Journal of Machine Learning Research 20 (2019) 1-21 Submitted 10/19; Published date published

Geometric goodness of fit measure to detect patterns in data point clouds

Alberto J. Hernández ALBERTOJOSE.HERNANDEZ@UCR.AC.CR
Centro de Investigación en Matemática Pura y Aplicada (CIMPA)
Escuela de Matemática
Universidad de Costa Rica
San José, Costa Rica

Maikol Solís MAIKOL.SOLIS@UCR.AC.CR (CORRESPONDING AUTHOR)
Centro de Investigación en Matemática Pura y Aplicada (CIMPA)
Escuela de Matemática
Universidad de Costa Rica
San José, Costa Rica

Ronald A. Zúñiga-Rojas RONALD.ZUNIGAROJAS@UCR.AC.CR
Centro de Investigaciones Matemáticas y Meta-Matemáticas (CIMM)
Escuela de Matemática
Universidad de Costa Rica
San José, Costa Rica

Editor:

Abstract

The curse of dimensionality is a commonly encountered problem in statistics and data analysis. Variable sensitivity analysis methods are a well studied and established set of tools designed to overcome these sorts of problems. However, as this work shows, these methods fail to capture relevant features and patterns hidden within the geometry of the enveloping manifold projected onto a variable. Here we propose an index that captures, reflects and correlates the relevance of distinct variables within a model by focusing on the geometry of their projections. We construct the 2-simplices of a Vietoris-Rips complex and then estimate the area of those objects from a data-set cloud. The analysis was made with an original R-package called TopSA, short for Topological Sensitivity Analysis. The TopSA R-package is available at the site https://github.com/maikol-solis/TopSA.

Keywords: Goodness of fit, R^2, Vietoris-Rip complex, Manifolds, Area estimation

Introduction

Data point cloud recognition is basic task in any statistical procedure. Draw a pattern into the data shed lights about the inherent phenomena trying to explain. In the literature are tools like linear regression or clustering to achieve this (Hastie et al., 2009; Hand, 2005). Other research branch uses data visualization techniques to highlight features hidden in the data (Tufte, 2001; Myatt and Johnson, 2009; Buja et al., 2005).

Professionals in computational modeling aim to reduce their problem choosing the most relevant factors of the problem. One way to tackle the problem is through goodness of fit measures to find the relevance or certain chosen model to explain a variable. One classic way to determine if this some variables fits inside a model is using the coefficient of

determination R^2. This quantity measure the amount of variance explained by some model against the variance explained by the model formed only by a constant. It measure how much fitting a model is preferable against fitting a single constant. If R^2 is near to one then the model fits well to the data. Otherwise, is better just to adjust a constant.

The quantity R^2 has been controversial since its origins (Barrett, 1974). For example, we can increase the R^2 score only by adding new variables to the model (even if they are irrelevant to the problem) or a model with high R^2 does not imply that the covariate explain the outcome (causation vs correlation).

Some extensions and properties have been discovered through the years. The work of Barten (1962) proposed a bias-reduced R^2. A Bayesian analysis was conducted by Press and Zellner (1978). The first two moment of the R^2 and the adjusted R^2 are studied by Cramer (1987). He showed that in small samples, the R^2 tends to be higher. Even those constraints, Barrett (1974) conclude the utility of these measures in applied research.

In any case, this goodness of fit measure overlook the geometric arrangement of the data. They build the statistical relation between X and Y and then present it in the form of an indicator. Depending on this simplification they do not consider the geometric properties of the data. For example, most indices will fail to recognize structure when the input variable is zero-sum, treating it as random noise.

The analysis of topological data is a recent field of research that aims to overcome these shortcomings. Given a set of points generated in space, it tries to reconstruct the model through an embedded manifold that covers the data set. With this manifold we can study the characteristics of the model using topological and geometrical tools instead of using classic statistical tools.

Two classical tools used to discover the intrinsic geometry of the data are Principal Components Analysis (PCA) and Multidimensional Scaling (MDS). PCA transforms the data into a smaller linear space preserving the statistical variance. The other approach, MDS, performs the same task but preserves the distances between points. Recent methods like the isomap algorithm developed in Tenenbaum (2000) and expanded in Bernstein et al. (2000); Balasubramanian (2002) unify these two concepts to allow the reconstruction of a low-dimensional variety for non-linear functions. Using geodesic distance isomap identifies the corresponding manifold and searches lower dimensional spaces where to embed it.

In recent years new theoretical developments have used tools such as persistent homology, simplicial complexes, and Betti numbers to reconstruct manifolds, the reconstruction works for clouds of random data and functional data, see Ghrist (2008), Carlsson (2009, 2014). In Gallón et al. (2013) and in Dimeglio et al. (2014) some examples are presented. This approach allows *"Big Data"* to be dealt with quickly and efficiently, see Snášel et al. (2017).

In this work we aim to connect the concept of goodness of fit with the analysis of topological data through a geometrical R^2 index. By doing this it will be possible to determine what variables has structured patterns using the geometric information extracted from the data.

The outline of this paper is: Section 1 deals with basic notions, both in sensitivity analysis and in topology. In Subsection 1.1 some of the most used and well-known statistical methods are reviewed and commented on. We finish this subsection with an example which motivated the work in this paper. Subsection 1.2 deals with preliminaries in homology, and describes the Vietoris-Rips Complex. Section 2 explains the method used to create our sensitivity index; Subsection 2.1 describes the construction of the neighborhood graph, and deals with different topics such as *the importance of scale*, the *Ishigami Model* and presents programming code to determine the radius of proximity. Subsection 2.2 describes the algorithm used to construct the homology complex through the *Vietoris-Rips complex* and Subsection 2.3 explains our proposed sensitivity index. Section 3 contains a description

of our results, it describes the software and packages used to run our theoretical examples. Subsection 3.1 is a full description of each theoretical example together with visual aids, such as graphics and tables describing the results. Subsection 3.2 is an application of our algorithm to a well-known hydrology model. Finally, Section 4 contains our conclusions and explores scenarios for future research.

1. Preliminary Aspects

In this section we will discuss the context and tools needed to implement our geometric goodness-of-fit.

1.1 Measure of goodness-of-fit

Let $(X_1, X_2, \dots X_p) \in \mathbb{R}^p$ for $p \geqslant 1$ and $Y \in \mathbb{R}$ two random variables. Define the non-linear regression model as

$$Y = \varphi(X_1, X_2, \dots X_p) + \varepsilon, \tag{1}$$

where ε is random noise independent of $(X_1, X_2, \dots X_p)$. The unknown function $\varphi : \mathbb{R}^p \mapsto \mathbb{R}$ describes the conditional expectation of Y given $(X_1, X_2, \dots X_p)$. Suppose as well that $(X_{i1}, X_{i2} \dots X_{ip}, Y_i)$, for $i = 1, \dots, n$, is a size n sample for the random vector$(X_1, X_2 \dots X_p, Y)$.

