifold reconstruction from noisy and non-uniform approximation...
also: Niyogi, Smale, Weinberger: Finding the homology of submanifolds with high confidence...

[1] i.e. all sets in the cover have small diameter, their pairwise intersections are contractible

Assume that the variance of the sampling noise is not too large, e.g. locally always smaller than the one eight of the reach of M. Given a finite sample F from a manifold M, we can approximate the homology of the manifold *thickened by noise* by performing Čech construction (the nerve of the cover of F by open balls). This is however unsuitable for computation, so we 'sandwich' the complex with Vietoris-Rips complexes by the virtue of the following theorem.

make this precise, ala [@@@de Silva, Ghrist, Coverage in sensor networks via persistent homology]

Theorem 4. *Let VR_ϵ denote the ε-Vietoris-Rips complex of K. ε'-Čech complex can be sandwiched between Vietoris-Rips complexes*

$$VR(\varepsilon) \subset \check{C}(\varepsilon') \subset VR(\sqrt{2}\varepsilon) \tag{1}$$

for all $\varepsilon \leqslant \varepsilon' \leqslant \sqrt{2}\varepsilon$.

Note that the inclusions above need not to be a homotopy equivalence for any ε', however if the complexes on the both sides of (1) are homotopy equivalent, they capture perfectly the homology of the Čech complex sandwiched in the middle. To achieve this one would have to assume $\varepsilon < 1/10$.

The summary of this section looks as follows. The Vietoris-Rips complex approximates topology of Čech complex which is the nerve of the open-ball covering of a noisy sample F. If the noise variance is bounded globally by reach of M (or locally by distance to medial axis), then the Čech complex captures the topological features of M. Therefore to estimate homology of M it is enough to estimate filtering functions defined on Vietoris-Rips complexes.

Estimation of persistence diagrams As we do not have access to manifold due to noise in sampling, from now on we will assume that the approximating Vietoris-Rips complex $\mathcal{X}$ build on the points in F is the ideal object which persistent homology we would like to analyse. We introduce new filtering functions on the complex that are suited for detection of non-injectivity. The persistence diagrams of these filtrations will serve as a description of shape of F despite sampling noise, due to result of Cohen-Steiner, Edelsbrunner and Harer.

@@@REF Cohen-Steiner, Edelbrunner, Harer: Stability of persistence diagrams@@@

Theorem 5. *Let M be a manifold and let $f, g\colon M \to \mathbb{R}$ be tame (e.g. smooth) functions. Then for any dimension k*

$$d_\infty\left(D_k(f), D_k(g)\right) \leqslant \|f - g\|_\infty,$$

where $\|\cdot\|_\infty$ denotes the sup*-norm over functions.*

Suppose that $\mathcal{X} \subset \mathbb{R}^2$. We endow it with four different filtrations. Let σ be a simplex in $\mathcal{X}$. We define the filtering functions as follows.

$$\begin{aligned} f_i^{\nearrow}(\sigma) &= \max_{x\in\sigma}\{\pi_i(x)\} \\ f_i^{\searrow}(\sigma) &= \max_{x\in\sigma}\{-\pi_i(x)\}, \end{aligned}$$

where $\pi_i(x)$ denote projection of x on i-th axis. Note that it is enough to evaluate π_i on vertices. We will denote filtrations induced by these functions as $\mathcal{X}_i^{\nearrow}$ and $\mathcal{X}_i^{\searrow}$ and refer as *increasing* and *decreasing*, respectively.

