Non-injectivity Approach to Causal Inference

Marek Kaluba, ...

Thursday 11th February, 2016

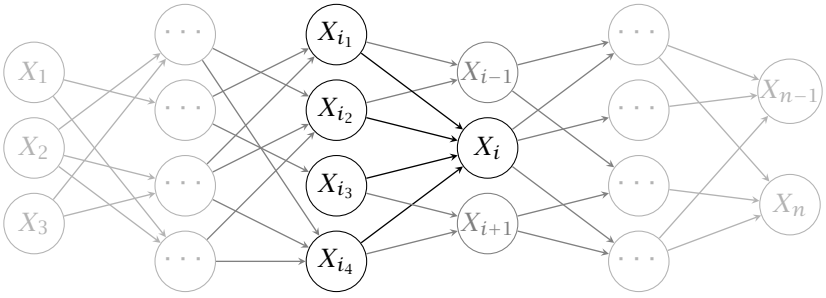
Abstract

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 finite sample. Although performing an intervention (controlled randomized 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.

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.



The existence of an edge indicates the dependence between variables, however there is no clear way to asses the direction. It turns out that the noise present in the system (or in the measurements) has strong impact on what can be inferred about the graph.

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 poset called structural equation model (SEM) encodes more information than the DAG itself. 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.

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

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

and linear non-gaussian additive model (LINGAM) which assumes that all the functions f_k are linear and the distribution of n_i is different from the Gaussian distribution.

The standard approach is to use a function from a fixed class to model the orientation of edges. Strong assumptions are imposed on the possible class of *funcional* dependencies and this allows statistically test hypotheses on the direction 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.

In this paper we propose a not-standard way of estimating the 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 $p(X_1, \ldots, X_n)$ of the random variables $\{X_i\}_{i=1}^n$ is given. The task is to recover a DAG G consistent with $p(X_1, \ldots, X_n)$, i.e. a graph such that repeated sampling from each node X_i converges to the initial joint distribution. We will use the following assumptions without implicit invocation.

1. the observed data has been generated by repeated sampling from a DAG G;
2. each sample noise (n_i) is drawn from jointly independent distribution with density $p_N(n) = \prod_i p_i(n_i)$;
3. all random variables are continuous;
4. the probability density functions p_{X_i} of each node variable and p_i of noise have the same support;

5. each probability density n_i and p_{X_i} has at most one maximum, i.e. X_i and N_i are uni-modally distributed random variables.

The first two assumptions are common in the causal inference framework, hence we will not comment on them. The third and forth are natural in the sense, that we do expect to observe the noise across the whole domain of the input variables. This assumption is similar in spirit to one of , where occurrence of noise and non-linearity of function have to coincide. We discuss the last assumption below.

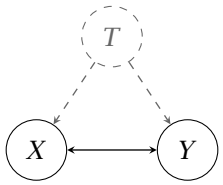
@@@REF@@@ non-linear causal discovery??

In the following we will restrict to a simple case of two variables X and Y and the task of inferring the causal relationship between them. The generalisation to multidimensional setting is possible and we provide additional details whenever the two settings diverge.

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. is smooth). Under the uni-modality 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 graph of the function f, with error determined by the noise.

In particular, if we can learn that the shape of the empirical density $\{(x_i, y_j)\}$ is close to graph of a non-injective function $y_i = f(x_i)$, this excludes the causal direction $Y \to X$. We assume therefore that there is always a causal relationship between X and Y and this allows us to conclude that the remaining possibility one is the true causal structure. To asses the non-injectivity of a function without regressing the function itself we will use tools of computational geometry (i.e. simplicial complexes) to approximate the graph and then algebraic topology (filtrations and persistence homology) to produce a non-injectivity score.

We note that the assumption above 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.



In high-dimensional setting we provide a similar framework. Given $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_k)$ 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 model

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

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. Next we prove basic facts about expected behaviour of topology of simplicial complexes. This theoretical results (section 2.3) serve as a backbone for further experiments with data in section 3. We use both simulated and real world data from . Other existing methods are quoted and comparisons are made in section 3.3.

pain?