If $p \gg n$ the model (1) suffers from the *"curse of dimensionality"*, term introduced in Bellman (1957) and Bellman (1961), where is shown that the sample size n required to fit a model increases with the number of variables p. Model selection techniques solve this problem using indicators as the AIC or BIC, or more advanced techniques such as Ridge or Lasso regression. For the interested reader there is a comprehensive survey in Hastie et al. (2009).

Suppose in the context of the linear regression we have the model

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ \vdots & & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and $\boldsymbol{\varepsilon}$ is a noisy vector with mean $(0, \dots, 0)^\top$ and identity covariance.

The least-square solutions to find the best coefficients is

$$\hat{\mathbf{b}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

If $p = 1$ the problem reduces to the equations,

$$\hat{b}_{i1} = \frac{\sum_{j=1}^{n} \left(X_{ij} - \bar{X}_i\right)\left(Y_j - \bar{Y}\right)}{\sum_{j=1}^{n} \left(X_{ij} - X_i\right)^2}$$
$$\hat{b}_{i0} = \bar{Y} - \hat{b}_1 \bar{X}_i$$

Notices that in the particular case $p = 0$ (the null model) then the estimated parameter simplifies into $\hat{b}_{0i} = \bar{Y}$.

The following identity holds in our context,

$$\sum_{j=1}^{n} \left(Y_j - \bar{Y}\right)^2 = \sum_{j=1}^{n} \left(\hat{Y}_j - \bar{Y}\right)^2 + \sum_{j=1}^{n} \left(\hat{Y}_j - Y_j\right)^2$$

One of the most used quantities to quantify if one covariate (or a set of them) are useful to explain an output variable is the statistic $R^2 \in [0, 1]$. We estimate it as

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - Y_j\right)^2}{\sum_{j=1}^{n} \left(Y_j - \bar{Y}\right)^2}$$

This value indicates how much the variability of the regression model explains the variability of Y. If R^2 is close to zero, the squared residuals of the fitted are similar to the residuals of a null model formed only for a constant. Otherwise, the residuals of the null model are greater than the residuals of the fitted values, meaning that the selected model could approximate better the observed values of the sample.

The R^2 has the deficiency that it increases if we add new variables to the model. A better statistic to measure the goodness of fit but penalizing the inclusion of nuisance variable is the Adjusted R^2,

$$R^2_{Adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

These measures could detect if a data point cloud could be fitted through some function. However, if the structure of the data has anomalous patterns R^2 could be insufficient.

For example in Figure 1 presents this phenomena. For two datasets the "Ishigami" and "Circle with one hole" models (they are presented in Section 3). We adjusted a polinomial of degree 8 to each one using least-square regression.

The Ishigami model presents strong non-linearity in their second variable. The model could capture the relevant pattern of the data and we got a $R^2 = 0.448$ and $R^2_{Adj} = 0.4435$ (panel **(c)** telling us that we could capture around the 44% of the total variability of the data. In the other cases notices how the classic regression failed to capture any of their features. In particular the R^2 and R^2_{Adj} are near to zero.

The issue with the panels **(a)**, **(b)** and **(d)** of Figure 1 is the fitted models are inflexible with respect to data used. In particular, the "Circle with one hole" model requires a better understanding about the anomalous geometry of the data cloud point. The next Section will be advocate to get a better insight in how to determine the geometric structure of the data.

1.2 Homology and Vietoris-Rips Complex

The seminal work of Carlsson (2014) presents the ground basic definitions to define a homological structure for a data cloud of points. In particular, we are interested in the reconstruction of the embedding manifold of the cloud of points using only the distance between each pair of points.

For the purpose of this paper a geometric object is either a connected surface or connected directed graph. Given a geometric object define a 0-simplex as a point, frequently called a *vertex*. Since we deal with finite sets of data, taking coordinates on the Euclidean Plane $\mathcal{E} = \mathbb{R}^2$, we denote a 0-simplex as a point $p_j = (x_j, y_j)$ for $j = 1, \ldots, n$.

If we join two distinct 0-simplices, p_0, p_1, by an oriented line segment, we get a 1-simplex called an *edge*: $\overline{p_0 p_1} = (p_1 - p_0)$.