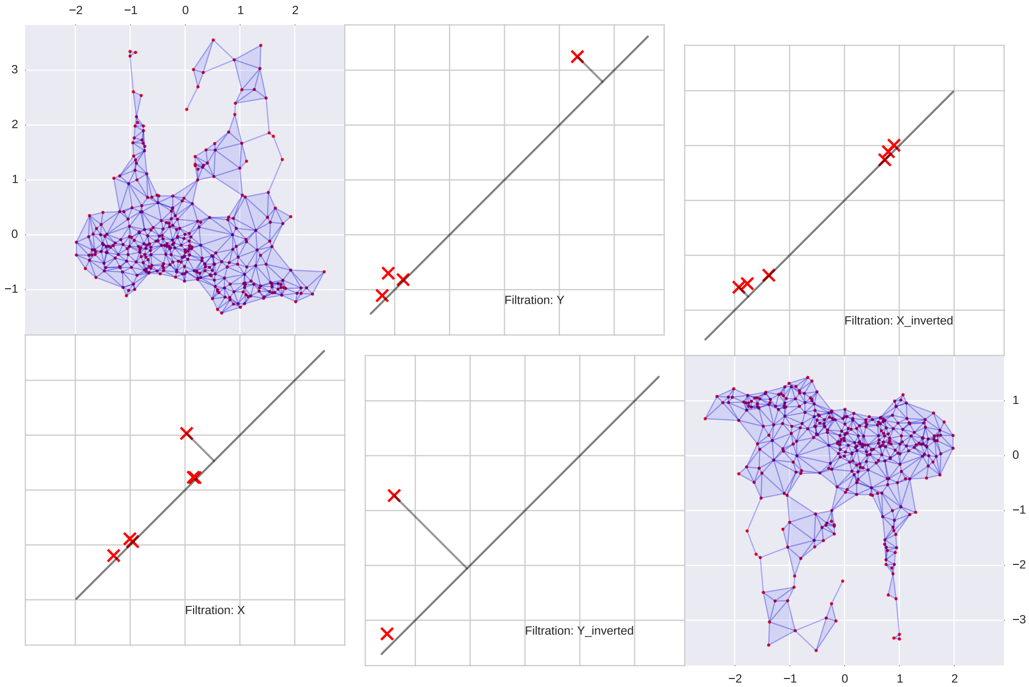


Figure 4: Persistence diagrams for the four filtrations on pair0021.txt from the CEP dataset. Original points (upper-left corner) have been inverted (lower-right corner) to make identification of the corresponding points in the persistence diagrams easier. The longest distance of a point in the diagram D to the diagonal is the bottleneck distance $d_\infty(D, D(\emptyset))$.

3 Topological Injectivity Test

In the paragraph below we provide quick overview of Topological Injectivity Test (TIT). Again, we restrict to the bivariate case and we will treat the first coordinate as values of random variable X and the second as Y. Suppose that a finite sample $S = \{(x, y)_i\} \subset \mathbb{R}^2$ is given. We build a geometric simplicial complex $\mathcal{X}$ on the sample and introduce four filtrations (on the same complex) by the filtering functions which are given by increasing or decreasing projections on different axes.

If S was sampled (without noise) from a graph of an injective function, the approximating complex would have similar shape. In particular a graph of an injective function (filtered by projections) has empty persistence diagrams, except one point of infinite life which corresponds to the fact that plot of a continuous function on a connected domain is connected itself. We forget the point of infinite life (e.g. using notion of *reduced homology*) and measure the distance between persistence diagrams of different filtrations on $\mathcal{X}$ with an empty diagram to produce a *non-injectivity score*.

Details of the TIT Suppose that a finite sample sample $S = \{(x, y)_i\} \subset \mathbb{R}^2$ is given. We create $\mathcal{X}$, the minimal connected subcomplex of the Delaunay complex on the points in S. We do it by first constructing the full Delaunay triangulation and then we continue to remove the longest edges (and triangles) as long as the comples is still connected. A similar construction for Vietoris-Rips complex is possible as well if $F \subset R^n$ for $n \geqslant 4$. Suppose that the multivariate random variables X and Y have dimension d_X and d_Y respectively. It is worth to limit the construction of $\mathcal{X}$ only to $\max(d_X, d_Y)$-dimensional simplices for computational reasons (we do not expect interesting homology in dimensions higher than the dimension of the domain).

As the score for non-injectivity hypothesis we propose

$$\begin{aligned} h_{X\to Y} &= \max\left\{d_\infty(D_0(\mathcal{X}_2^{\nearrow}), D_0(\varnothing)), d_\infty(D_0(\mathcal{X}_2^{\searrow}), D_0(\varnothing))\right\} \\ h_{Y\to X} &= \max\left\{d_\infty(D_0(\mathcal{X}_1^{\nearrow}), D_0(\varnothing)), d_\infty(D_0(\mathcal{X}_1^{\searrow}), D_0(\varnothing))\right\}. \end{aligned}$$

In the multivariate case we take maximum also over all projections (in the ranges of X or Y) as well as over higher dimensional diagrams. Note that this is a simplified version, since we could be looking at projections on any (injective) functional combination of variables (see Section 1).