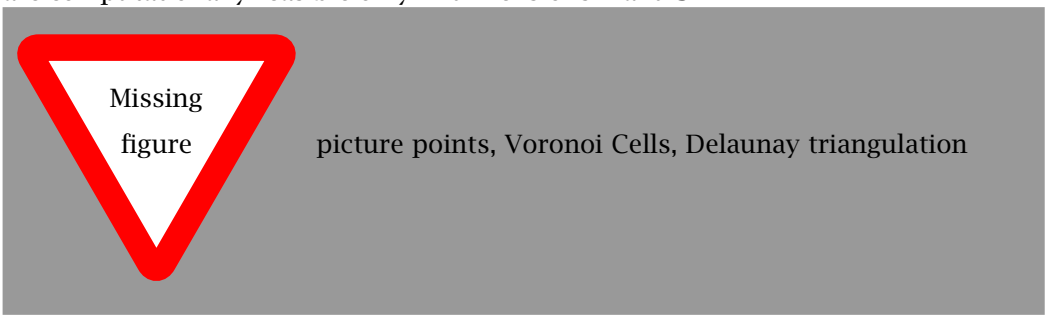
@@@REF@@@ Cause-Effect-Pairs, CE-Benchmarks

2 Algebraic topology

2.1 Informal exposition

We will only treat triangulated shapes, i.e. objects given as a form of triangular mesh in $\mathbb{R}^n$. Usually it is enough to specify set F of vertices of the mesh and a list of maximal simplices (a sets of subsets of F). E.g. to specify triangulated surface in $\mathbb{R}^3$ it is enough to specify vertices of the mesh and list of triangles spanned by triples of vertices. Union of all vertices, edges and triangles forms in this case an example of simplicial complex defined in the next paragraph.

An common example of 1-dimensional simplicial complex is a graph. In computational geometry commonly used is Delaunay triangulation on the set of points, which may be defined as dual of the Voronoi tessellation, or the triangulation of the convex hull on F which minimises the shortest paths distance (or maximises the smallest angles in each triangle). These triangulations exists in any dimensions, however they are computationally feasible only in dimensions 2 and 3.



picture points, Voronoi Cells, Delaunay triangulation

Given such regularly structured objects one can try to assign some well behaved *fingerprint*, or *signature* in the terms of an algebraic object. In purely mathematical setting groups and rings are natural choices, however these objects are very hard to compute with. Thus computational algebraic topology usually settles on vector spaces over the field of 2 elements (integers modulo 2).

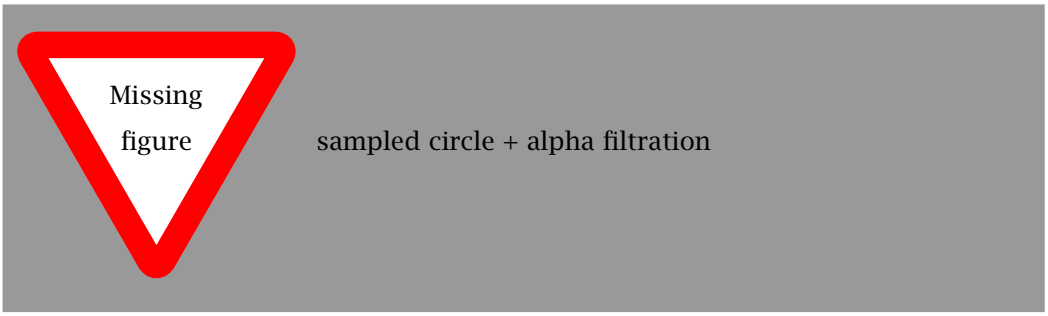
Homology of a triangulated shape $\mathcal{X}$ is a sequence of vector spaces $H_0(\mathcal{X}), H_1(\mathcal{X}), \ldots$ is meant to capture qualitative information about geometry of the object. To keep the geometric intuition we note that $H_0(\mathcal{X})$ is a vector space with basis the *connected components* of $\mathcal{X}$. Similarly if $\mathcal{X}$ is a graph, $H_1(\mathcal{X})$ is spanned by *linearly independent cycles* of $\mathcal{X}$ modulo boundaries. The void in 2-sphere (approximated e.g. as triangulated icosahedron) results in non-zero $H_2(S^2)$ and similarly Higher homology groups capture higher-dimensional features. Homology groups does 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 geometry, this is usually very hard to interpret without additional information. Usually one resolves to look at *persistent homology* which contains information on the evolution of a 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 fill a cycle – in some cases this can be close to its simplicial diameter. 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". For example distance can be considered as a filtering function on a graph. This filtration arises when we initially accept all points and then in each step we add edges which are shorter than certain value. 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 section, and analyse the following example.