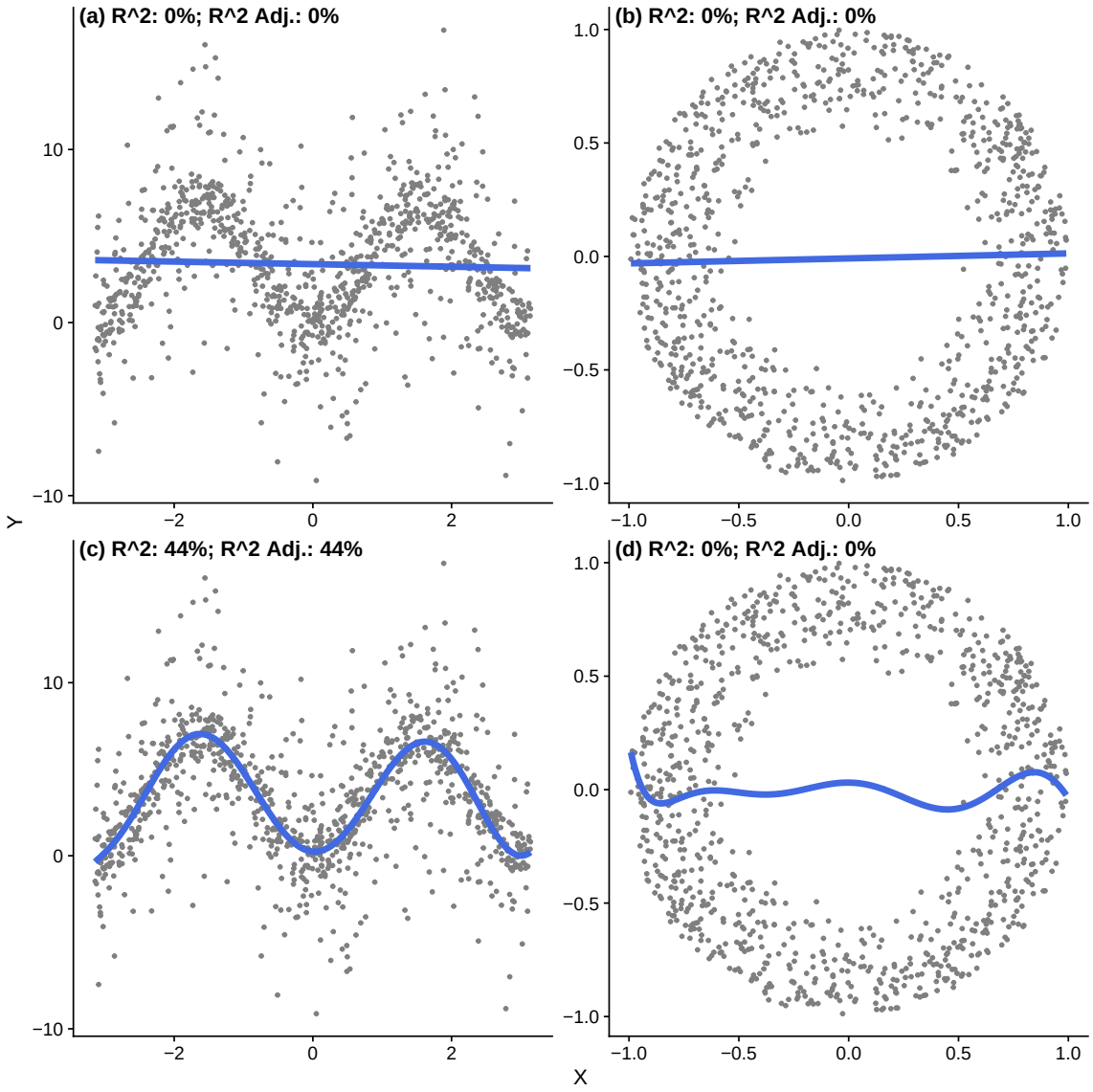


Figure 1: *Linear regression* $Y = a_0 + a_1 X$ with **(a)** Ishigami model, and **(b)** Circle with hole model. *Polynomial regression* $Y = a_0 + \sum_{k=1}^{8} a_k X_1^k$ with **(c):** Ishigami model **(d):** Circle with one hole model.

Consider now three non-colinear points $p_0,\ p_1,\ p_2$ as 0-simplices, together they form three 1-simplices: $\overline{p_0p_1}$, $\overline{p_0p_2}$ and $\overline{p_1p_2} = \overline{p_0p_2} - \overline{p_0p_1}$. This last equation shows that only two of them are linearly independent and span the other. The union of these three edges form a triangular shape, a 2-simplex called a *face*, denoted as $\triangle(p_0p_1p_2)$ that contains all the points enclosed between the edges:

$$\triangle(p_0p_1p_2) = \Big\{p \in \mathbb{R}^2 \colon p = \sum_{j=0}^{2} \lambda_j p_j \colon \sum_{j=0}^{2} \lambda_j = 1, \lambda_j \geqslant 0\Big\}.$$

Definition 1 (VR neigboorhood graph) *Given $S \subseteq \mathbb{R}^{ny}$ and scale $\varepsilon \in \mathbb{R}$, the VR neighborhood graph is a graph where $G_\varepsilon(V) = (V, E_\varepsilon(V))$ and*

$$E_\varepsilon(V) = \{\{u, v\} \ \mid \ d(u, v) \leqslant \varepsilon, u \neq v \in V\}.$$

Definition 2 (VR expansion) *Given a neigborhood graph G, their Vietoris-Rips complex $\mathcal{V}(G)$ is defined as all the edges of a simplex σ that are in G. In this case σ belongs to $\mathcal{V}(G)$. For $G = (V, E)$, we have*

$$\mathcal{V}(G) = V \cup E \cup \left\{\sigma \mid \binom{\sigma}{2} \subseteq E\right\}.$$

where σ is a simplex of G.

2. Methodology

Recall the model (1). The random variables $(X_1, \ldots, X_p)$ are distorted by the function m and its topology. Our aim is to measure how much each of the X_i influence this distortion, i.e. we want to determine which variables influence the model the most.

In Section 1.1 we review the a popular statistic to estimate the correspondence between X_i, for $i = 1, \ldots, p$, with respect to Y. In this paper, we want to consider the geometry of the point-cloud, their enveloping manifold and create an index that will reveal information about the model.

The first step is to create a neighborhood graph for the point-cloud formed by (X_i, Y) where an edge is set if a pair of nodes are within ε euclidean distance. In this way we connect only the nodes nearby within a fixed distance. With this neighborhood graph we construct the persistent homology using the method of Zomorodian (2010) for the Vietoris-Rips (VR) complex.

The algorithm of Zomorodian (2010) works in two-phases: First it creates a VR neighborhood graph (Definition 1) and then builds the VR complex step-by-step (Definition 2).

The definitions imply a two-phase procedure to construct the Vietoris-Rips complex with resolution ε:

1. Using Definition 1 compute the neighborhood graph $G_\varepsilon(V)$ with parameter ε.
2. Using Definition 2 compute $\mathcal{V}(G_\varepsilon(V))$. From now on set $\mathcal{V}_\varepsilon(V) = \mathcal{V}(G_\varepsilon(V))$.

This procedure provides us with a geometrical skeleton for the data cloud points with resolution ε. In case the parameter ε is large, there will be more edges connecting points. We could have a large interconnected graph with little information. Otherwise, if the parameter ε is small, there may be fewer edges connecting the points, resulting in a sparse graph and missing relevant topological features within the data cloud.

In the second step we unveil the topological structure of the neighborhood graph through the Vietoris-Rips complex. The expansion builds the cofaces related to our simplicial complex. In Section 2.2 we will further discuss the algorithm to achieve this.

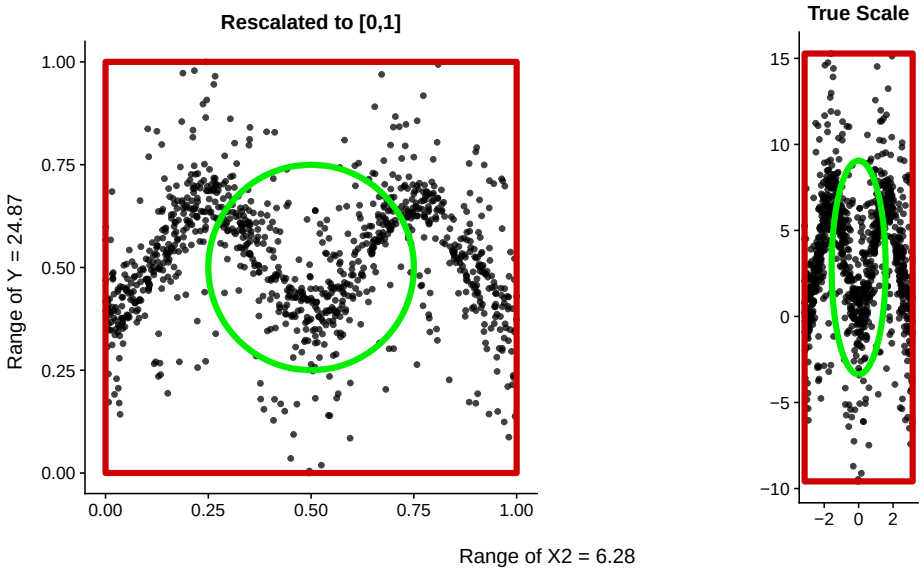


Figure 2: The Second variable of the Ishigami model scaled to $[0, 1]$ with a circle centered at $(0.5, 0.5)$ and radius 1 (left). The same circle draw at the true scale of the data (right).