Dealing with outliers The above method is heavily influenced by outliers in the data. Indeed, a single outlier can force us to accept very long edges (or triangles) in the complex $\mathcal{X}$ which will engulf the features. To remedy this we propose the following procedure inspired by persistence.

Instead of choosing a fixed number of points to be removed as outliers we remove them one-by-one, analysing the geometry of the arising complex each time and producing non-injectivity scores. We sort the points in S according to distance to their k-nearest neighbours. We used $k = \lceil 0.02n \rceil + 5$ for the experiments. Then the complex $\mathcal{X}(i)$ is constructed as described above using all point from S except last i (these are most significant outliers). To add more stability to scores we standardise (that is apply an affine map which brings mean to 0 and standard deviation to 1) data prior to each removal.

This produces real-valued functions $h_{X\to Y}(i), h_{Y\to X}(i)$ of non-injectivity scores over the range of removed outliers. The hope is that while outliers can induce fluctuations of scores, these should stabilise over time (i.e. number of removed outliers). To produce final *stable non-injectivity score* of the sample S we sum the differences of scores :

$$\mathrm{Score}(S) = \frac{1}{0.2n} \sum_{i=0}^{\lceil (0.2n) \rceil} (h_{X\to Y}(i) - h_{Y\to X}(i))$$

Causal Inference It is easy to see that when the complex has large non-injectivity score when filtered along y-axis (in either direction), the (functional) causal direction $Y \to X$ is implausible. Similarly for the x-axis and $X \to Y$ direction. If both scores seem large the most sensible solution would be to opt for the existence of a confounding variable T such that the observed sample is of the form $(X(t), Y(t))$. However, we just compare the two scores and choose the larger to be the driving factor, while the absolute value of difference between scores serves as confidence of our decision.

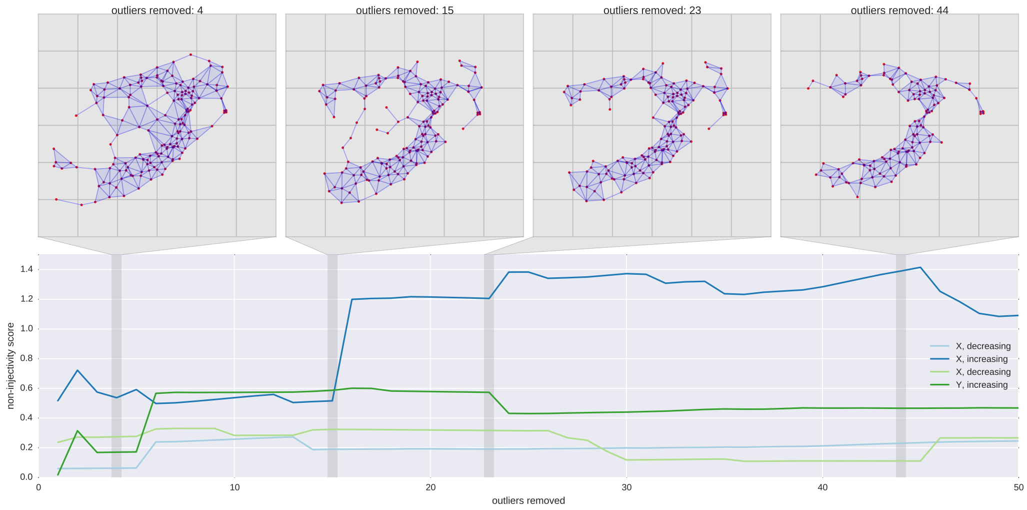


Figure 5: The limited Delaunay triangulation on the points of pair0048.txt from the CEP dataset. The changes in the scores reflect changes in geometry of the occuring complex. Note that the most prominent feature (arc of higher density of points) when reflected in the geometry of the complex after removing the 16-th outlier becomes the dominating factor.