Example. Consider a noisy sample of 100 points drawn from a circle in the plane of radius 1000 with noise $N(0, 300)^2$. We build a full Delaunay traingulation $\mathcal{X}$ on them and filter edges (and triangles) by distance between points. The filtration comes from limiting the length of edges we would like to 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. While reconstruction of the original density would be quite hard, the existence of a long-lived cycle indicates that the underlying density has "circular shape". Whereas it is not possible to reconstruct the actual shape from this homological information, it is enough to e.g. claim that the density is not a simple Gaussian, nor a sum of a few of them.



In the next section we provide some definitions of the tools of computational algebraic topology. 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

2.2 Language of algebraic topology

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

Definition 1 (Simplex)**.** An n-dimensional simplex σ is the convex hull spanned by 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 2 (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.

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.

Note that this simplicial complex is *abstract* in a sense that its vertices need not to be points in $\mathbb{R}^n$ and its edges (or higher dimensional simplices) represent similarities between its vertices (e.g. subjects). Commonly used for clustering ε-neighbouring graph is a special case of the Vietoris-Rips complex when we limit construction to $k = 1$. In the following we will use the Delaunay complex in dimensions 2 and 3 and Vietoris-Rips complex in higher dimensions.

Chain complexes and homology

Definition 3 (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 4 (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 sidenote 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 5 (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 The changes in geometry are best spoken of using the language of Morse functions which we describe in visual terms in 2.3. Below we provide some motivation behind the idea of persistence. For more rigorous treatment of the subject please refer to **??**.

missing ref for theory of persistent homology

An n-dimensional persistence diagram is a multiset of pairs $(t_b, t_d) \in \mathbb{R}^2$. Consider the case of $n = 0$. 0-dimensional features of $\mathcal{X}_t$ are the connected components of sub-level set $\{x \in \mathcal{X}\colon f(x) \leqslant t\}$. A 0-dimensional feature emerges when at time t_b

(*birth time*) a new cluster of points is added to X_{t_b} that was not present for $t < t_b$. Similarly when the cluster merges with another (because we kept adding points) we say that the 0-feature dies 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 $(\mathcal{X}, f)$. 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 in the last stage of the filtration we say that it is of infinite life. Note that in the 0-th persistence diagram there will be an infinite life point for every connected component of the final shape $\mathcal{X}_\infty = \mathcal{X}$.

Of course such diagrams exist in any dimension. The more formal treatment is as follows. 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 6 (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 $\mathcal{X}_{t_b} \hookrightarrow \mathcal{X}_{t_d}$. For technical reasons we also include points on the diagonal with infinite multiplicity. These diagrams provide a natural way of comparing two filtered simplicial complexes. We have a natural notion of distance between the diagrams and, in certain cases, robustness to noise.

Definition 7 (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 some points to the diagonal is still a bijection.

2.3 Theory behind the method

Below we quote several results which we will need to justify the method described in section 2.4. In the first part we assume that we deal with perfect objects (manifolds and smooth functions). In the second we will show how to reconstruct a simplicial complex from a finite noisy sample drawn from M which is close in homological properties to M.