2.1 Neighborhood graph

The neighborhood graph collects the vertices V and for each vertex $v \in V$ it adds all the edges $[v, u]$ within the set $u \in V$, satisfying $d(v, u) \leqslant \varepsilon$. This brute-force operation works in $O(n^2)$ time. We considered a variety of theoretical examples and it becomes clear that the scale factor in the data set is relevant. The scale in one variable may differ with the scale of the output by many orders of magnitude. Thus, proximity is relative to the scale of the axis on which the information is presented and the proximity neighborhood may be misjudged.

The Ishigami model presented in Figure 2 below shows how the proximity neighborhood becomes distorted when scale is taken into consideration.

We conclude that the aspect ratio between both variables defines how the algorithm constructs the neighborhood graph. Therefore, to use circles to build the neighborhood graph we would need to set both variables to the same scale. Algorithm 1 constructs the VR-neighborhood graph for a cloud of points with arbitrary scales.

1. Re-scale the points (X_i, Y), $i = 1, \ldots, n$, onto the square $[0, 1] \times [0, 1]$.
2. Estimate the distance matrix between points.
3. With the distance chart estimate the α quantile of the distances. Declare the radius ε_i as this quantile.
4. Using Definition 1 build the VR-neighborhood graph with ε changed by ε_i for each projection.
5. Re-scale again the data points to their original scale.

2.2 VR expansion

In Zomorodian (2010) the author describes three methods to build the Vietoris-Rips complex. The first approach builds the complex by adding the vertices, the edges and then

Data: A set of points $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$
A value $0 < \alpha < 1$.
Result: The Neighborhood Graph.

Function Create-VR-Neighborhood((X, Y), α)**:**
  $X_r \leftarrow \dfrac{X - \min X}{\max X - \min X}$
  $Y_r \leftarrow \dfrac{Y - \min Y}{\max Y - \min Y}$
  $n \leftarrow \text{length}(Y)$
  DistanceMatrix $\leftarrow$ Matrix$(n \times n)$
  for i = *1:n* **do**
    for j = *1:n* **do**
      DistanceMatrix[i,j] $\leftarrow \sqrt{(X_r[\text{i}] - X_r[\text{j}])^2 + (Y_r[\text{i}] - Y_r[\text{j}])^2}$
    end
  end
  $\varepsilon \leftarrow$ Quantile (DistanceMatrix, α)
  AdjacencyMatrix $\leftarrow$ DistanceMatrix $\leqslant \varepsilon$
  NeighborhoodGraph $\leftarrow$ CreateGraph (AdjacencyMatrix, xCoordinates $= X$, y Coordinates $= Y$)
  return (NeighborhoodGraph)
end

Algorithm 1: Procedure to estimate the neighborhood graph given a set of points in the plane.

increasing the dimension to create triangles, tetrahedrons etc. The second method starts with an empty complex and adds all the simplices step-by-step stopping at the desired dimension. In the third method one takes advantage of the fact that the VR-complex is the combination of cliques in the graph of the desired dimension.

Due to its simplicity we adopt the third approach and detect the cliques in the graph. We use the algorithm in Eppstein et al. (2010) which is a variant of the classic algorithm from Bron and Kerbosch (1973). This algorithm orders the graph G and then computes the cliques using the Bron-Kerbosch method without pivoting. This procedure reduces the worst-case scenario from time $\mathcal{O}(3^{n/3})$ to time $\mathcal{O}(dn3^{d/3})$ where n is the number of vertices and d is the smallest value such that every nonempty sub-graph of G contains a vertex of degree at most d.

Constructing a manifold via the VR-complex is not efficient in that the co-faces may overlap, increasing the computational time. One can overcome this by creating an ordered neighborhood graph.

2.3 Geometrical goodness-of-fit construction

The main use of the R^2 is to gauge the variability explained by a chosen model against a null one. In our case, the null model is the box contained all our the data contained in the Vietoris-Rips complex. This is how we envelop the data in the most basic way possible. Our model will be the Vietoris-Rips complex itself which give us a representation of our data through a clearer structure.

The patterns in the data emerge through the empty spaces in the projection space generated by each individual variable. When the point-cloud fills the whole domain then the unknown function φ applied to X_i produces erratic Y values. Otherwise, the function yields a structural pattern which can be recognized geometrically.

The VR-complex $\mathcal{V}(G)$ estimates the geometric structure of the data by filling the voids in-between close points. We may then estimate the area of the created object. This value will not give much information about the influence of the variable within the model. Therefore, we have to estimate the area of the minimum rectangle containing the entire object. If some input variable presents a weak correlation with the output variable, its behavior will be almost random with uniformly distributed points into the rectangular box. For other case, it has some relevant correlation it will create a pattern causing empty spaces to appear across the box.

To clarify the notation we will denote by $G_{\varepsilon,j}$ the neighborhood graph generated by the pair of variables (X_j, Y) and radius ε. Denote by $\mathcal{V}(G_{\varepsilon,j})$ the VR-complex generated by $G_{\varepsilon,j}$. We also denote the geometrical area of the object formed by the VR-complex $\mathcal{V}(G_{\varepsilon,j})$ by $\text{Area}(\mathcal{V}(G_{\varepsilon,j}))$.

We define the rectangular box for the projection the data (X_j, Y) as

$$B_i = \Big[\min_{X_i}(\mathcal{V}(G_{\varepsilon,j})), \max_{X_i}(\mathcal{V}(G_{\varepsilon,j}))\Big] \times \Big[\min_{Y}(\mathcal{V}(G_{\varepsilon,j})), \max_{Y}(\mathcal{V}(G_{\varepsilon,j}))\Big].$$

The geometrical area of B_j will be denoted by $\text{Area}(B_j)$.

Therefore, we can define the measure

$$R^2_{\text{Geom},j} = 1 - \frac{\text{Area}(\mathcal{V}(G_{\varepsilon,j}))}{\text{Area}(B_j)}.$$

Notice that if the areas of the object and the box are similar then the index $R^2_{\text{Geom},j}$ is close to zero. Otherwise, if there is a lot of empty space and both areas differ the index will approach 1.

3. Results

To measure the quality of the index described above we work concrete examples. The software used was *R* (R Core Team, 2019), along with the packages `igraph` (Csárdi and Nepusz, 2006) for all graph manipulations, and the packages `rgeos` and `sp` (Pebesma and Bivand, 2005; Bivand et al., 2013; Bivand and Rundel, 2013) for all the geometric estimations. A package containing all these algorithms will be available soon in CRAN.

3.1 Theoretical examples

In the examples which follow we sample $n = 1000$ points with the distribution specified in each case. Due to the number of points in every example we choose the quantile 5% to determine the radius of the neighborhood graph. Further insights into this choice will be presented in the conclusions section.

We will consider five unique settings for our examples, each one with different topological features. These settings are not exhaustive and there are others with interesting features. However through this sample we do show how the method captures the geometrical correlation of the variables where other classical methods have failed, as well as making a case for which method fails to retrieve the desired information.