4 Experiments with data

@@@REF: Dionysus

We use python scripts for preprocessing (i.e. standarisation and finding the outliers) and the Dionysus C++ library for topological constructs and calculation of persistence diagrams.

Both simulated and real-world dataset consist of 100 samples, each containing between 300 and 16000 samples from the joint distribution. For a more concrete description see paragraphs below. For each sample S and a given threshold $t \in [0, 1]$ we perform decisions based on the rule

$$\begin{cases} X \to Y, & \text{if } S(F) \geqslant t, \\ Y \to X, & \text{if } S(F) \leqslant -t, \\ \text{no decision}, & \text{if } |S(F)| \leq t. \end{cases}$$

Since prevailing part (over 90% of real-world and 100% of simulated) of the pairs were 2-dimensional, we implemented only analysis of H_0, even for multivariate datasets. It is highly unlikely that information contained in H_0 alone can reveal much about such pairs, hence it serves as a proof-of-concept. Moreover assessing accuracy of the method based on such a limited number of pairs is error prone. We want to stress that in principle, it is possible to examine high-dimensional pairs using projections on appropriate combinations of dimensions as explained at the end of section 1. It is, however unclear how to find the most pomising functional combinations based only on the homological data.

ref: complexity of Smith Normal Form??

In the pipeline there is no re-usage of the generated complexes and boundary matrices which results in simple implementation and allows for easy parallelisation, at the expense of memory and possibly running time. As the homology computation algorithm has worst-case complexity of $O(n^3)$it is plausible that the re-usage might prove useful for larger pairs. In our experiments simulated data presented no significant computational difficulty. Our naive implementation was able to process all simulated pairs from every SIM dataset in less than 10 minutes on a quad-core mobile processor.

However the computational complexity and lack of re-usage of generated structures takes its toll for larger pairs from the CEP dataset. In the processing of full size pairs just three high-dimensional pairs were responsible for 5 hours of computation. The remaining hour was enough for the rest. When sub-samples of size 2000 were drawn running time shortened to less than 20 minutes.

4.1 Simulated data

We used the same four sets of data as in [**?**]. The sets of data consists of 100 pairs (samples from joint distributions) in 2-dimensions, generated under different schemes. Nodes without parents are represented by samples drawn from normal distributions and mapped to the domain by a Gaussian process function. Functions mapping nodes are sampled from a Gaussian processes as well. At the end Gaussian (measurement) noise is added to both coordinates.

SIM is generated using simple relation $Y = f(X)$;

SIM-G has approximately gaussian distribution of cause and follows (approximately) $Y = f(X) + N$, where N is drawn from Gaussian distribution;

SIM-LN is similar to SIM but the noise is reduced, hence the sample is close to deterministic relationship;

SIM-C is confounded using rule $X = f(Z, N_X)$, $Y = g(X, Z)$, where Z and X has similar influence on Y.

The shapes of the joint distributions and more details can be found in [**?** , Appendix C].

For all of these datasets the noninjectivity score clearly obtains accuracy significantly better than chance. The performance is clearly based on non-injective functions relating X and Y, as in *all pairs* where non-injectivity is visible to the human observer were decided correctly. On the other hand, the other pairs which seem close to an injective function seem to be decided purely by chance (as expected). Their confidence is similarly low, hence it should be easy to choose a threshold t for the decision algorithm.

4.2 Real-world data

We used the Tübingen CEP dataset (Cause-Effect Pairs, at version 1.0) available at [**?**]. The dataset contains 100 weighted pairs where weights are imposed to counter the dependence existing in the pairs (e.g. some of them were drawn from in the same