Suppose that $f = (f_1, \ldots, f_k)\colon U \to \mathbb{R}^k$ is a smooth (at least C^2) function defined on a compact manifold (possibly with boundary) $U \subset \mathbb{R}^n$. We will consider the graph of f, i.e. the manifold defined as

$$G(f) = \left\{(x, f(x)) \in \mathbb{R}^{n+k}\colon x \in U\right\},$$

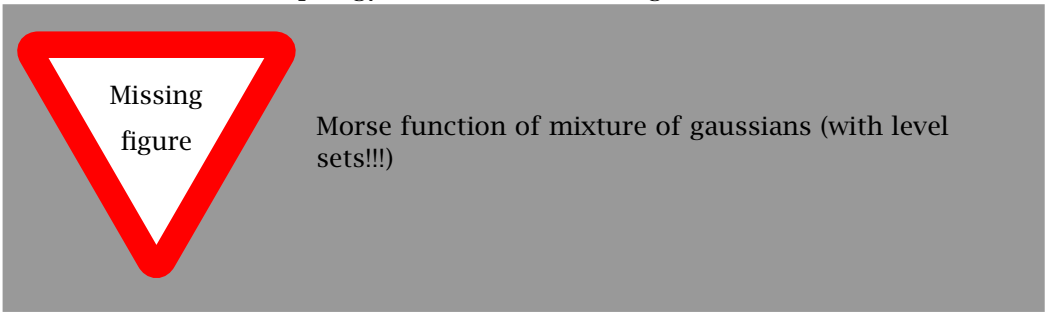
Figure 1: A graph of mixture of gaussians. In this the case sub-level set are $\mathcal{X}_t = \{(x, y, p(x, y)) : p(x, y) \geqslant 1 - t$. Critical points of f are points where tangent plane is parallel to the XY-plane. Note how topology of X_t changes as we increase t.

filtered by the function $\pi_j \circ f = f_j$. It will follow that if f_j is an injective function, then there are no (non-diagonal) points in persistence diagram (of any dimension) of $G(f)$ filtered by f_j. Therefore, later on, we can conclude that if the persistence diagram of a noisy approximation of f contains points *far from the diagonal*, this can be attributed to non-injectivity of the function f_j rather than the noise or the manidold U itself.

Finally we will show that one can reconstruct a shape close to M and presistence diagrams of some filtrations on M from a finite, noisy sample drawn from M.

Morse functions and filtrations In the case where U is a manifold, the geometric features that took part in the definition of persistence homology can be characterised using notion of Morse functions. Consider the following example.

Example. Suppose a mixture of gaussians $\mathcal{T} \subset \mathbb{R}^2$ is given as a set of points $(x, y, p(x, y))$ as depicted in figure 2.3. The (inverse) height function $f(x, y, p(x, y)) = 1 - p(x, y)$ is a valid filtering function. The presistent homology captures exactly those moments when topology of the level-sets changes.



Morse function of mixture of gaussians (with level sets!!!)

We say that a C^2 function on a manifold $f : M \to \mathbb{R}$ is *Morse* if the gradient Df has only isolated zeros (so called critical points of f) and at those points the Hessian matrix $D^2 f$ is non-degenerate. We note that Morse functions are dense in the set of all functions on M. 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.

We set the notation for this paragraph. Let $f = (f_1, \ldots, f_k) : U \to \mathbb{R}^k$ be a C^2 function, where $U \subset \mathbb{R}^n$ is a compact manifold of dimension u, and let $G(f)$ denote the graph of f. Let $\pi_j : \mathbb{R}^{n+k} \to \mathbb{R}$ denote projection on the j-th coordinate.

Lemma 1. *Suppose that f_j is injective and that $\pi_j : G(f) \to \mathbb{R}$ for $j \geqslant n + 1$ is a Morse function. Then $\pi_j(x, f(x)) = f_j(x)$ has no critical points and hence persistence diagrams in any dimension of $G(f)$ filtered by π_j are empty.*

The above lemma is a folklor knowledge, but we provide a proof for completeness.