The examples considered are the following:

Linear: This is a simple setting with

$$Y = 2X_1 + X_2$$

and X_3, X_4 and X_5 independent random variables. We set $X_i \sim \text{Uniform}(-1, 1)$ for $i = 1, \ldots, 5$.

Quartic: This is another simple case with link function

$$Y = X_1 + X_2^4$$

with $X_i \sim \text{Uniform}(-1, 1)$ for $i = 1, 2$.

Circle with hole: The model in this case is

$$\begin{cases} X_1 = r\cos(\theta) \\ Y = r\sin(\theta) \end{cases}$$

with $\theta \sim \text{Uniform}(0, 2\pi)$ and $r \sim \text{Uniform}(0.5, 1)$. This form creates a circle with a hole in the middle.

Connected circles with holes: The model consists in two connected pieces, where in both we set $\theta \sim \text{Uniform}(0, 2\pi)$:

1. Circle centered at $(0, 0)$ with radius between 1 and 2:

$$\begin{cases} X_1 = r_1\cos(\theta) \\ Y = r_1\sin(\theta) \end{cases}$$

where $r_1 \sim \text{Uniform}(1, 2)$.

2. Circle centered at $(1.5, 1.5)$ with radius between 0.5 and 1:

$$\begin{cases} X_1 - 1.5 = r_2\cos(\theta) \\ Y - 1.5 = r_2\sin(\theta) \end{cases}$$

where $r_2 \sim \text{Uniform}(0.5, 1)$.

Ishigami: The final model is

$$Y = \sin X_1 + 7\ \sin^2 X_2 + 0.1\ X_3^4 \sin X_1$$

where $X_i \sim \text{Uniform}(-\pi, \pi)$ for $i = 1, 2, 3$, $a = 7$ and $b = 0.1$.

In order to compare the results with our method, we fit for each case the regression $Y = a_0 + \sum_{k=1}^{10} a_k X_i^k$ for each variable X_i. Some type of regressor are possible in order to fit each case. For this paper, we only need a comparison point to determine how our method performs.

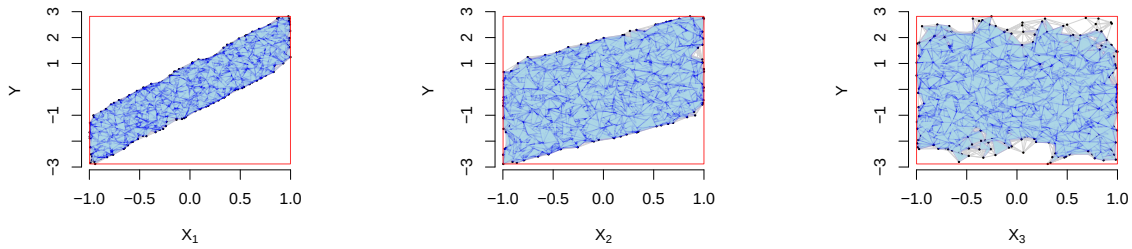
The figures presented in this section represent the estimated manifold for each input variable X_i with respect to the output variable Y. The table below each figure presents the radius used to build the neighborhood graph, the estimated areas of the manifold object, of the reference square, and the proposed index.

The linear model in Figure 3 is simple enough to allow us to directly see that the variable X_1 explains almost the double of variability than the variable X_2. Variables X_3 to X_5 will have less important indices. In this case, given the linearity, the normal R^2 covers almost 81% of the variance for the X_1. The rest of the total variance is covered by X_2. The rest of variables has depreciable amounts. We conclude how the empty spaces are present according to the relevance level of the variable. For the Quartic model in Figure 4 we observe a similar behavior as in the previous case even if the second variable has non linear pattern

The model of circle with a hole in Figure 5 was discussed in the preliminaries. Recall that in this case both variables have R^2 near to zero for our model even if the geometric shape showed the contrary. Observe how the first variable has index equal to 0.46 and the second one 0.07. This allows us to say that the VR-complex built for X_1 perform better explaining the data than only a close it in a square. For the second variable the VR-complex and the enclosed square perform similar, thus the geometric goodness-of-fit is almost zero.

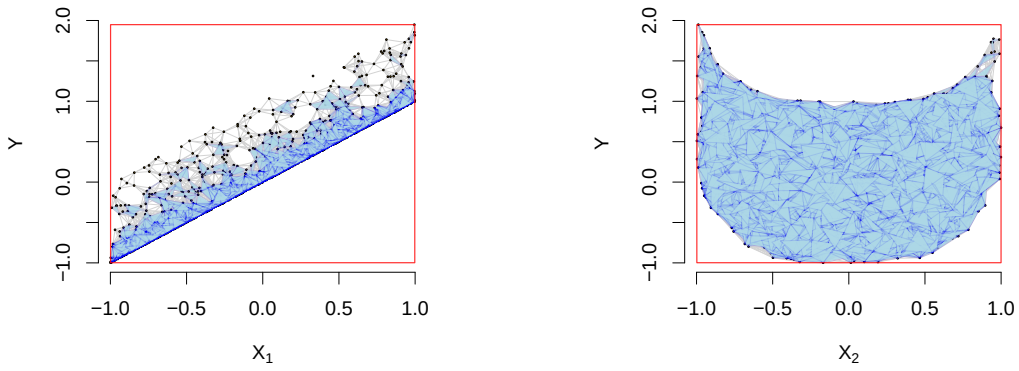
To test our algorithm further we present the model of connected circles with holes in Figure 6. Here we created two circles with different scales and positions. We could capture the most relevant features for each projection. In this case the classic R^2 could capture only a 19% of the explained variance of the model for the first variable. For the second one the R^2 is near to zero. Meanwhile, the R^2_{Geom} could detect almost a 53% of correspondence between the first variable and the outcome. In this case, our method performs better to detect the anomalous pattern.

The final model is produced by the Ishigami function, Figure 7. This is a popular model in sensitivity analysis because it presents a strong non-linearity and non-monotonicity with interactions in X_3. With other sensitivity estimators the variables X_1 and X_2 have great relevance to the model, while the third one X_3 has almost zero. For a further explanation of this function we refer the reader to Sobol' and Levitan (1999). The first, second and third variables explain 33%, 48% and 1% of the variance respectively. In particular, notices how the third variable is considered pure noise for the regression polynomial model. Also, notice how in this case, the R^2_{Geom} for the three variable are around the 50% and 60%. It indicates, for the three variables, that geometric pattern was revealed by the VR-complex. The three variables presents large areas of blank spaces inside their boxes.