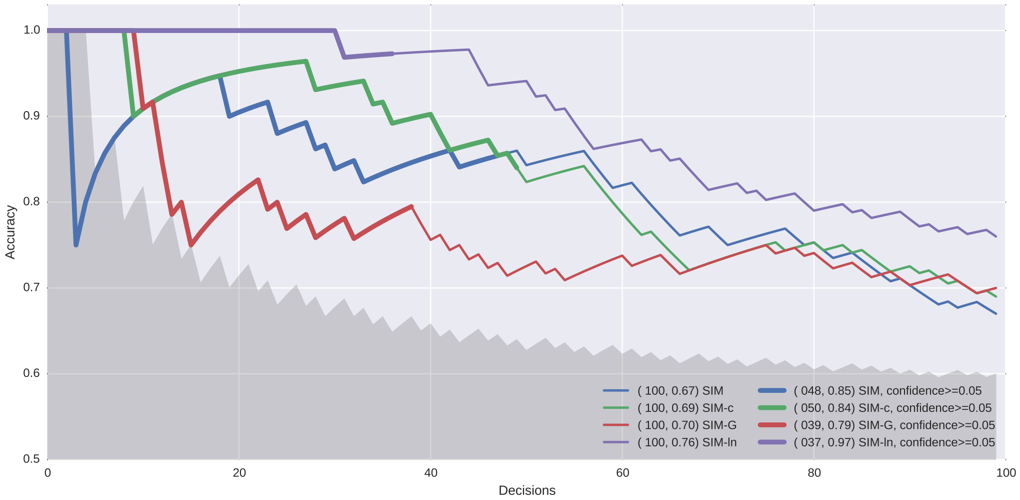


Figure 6: The accuracy plot for SIM datasets. Decisions on the X-axes are sorted according to their inverse confidence score. The thickened lines corresponds to the scenario when we do not resolve the pairs of confidence under threshold $t = 0.05$. In the parentheses are number of decisions taken and the final accuracy. We remark that for the dataset SIM-LN accuracy of $(51, 0.94)$ is obtained with $t = 0.01$. The gray area is the $p > 0.05$ significance level.

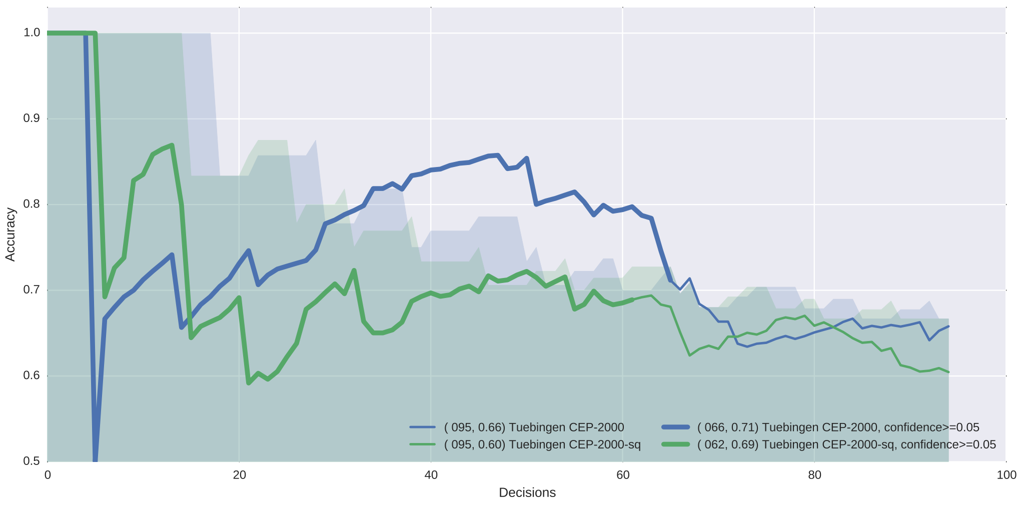


Figure 7: The accuracy plot for CEP datasets: unaltered and quantised. Decisions on the X-axes are sorted according to their inverse confidence score. The thickened lines corresponds to the scenario when we do not resolve the pairs of confidence under threshold $t = 0.05$. In the parentheses are number of decisions taken and the final accuracy. to weights associated with pairs the actual significance depends on the order of decisions taken. Note that seemingly large distances between curves are due to different ordering. However the CEP dataset consists of only about 35 *independent* pairs of which 33 were used hence is too small to judge the significance of the results.

experiment or gathered in similar conditions). Note that weights has been introduced in version 0.8 of the dataset and were severly altered (mostly down-weighted) in version 0.99, hence it is hard to compare accuracies between different versions of the dataset.

Out of those pairs we excluded four:

- pairs: 0047, 0070 and 0071 violate the continuity assumption (variables are binary);
- pair 0095 does not satisfy the unimodality of noise assumption.