Proof. Suppose that there is a point (t_b, t_d) (possibly $t_d = \infty$) in ℓ-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 ℓ in $\pi_j^{-1}(t_b)$. Again Morse theory tells us that there exists a local chart $\psi\colon \mathbb{R}^{u+k} \to G(f)$ around y_b such that $\pi_j(x, f(x)) = f_j(x)$ can be expressed as

$$\pi_j(\psi(y_1, \dots, y_{n+k})) = f_j(x_b) - \sum_{i=1}^{\ell} y_i^2 + \sum_{j=\ell+1}^{n+k-l} y_j^2.$$

As ψ is a homeomorphism its f_j responsible for the (local) non-invertibility of $\pi_j \circ \psi$. Thus we arrive at a contradiction as f is smooth and injective hence (locally) invertible by the inverse function theorem. □

Even if $\pi_j = f_j$ does not satisfy Morse conditions, one can find a Morse function $\widehat{f}_j\colon G(f) \to \mathbb{R}$ such that $\widehat{f}_j$ and f_j are close in the sup-norm, since Morse functions are dense in the set of smooth functions on every manifold. Note that we can estimate the persistence diagram of f_j by using $\widehat{f}_j$ instead with precision related to ε by Theorem 5.

Lemma 2. *Choose $n < j \leq k$ and adjust f_j to be a Morse function $\widehat{f}_j$ such that $\|\widehat{f}_j - f_j\|_\infty < \varepsilon$. If f_j is injective no persistence diagram of $G(f)$ filtered by $\widehat{f}_j$ contains points of lifespan greater than 2ε.*

Proof. Suppose that an ℓ-th persistence diagram of the function $\widehat{f}_j\colon G(f) \to \mathbb{R}$ contains point 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 $\gamma\colon [0,1] \to G(f)$ of negative gradient flow connecting y_d and y_b. Moreover while $f_j \circ \gamma$ is increasing, $\widehat{f}_j$ is strictly decreasing when restricted to the image of γ.

Therefore

$$\left|\widehat{f}_j(x_d) - \widehat{f}_j(x_b)\right| \leqslant \left|\widehat{f}_j(x_b) + f_j(x_d) - f_j(x_b) - \widehat{f}_j(x_d)\right| \leqslant$$

However

$$\sup_{G(f)} |f_j - \widehat{f}_j| \leqslant \sup_{\gamma} |f_j - \widehat{f}_j| \leq f(y_b)$$

, i.e. that locally, around y_b $\widehat{\pi}_j$ can be expressed as

$$\widehat{\pi}_j(\gamma) = \widehat{\pi}_j(y_b) - \sum_{i=1}^{\ell} y_i^2 + \sum_{j=1}^{u-\ell} y_j^2.$$

Note that since $y_b = (x_b, f(x_b))$ locally $|\widehat{\pi}_j(x, f(x)) - f_j(x)| < \varepsilon$. Similarly there exists a critical point y_d of index $\ell + 1$ such that in its neighbourhood

$$\widehat{\pi}_j(\gamma) = \widehat{\pi}_j(y_d) - \sum_{i=1}^{\ell+1} y_i^2 + \sum_{j=1}^{u-\ell-1} y_j^2,$$

and locally $|\hat{\pi}_j(x, f(x)) - f_j(x)| < \varepsilon$.

Since the cell attachment at y_d kills cycle created at y_b there is a negative griadient flowline connecting y_d and y_b of length at least $\hat{\pi}_j(y_d) - \hat{\pi}_j(y_b) > 2\varepsilon$.

□

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

Approximation of manifolds by simplicial complexes It was shown by de Rham in [] that the homology of a manifold M 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 noisy sample $F \subset R^n$ of points sampled from M is given. Of course points in F will not necessarily belong to M, but their expected distance is bounded by noise. We investigate under which conditions we can use Vietoris-Rips complex build on F to approximate the homology of M.

Definition 8 (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.

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

For a proper definition of reach (using distance to medial axis) please refer to . The reach intuitively means the maximal radius r such that thickening of M in the normal direction by $\varepsilon < r$ does not change the topology.

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

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.*

Therefore, if we are given a finite sample F from a manifold M and we know that the variance of noise is not too large, e.g. locally always smaller than the one eight of the reach of 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 homology 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 homological features of M. Therefore to estimate homology of M it is enough to estimate filtering functions defined on Vietoris-Rips complexes.

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

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 $f, g : M \to \mathbb{R}$ 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.

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

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.