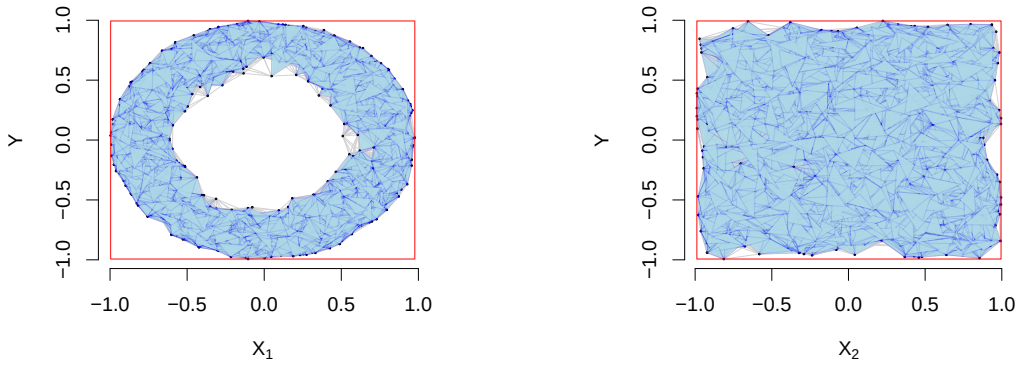
Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	R^2_{Geom}	R^2
X_1	0.08	3.79	11.38	0.67	0.81
X_2	0.11	7.69	11.39	0.32	0.19
X_3	0.12	9.78	11.39	0.14	0.01
X_4	0.12	10.18	11.38	0.10	0.01
X_5	0.12	10.16	11.39	0.11	0.01

Figure 3: Results for the linear case.



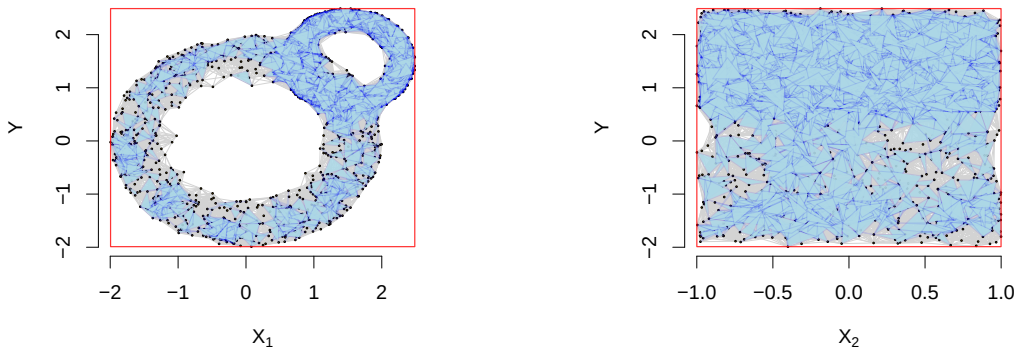
Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	R^2_{Geom}	R^2
X_1	0.06	1.50	5.88	0.75	0.82
X_2	0.11	3.83	5.89	0.35	0.19

Figure 4: Results for the quartic case.



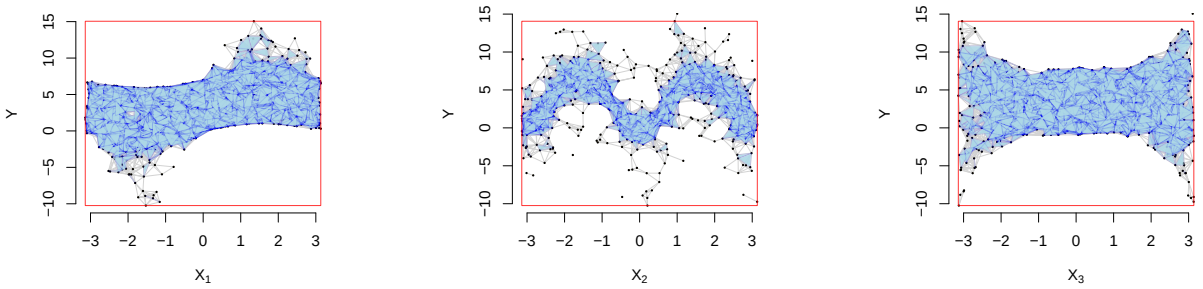
Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	R^2_{Geom}	R^2
X_1	0.11	2.10	3.92	0.46	0.01
X_2	0.13	3.67	3.94	0.07	0.01

Figure 5: Results for the circle with 1 hole case.



Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	R^2_{Geom}	R^2
X_1	0.08	9.44	20.07	0.53	0.19
X_2	0.12	8.60	8.96	0.04	0.00

Figure 6: Results for the circle with 2 holes case.



Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	R^2_{Geom}	R^2
X_1	0.08	64.83	158.65	0.59	0.33
X_2	0.07	56.78	152.50	0.63	0.48
X_3	0.09	70.96	152.59	0.53	0.01

Figure 7: Results for the Ishigami case.

3.2 Possible caveats: An Hydrology model

One academic case model which tests performance in sensitivity analysis is the dyke model. This model simplifies the 1D hydro-dynamical equations of Saint Venant under the assumptions of uniform and constant flow rate and large rectangular sections.

The following equations recreate the variable S which measures the maximal annual overflow of the river (in meters) and the variable C_p which is the associated cost (in millions of euros) of the dyke.

$$S = Z_v + H - H_d - C_b \tag{2}$$

with

$$H = \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)$$

$$C_p = \mathbf{1}_{S>0} + \left[0.2 + 0.8 \left(1 - \exp \frac{-1000}{S^4} \right) \right] \mathbf{1}_{S \leqslant 0}$$

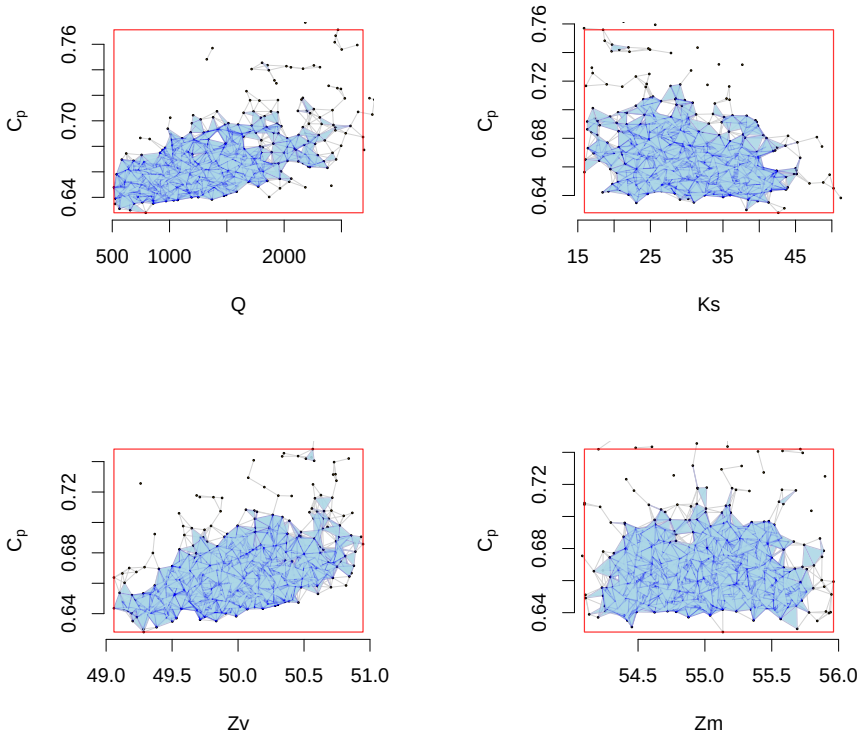
$$+ \frac{1}{20} \left(H_d \mathbf{1}_{H_d > 8} + 8 \mathbf{1}_{H_d \leqslant 8} \right) \tag{3}$$

Table 1 shows the inputs ($p = 8$). Here $\mathbf{1}_A(x)$ is equal to 1 for $x \in A$ and 0 otherwise. The variable H_d in Equation (2) is a design parameter for the dyke's height set as Uniform$(7, 9)$.