It is interesting to note that in the case of pair 0095 uniform subsampling of 2000 points allows the algorithm to pick up the right non-injectivity feature. This suggests that some of the few arbitrary choices made in the algorithm (e.g. the number of outliers considered, the number of nearest neighbours in KNN, uniform measure in the score integral) can be better tuned.

We note that in some of the pairs marginal densities were discrete (i.e. not continuous). If the dataset is sufficiently large this should not be a problem. However if the variables quantisation is uneven this may lead to the score driven by the (observed) sudden changes in density itself rather than the “trend” function. To minimise the potential influence of such cases (i.e. when one of the variables is heavily quantised) we also align points to rectangular grid. We do not expect significant difference in accuracy score: since quantisation introduces small error in estimation of filtering functions, we expect persistence diagrams (hence non-injectivity scores) to be close to each other by Theorem 5. distribution of residuals with mean 0.01 and standard deviation 0.1.

How to asses this eg. from the residuals??

4.3 Comparison with other methods

5 Conclusions

We proposed a non-injectivity based method for causal inference. Methods advantages include relatively short running time, model semi-agnosticism and high accuracy for high confidence pairs. The running time however depends significantly on the sample size and our naïve implementation was not able to cope with more than 10000 points in dimensions greater than 3. There is however room for improvement not only on the software side.

For better approximation of the geometry of the shape one could use kernel density estimation and add weights in complex reconstruction **??**. The first method should be usefull when subsampling the data, or choosing witnesses for simplification of the complex. The second method should be able to (somewhat) tame the combinatorial explosion of the Vietoris-Rips complex for larger datasets and dimensions, while capturing the shape accurately. Moreover one could use distance to measure **??** to obtain proper statistical bounds on the diagrams used to produce scores.

weighted alpha complexes @@@ref

distance to measure @@@ref

As a design feature we anticipate approximately injective relations to have a very low confidence score, thus rendering the persistence homology-based judgement prone to error. While this can be seen as methods weakness, we could take advantage of it. One could chain two (or more) score-based methods and use a blend of the algorithms when the confidence of the topological method falls below a threshold. This allows to have the best of the two worlds: cheap answer when causal direction is evident (i.e. a non-injective relation) and traditional statistical inference when topological methods do not provide a clear answer.

References

A Language of algebraic topology

We provide some of the necessary language of algebraic topology and facts used in the paper. This is by no means a rigorous treatment of the theory. The interested reader may have a look at introductory chapters of standard textbooks on homology theory or modern books on topology for computation .

@@@REF@@@ Munkres, Hatcher

@@@REF@@@ Mrozek, Zomorodian

Simplicial topology We begin with a definition of a single building block, a simplex.

Definition 3 (Simplex)**.** An n-dimensional (geometric) simplex σ is the convex hull spanned by the standard basis and 0 in $\mathbb{R}^n$. When we speak about simplex spanned by points $(p_0, \ldots, p_n)$ we only mean a (abstract) simplex whose vertices has been only *labeled* by the points.

Spaces in simplicial topology are build out of simplices which have to interact in a regular way. Given a family of simplices we would like to take a union of them, but to make the forthcoming theory easier we should require that the intersections of every two simplices is again a (lower dimensional) simplex.

Definition 4 (Simplicial complex)**.** A simplicial complex Σ is a set of simplices $\{\sigma_i\}_{i \in I}$ such that whenever σ_i and σ_j intersect, their intersection is again in the collection.

Note that simplicial complex is *abstract* in the sense that its vertices need not to be points in $\mathbb{R}^n$ and its edges (or higher dimensional simplices) represent merely similarities between its vertices. In the following we will use the Delaunay complex in dimensions 2 and 3 and Vietoris-Rips complex in higher dimensions.

Example (ε-Čech complex)**.** Suppose that a finite set $F \subset \mathbb{R}^n$ is given. Fix $\varepsilon > 0$ and consider the following simplicial complex $\check{C}(\varepsilon)$:

- we add a vertex v_x for every point $x \in F$;
- for $x, y \in F$ we place an edge between the the corresponding vertices v_x and v_y whenever $d(x, y) < \varepsilon/2$;
- for every set of $(k+1)$-points $V \subset F$: if sphere circumscribed on points of V has radius smaller than $\varepsilon/2$, we add the k-dimensional simplex spanned by the corresponding vertices.