2.4 Algorithm used for causal inference

Again we will restrict to the 2-dimensional case and provide necessary modifications as we proceed. We will treat the first coordinate as (values of) random variable X and the second as Y.

We propose the following algorithm for assessing the non-injectivity of a function without regressing it first. Suppose that a finite set $F \subset \mathbb{R}^2$ 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 F was sampled (without noise) from a graph of an injective function, the approximating complex would have similar shape. Note that graph of an injective function (filtered by projections) have 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 disregard the point of infinite life and measure the distance between persistence diagrams of different filtrations on $\mathcal{X}$ with an empty diagram to produce a *non-injectivity score*.

It is easy to see that if the complex has large non-injectivity score when filtered along OY-axis, the (functional) causal direction $Y \to X$ is implausible. Similarly for the OX-axis and $X \to Y$ direction. If both scores seem large the most sensible would be to opt for 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 absolute value of difference between scores serves as confidence score of our decision.

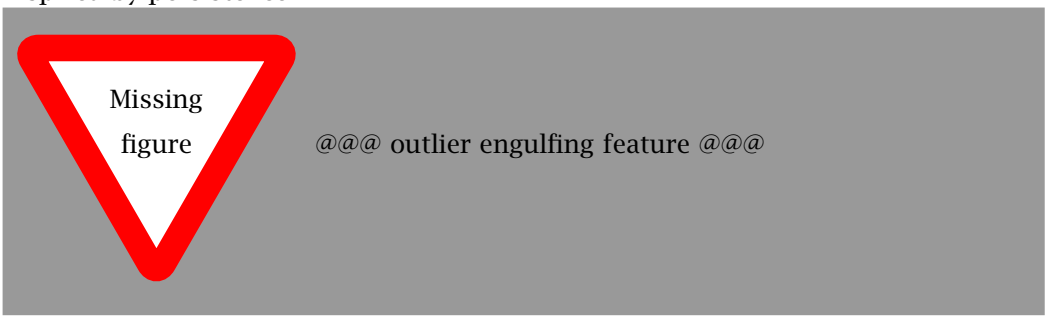
Non-injectivity score Suppose that a finite set $F \subset \mathbb{R}^2$ is given. We create $\mathcal{X}$, the minimal connected subcomplex of the Delaunay complex on points in F. We do it by first constructing the full Delaunay triangulation and then we continue to remove the longest edge as long as the graph 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 n_X and n_Y. respectively. It is worth to limit the construction of $\mathcal{X}$ only to $\max(n_X, n_Y)$-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

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

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.

Noisy sample 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.

The outliers are found one-by-one using KNN method and stored in list. Then complex $\mathcal{X}(n)$ is constructed as described in the previous paragraph using all point from F except the n-furtherest outliers. Non-injectivity scores for the complex are computed.

To add more stability to scores we standardise data prior to each removal. This produces real-valued functions $h_{X\to Y}(n), h_{Y\to X}(n)$ 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 sample F we integrate the difference of scores:

$$S(F) = \int_{i\in\mathcal{O}} h_{X\to Y}(i) - h_{Y\to X}(i)$$

3 Experiments

@@@REF: Dionysus

We use python for pre-processing and Dionysus C++ library for topological constructs and calculation of persistence diagrams.

Both simulated and real-world dataset consist of around 100 *pairs*, each containing between 300 and 16000 samples from joint distribution. For each pair and given threshold $t > 0$ we perform decisions based on the rule

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

Since prevailing part of the examples were 2-dimensional we implemented only analysis of H_0, even for multidimensional pairs. It is highly unlikely that analysis of H_0 alone can reveal much about such examples, hence it serves as a proof-of concept. It is, however, possible to examine high-dimensional pairs using projections on appropriate combinations of dimensions.

In the pipeline there is no re-usage of the generated complexes and boundary matrices which allows for greater parallelisation, at the expense of memory. As the homology computation algorithm has worst-case complexity of 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 each simulated dataset (i.e. $\sim$100 pairs) on quad-core mobile processor in less than 10 minutes. Clearly the $O(n^3)$ complexity and naive implementation takes its toll for larger pairs. The running time the same when subsamples of size 2000 were drawn from experimental pairs. For processing full size pairs about 6 hours is required.