In Equation (3.2) the first term represents a cost of 1 million euros due to flooding ($S > 0$). The second term corresponds to the cost of the dyke maintenance ($S \leqslant 0$) and the third term is the construction cost related to the dyke. The latter is constant for a dyke height of less than 8m, and grows with the dyke height otherwise.

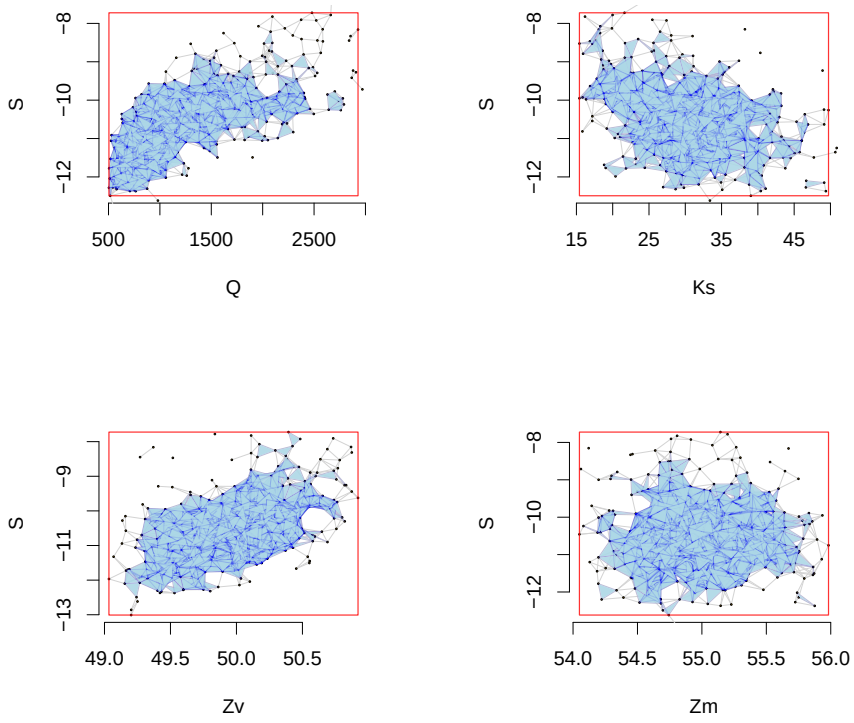
For a complete discussion about the model, parameters and meanings, the reader should see Iooss and Lemaître (2015), de Rocquigny (2006) and their references.

The work of Iooss and Lemaître (2015) detects the most influential variables for models (3.2) and (2). They use a combination of a Morris screening method, standardized



Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	R^2_{Geom}	R^2
Q	0.07	96.81	311.84	0.69	0.39
K_s	0.07	1.58	4.39	0.64	0.16
Z_v	0.08	0.09	0.23	0.61	0.24
Z_m	0.08	0.10	0.21	0.53	0.00
H_d	0.09	0.06	0.13	0.55	0.07
C_b	0.08	0.05	0.09	0.51	0.04
L	0.08	1.13	2.26	0.50	0.00
B	0.08	0.45	1.06	0.57	0.01

Figure 8: Results for the Dyke Cp case.



Variable	ε	Area($\mathcal{V}(G)$)	Area(B)	R^2_{Geom}	R^2
Q	0.08	4289.10	11554.51	0.63	0.47
K_s	0.09	80.24	163.61	0.51	0.19
Z_v	0.09	3.89	9.98	0.61	0.23
Z_m	0.09	4.77	9.46	0.50	0.00
H_d	0.11	3.01	5.13	0.41	0.10
C_b	0.09	2.35	4.71	0.50	0.04
L	0.09	46.66	83.16	0.44	0.00
B	0.09	23.11	45.34	0.49	0.00

Figure 9: Results for the Dyke S case.

Input	Description	Unit	Probability Distribution
Q	Maximal annual flowrate	m^3/s	Gumbel$(1013, 558)$ truncated on [500, 3000]
K_s	Strickler coefficient	—	$\mathcal{N}(30, 8)$ truncated on $[15, \infty)$
Z_v	River downstream level	m	Triangular$(49, 50, 51)$
Z_m	River upstream level	m	Triangular$(54, 55, 56)$
H_d	Dyke height	m	Uniform$(7, 9)$
C_b	Bank level	m	Triangular$(55, 55.5, 56)$
L	Length of the river stretch	m	Triangular$(4990, 5000, 5010)$
B	River width	m	Triangular$(295, 300, 305)$

Table 1: Input variables and their probability distributions.

regression coefficients and sobol indices. The important most correlated variables influential to the outputs C_p and S are: Q, H_d, Z_v and K_s

Figures 8 and 9 present the results of this model for the variables Q, K_s, Z_v and Z_m. For the output C_p the first three variables present a clearer geometric pattern than the others. In the case of the output S we recover the most correlated variables as Q and Z_v. The other ones do not display a clear pattern in which we could discriminate their influence, however they have values of R^2_{Geom} near to 50%. It says that our method recognize the geometric patterns of the data but no their real influence in the model.

The variable B has a higher value of R^2_{Geom} even not being correlated under the classic models. Figure 10 presents a zoomed-out result, the bounding box loosely enclosing the manifold. The reason: isolated points near to the box frontier create 2-simplexes (triangles) far away from the concentrated data. Therefore those simplexes artificially increase the area of the bounding box B. There are still improvements needing to be researched in order to create a robust version of the algorithm.

4. Conclusions and further research

As mentioned above the aim of this paper was to build a goodness of fit index relied solely on the topological and geometrical features of a given data-cloud. It is clear that purely analytic or statistical methods fail to recognize the structure within the projection of certain variables, primarily when the input is of zero sum, which might be considered artificial noise. In such cases those projections, or the variables in question, have positive conditional variance that might contribute to the model in ways that had not been explored.

Our index proved to be reliable in detecting this the variability of the data when the variable is of zero-sum, differentiating between pure random noise and well structured inputs. In the cases where the model presents pure noise our index coincides fully with other methods' indexes in detecting relevant structured inputs, in the other cases our index reflects the presence of structure in all the variables, which was the notion we wanted to explore.