The Čech complex, although very useful from theoretical point of view is hard to compute, as it requires computations of spheres. An variant of a similar complex which is easier to compute was proposed (independently) by Vietoris and Rips.

Example (ε-Vietoris-Rips complex)**.** Suppose that a finite set $F \subset \mathbb{R}^n$ is given. Fix $\varepsilon > 0$ and consider the following simplicial complex $VR(\varepsilon)$:

- we add a vertex v_x for every point $x \in F$;
- whenever $d(x, y) < \varepsilon$ for $x, y \in F$ we place an edge between the corresponding vertices v_x and v_y.
- for every set of $(k+1)$-points $V \subset F$: if all the pairwise distances in V are smaller than ε, we add the k-dimensional simplex spanned by the corresponding vertices.

Commonly used for clustering ε-neighbouring graph is a special case of the Vietoris-Rips complex when we limit construction to $k = 1$.

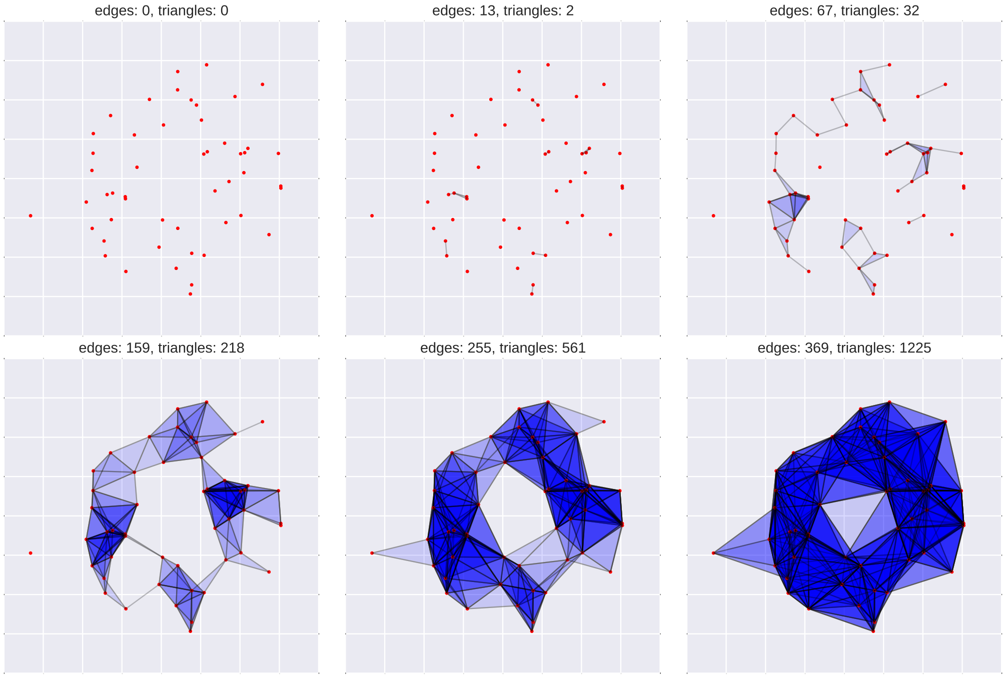

edges: 0, triangles: 0
edges: 13, triangles: 2
edges: 67, triangles: 32
edges: 159, triangles: 218
edges: 255, triangles: 561
edges: 369, triangles: 1225

Figure 8: The ϵ-filtered Vietoris-Rips complex on the same sample as in figure **??**, limited to $k \leq 2$. The complex captures the initial shape accurately, however due to its nature it suffers from the combinatorial explosion for larger epsilonss. Unlike the Delaunay Triangulation its construction is fast as it depends only on the pairwise distances.

Chain complexes and homology

Definition 5 (Chain complex over of $\mathcal{X}$)**.** A chain complex

$$\cdots \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \to 0$$

is a sequence of vector spaces and linear operators, where each C_k is spanned by all k-dimensional simplices present in $\mathcal{X}$. The linear maps ∂_k are the boundary operators (assign to every simplex its boundary as a linear combination of simplices of lower dimension).