3.1 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 (random variables) are 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.

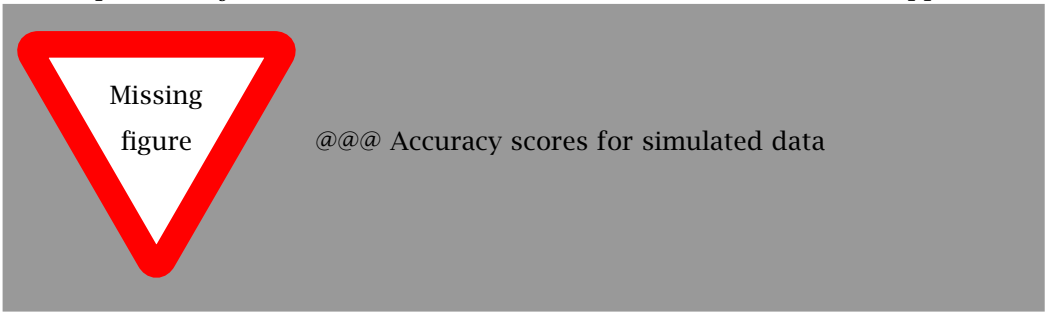
SIM is generated using simple relation $Y = f(X)$;

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. What is worth noticing, the performance is clearly based on non-injective functions relating X and Y. In particular *all cases* where non-injectivity is visible to human observer were decided correctly. On the other hand, the injective cases 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. In the SIM-LN scenario 52 most confident pairs were decided with accuracy 92%. For the remaining 48 pairs the (normalised) decision confidence was below 0.05.

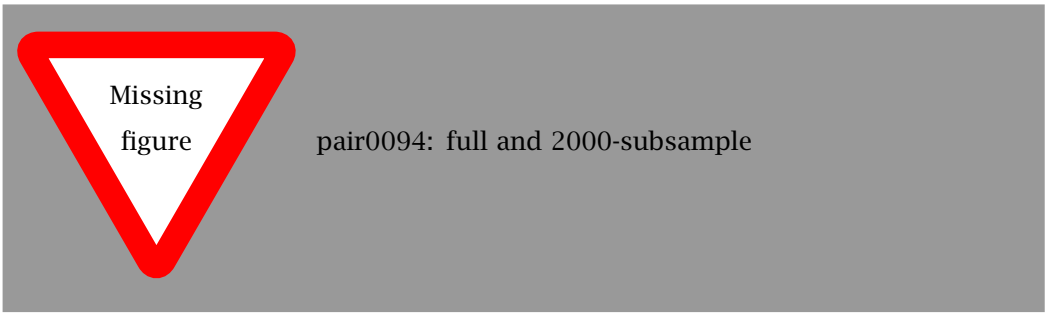
3.2 Real-world data

We used the Tuebingen CAUSE-EFFECT PAIRS as available in [**?**]. The newest version (0.9999) contains weighted 98 pairs to counter the dependence existing in the pairs (a few of them were drawn from similar experiment or gathered in similar conditions).

Out of those pairs we excluded four:

- pairs: 0047, 0070 and 0071 violate the continuity assumption (variables are binary);
- pair 0094 does not satisfy the unimodality of noise assumption.

It is interesting to note that in the case of pair 0094 uniform subsampling of 2000 points allowed the algorithm to pick up the right non-injectivity feature. This suggests that some of few arbitrary choices made in the algorithm (i.e. the number of outliers considered, the number of nearest neighbours in KNN, uniform measure in the score integral, etc) can be better tuned.



pair0094: full and 2000-subsample

We note that in some of the pairs marginal densities were not unimodal as well, which in the case of significant or uneven noise may lead to answer driven by (observed) sudden changes in noise rather than the "trend" itself.

To minimise the potential influence of data gathering methods (i.e. one of the variables is heavily quantised) we also align points to rectangular grid. We do not expect significant difference in accuracy score. This is indeed the case: 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.

3.3 Comparison with other methods

4 Conclusions

References