We can not have yet an efficient and transparent index that measure the geometric goodness-of-fit through the VR-complex. To reach this point we identify a series of research problems to be dealt with in the near future: Improving the algorithm for the construction of the base graph. Alternatively, change it to make the process more efficient and hence allow ourselves to run more sophisticated examples, both theoretical and real data examples from controlled examples. One of the central points to be discussed and studied further is determining the radius of proximity, which we believe must be given by the data-set itself,

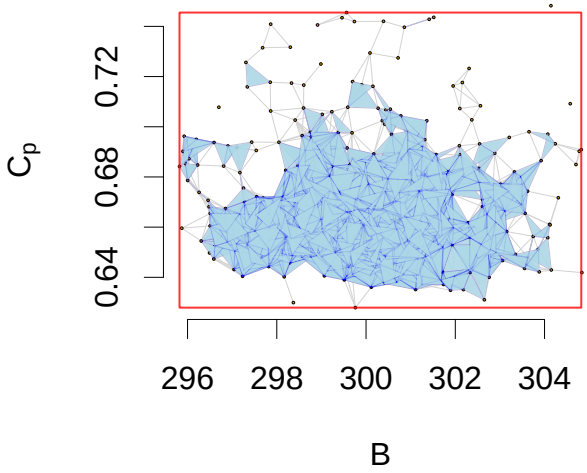


Figure 10: Scaling issue with irrelevant variables. The VR-complex spreadout causes bounding boxes larger than the geometrical structure of the estimated manifold.

probably by a more detailed application of persistent homology. Finally, we look forward to extending our method to more than one variable at a time, to be able to cross-check relevance between variables.

As stated we do not claim ours to be an exhaustive list of problems related to the improvement of our method, but are confident that they will help us run more examples, and for them to be more sophisticated, as well as helping us get more data to compare our results with other methods.

One area that we will be focused in the near future is to explore if using these constructions, we can achieve to determine if one variable is really relevant to model. It means, not only say that the variable has a geometric structure, but also say if the there is a noisy structured pattern or there is a correlated effect between input and output.

Acknowledgments

We acknowledge Santiago Gallón for enlightening discussions about the subject, whose help has been very valuable.

The first and second authors acknowledge the financial support from CIMPA, Centro de Investigaciones en Matemática Pura y Aplicada through the projects 821–B7–254 and 821–B8–221 respectively.

The third author acknowledges the financial support from CIMM, Centro de Investigaciones Matemáticas y Metamatemáticas through the project 820–B8–224.

The three authors also acknowledge Escuela de Matemática, Universidad de Costa Rica for its support.

Supplementary Material

R-package for TopSA routine: R-package "TopSA" estimates sensitivity indices reconstructing the embedding manifold of the data. The reconstruction is done via a Vietoris Rips with a fixed radius. Then the homology of order 2 and the indices are estimated. `https://github.com/maikol-solis/topsa`

References

M. Balasubramanian. The Isomap Algorithm and Topological Stability. *Science*, 295(5552):7a–7, jan 2002.

James P. Barrett. The coefficient of determination-some limitations. *American Statistician*, 28(1): 19–20, feb 1974.

A. P. Barten. Note on unbiased estimation of the squared multiple correlation coefficient. *Statistica Neerlandica*, 16(2):151–164, jun 1962.

R. Bellman. *Dynamic Programming*. Dover Books on Computer Science Series. Princeton University Press, 1957. ISBN 978-0-486-42809-3.

R. Bellman. *Adaptive control processes: A guided tour*, volume 4 of *'Rand Corporation. Research studies*. Princeton University Press, 1961.

Mira Bernstein, Vin de Silva, John C. Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds. *Igarss 2014*, 01(1):1–5, 2000.

Roger Bivand and Colin Rundel. rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-2. *R package version 0.1-8*, page 61, 2013.

Roger Bivand, Edzer J Pebesma, Virgilio Gómez-Rubio, and Corporation Ebooks. Applied spatial data analysis with R. 10(Book, Whole), 2013.

Coen Bron and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973.

Andreas Buja, Dianne Cook, Daniel Asimov, and Catherine Hurley. Computational Methods for High-Dimensional Rotations in Data Visualization. pages 391–413. 2005.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, jan 2009.

Gunnar Carlsson. Topological pattern recognition for point cloud data. *Acta Numerica*, 23(23): 289–368, may 2014.

J.S. Cramer. Mean and variance of R2 in small and moderate samples. *Journal of Econometrics*, 35(2-3):253–266, jul 1987.

Gábor Csárdi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal Complex Systems*, 1695(1695):1695, 2006.

Étienne de Rocquigny. La maîtrise des incertitudes dans un contexte industriel. 1re partie : une approche méthodologique globale basée sur des exemples. *Journal de la société française de statistique*, 147(3):33–71, 2006.

Chloé Dimeglio, Santiago Gallón, Jean Michel Loubes, and Elie Maza. A robust algorithm for template curve estimation based on manifold embedding. *Computational Statistics and Data Analysis*, 70:373–386, feb 2014.

David Eppstein, Maarten Löffler, and Darren Strash. Listing All Maximal Cliques in Sparse Graphs in Near-Optimal Time. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6506 LNCS, pages 403–414. 2010. ISBN 3642175163.

Santiago Gallón, Jean-Michel Loubes, and Elie Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129–142, apr 2013.

Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, oct 2008.

David J. Hand. Pattern Recognition. In C.R. Rao, E.J. Wegman, and J.L. Solka, editors, *Handbook of Statistics*, pages 213–228. Elsevier B.V, volumen 24 edition, 2005.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, volume 1 of *Springer Series in Statistics*. Springer-Verlag New York, New York, NY, 2009. ISBN 978-0-387-84857-0.

Bertrand Iooss and Paul Lemaître. *A Review on Global Sensitivity Analysis Methods*, pages 101–122. Springer US, Boston, MA, 2015. ISBN 978-1-4899-7547-8.

Glenn J. Myatt and Wayne P. Johnson. *Making Sense of Data II*. John Wiley & Sons, Inc., Hoboken, NJ, USA, jan 2009. ISBN 9780470417409.

E.J. Pebesma and R. Bivand. Classes and methods for spatial data in R. *The Newsletter of the R Project*, 5(2):9–13, 2005.

S.James Press and Arnold Zellner. Posterior distribution for the multiple correlation coefficient with fixed regressors. *Journal of Econometrics*, 8(3):307–321, dec 1978.

R Core Team. R: A language and environment for statistical computing, 2019.

Václav Snášel, Jana Nowaková, Fatos Xhafa, and Leonard Barolli. Geometrical and topological approaches to Big Data. *Future Generation Computer Systems*, 67:286–296, feb 2017.

I. M. Sobol' and Yu.L. Levitan. On the use of variance reducing multipliers in Monte Carlo computations of a global sensitivity index. *Computer Physics Communications*, 117(1-2):52–61, mar 1999.

J. B. Tenenbaum. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, dec 2000.

Edward R. Tufte. *The visual display of quantitative information*. Graphics Press, 2001. ISBN 9780961392147.

Afra Zomorodian. Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3): 263–271, jun 2010.