In more rigorous way we may define $\partial_k \colon C_k \to C_{k-1}$ as follows. Since this is a map between vector spaces it suffices to define ∂_k on basis of C_k and then extend linearly. Let $\sigma(\{x_0, \ldots, x_k\})$ be a k-simplex in $\mathcal{X}$, i.e. a convex hull of the points $x_0, \ldots, x_k$. Its boundary $\partial_k(\sigma)$ is a linear combination of all of its $(k-1)$-simplices

$$\partial_k(\sigma) = \sum_{i=0}^{k} \tau(\{x_0, \ldots, x_k\} - \{x_i\}).$$

As such chain complexes are not very robust: a small change in the geometry of shape we approximate with a simplicial complex (e.g. addition a new point to F) may result in completely different chain complex with no easy way of comparison. Therefore we define homology which nullifies these changes.

Definition 6 (Homology of $\mathcal{X}$)**.** We define k-th homology of $\mathcal{X}$ as

$$H_k(\mathcal{X}) = {}^{\ker \partial_k}\big/_{\operatorname{im} \partial_{k+1}}$$

In our setting homology groups of a simplicial complex are vector spaces. As a side-note we provide connections to other graph-theoretic notions: the graph Laplacian can also be defined in the terms of boundary operator: $\mathcal{L} = \partial_1 \partial_1^T$. Higher homologies capture more complex information about geometry of simplicial complex, e.g. one can define higher Laplacians using higher boundary operators.

Filtrations and persistence

Definition 7 (Filtration)**.** A filtration on space $\mathcal{X}$ is an increasing family of spaces

$$\varnothing = \mathcal{X}_{-1} \subseteq \mathcal{X}_0 \subseteq \mathcal{X}_1 \subseteq \cdots \subseteq \mathcal{X}_{n-1} \subseteq \mathcal{X}_n = \mathcal{X}.$$

A practical way of describing a filtration is by specifying a (*filtering function*) $f \colon \mathcal{X} \to \mathbb{R}$ and setting $\mathcal{X}_t = \{x \in \mathcal{X} \colon f(x) \leqslant t\}$.

If $\mathcal{X}$ is a simplicial complex we would like the filtration to preserve its structure, i.e. we would like each $\mathcal{X}_t$ to be a simplicial complex as well. Therefore we require additionally that filtering function is increasing on every simplex, that is if τ is a face of σ then $f(\tau) \leqslant f(\sigma)$. It is easy to see that for a complex with finite number of simplices filtering function is equivalent to an increasing sequence of spaces as in the definition. Vietoris-Rips complexes come naturally as examples of filtered spaces. By choosing different values of ε we create an increasing sequence of simplicial complexes – the function which assigns to every simplex diameter of the set of its vertices is a properly defined filtering function.

Instead of fixing (e.g. learning in some way) ε and analysing the ε-neighbouring graph on F, it might be more insightful to look at the whole spectrum of possible ε's and observe changes in geometry of the arising complex as we increase the epsilon. These changes are captured by persistent homology. A very similar concept for Delaunay complexes is filtration by α-complexes which we will not describe here.

Persistence homology Suppose that a k-cycle z is a generator of $H_k(\mathcal{X}_{t_b})$ and this cycle is not homologuous to any other generator of $H_k(\mathcal{X}_t)$ for any $t < t_b$. Similarly if $t_d \geqslant i$ is the minimal filtration level in which z becomes homologuous to 0 we say that z dies at time t_d.

Definition 8 (Persistent homology)**.** Persistent homology of a filtered complex $\mathcal{X}$ is the bi-gradation on the homology of $\mathcal{X}$ given by

$$H_k^{i,j}(\mathcal{X}) = \operatorname{im}\left(H_k(\mathcal{X}_i) \hookrightarrow H_k(\mathcal{X}_j)\right).$$

Technically speaking persistence diagram in dimension k is a multiset of pairs (t_b, t_d) where every pair corresponds to a generator z in the kernel of the map on homology induced by inclusion $\mathcal{X}_{t_b} \hookrightarrow \mathcal{X}_{t_d}$.