Topological Sensitivity Analysis Using R

Maikol Solís[1,2], Alberto Hernández[1,3], and Ronald Zuñiga[1,4]

[1] Universidad de Costa Rica,
Centro de Investigación en Matemática Pura y Aplicada (CIMPA)
Escuela de Matemática
[2] Email: maikol.solis@ucr.ac.cr
[3] Email: albertojose.hernandez@ucr.ac.cr
[4] Email: ronald.zunigarojas@ucr.ac.cr

Abstract. The aim of our investigation is to use R to create a tool to estimate the relevance of random variables within a model. The algorithm will extract the geometric information from a data cloud by exploiting its topological features. It will reveal the reconstruction of the corresponding enveloping manifold.

Keywords: Sensitivity Analysis, Topological Data, Persistent Homology, Betti Numbers, Simplicial Complex

Overview

Let be $(X_1, X_2, \dots X_p) \in \mathbb{R}^p$ for $p \geq 1$ and $Y \in \mathbb{R}$ two random variables. Define the model,

$$Y = m(X_1, X_2, \dots X_p). \tag{1}$$

The unknown function $m : \mathbb{R}^p \mapsto \mathbb{R}$ describes the conditional expectation of Y given $(X_1, X_2, \dots X_p)$. Suppose also that $(X_{i1}, X_{i2} \dots X_{ip}, Y_i)$ for $i = 1, \dots, n$ is a sample of size n of the random vector$(X_1, X_2 \dots X_p, Y)$. If $p \gg n$ we will require to increase n exponentially to fit the model.

One option to select the variables is to rank them according to their impact inside the model. In [1] different techniques to estimate such indicators are studied. Cited examples are: parametric and non-parametric settings, variance-based measures, moment independent measures, and graphical techniques.

Using topological tools, we aim to exploit the geometrical information from clouds of random and functional data to build the afore mentioned sensitivity indicators.

The Model

Consider now the model from Equation (1). Our method creates the neighbourhood graph of the cloud formed by (X_i, Y). An edge connects two points if they are "close enough" with respect to the euclidean distance. This graph gives us the skeleton of the geometric structure of the data.

Then, we construct in R the persistent homology using the method of [2] to get the Vietoris-Rips Complex. With this, we estimate the area of the enveloping manifold of the data. To normalize this value, we estimate the area of the minimum rectangle containing the whole object.

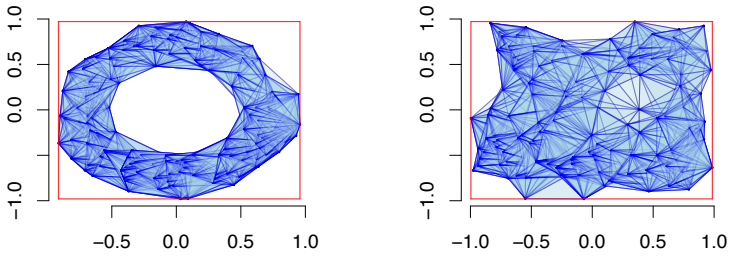
Strong influential variables have many empty spaces inside their reference box, implicating a clear pattern. Otherwise, weak influential variables have an almost random behavior inside their reference box. We can measure this difference by

$$S_i^{\text{Area}} = 1 - \frac{\text{Object Area for variable i}}{\text{Box Area for variable i}}.$$

Therefore, strong influential variables will have S_i^{Area} near to 1 (100%) and otherwise will be near to 0.

For example, suppose that $n = 150$, and generate the following random variables $r \sim \text{Uniform}(0.5, 1)$ and $\theta \sim \text{Uniform}(0, 2\pi)$. Set the variables $X_1 = r\cos(\theta)$, $Y = r\sin(\theta)$ and $X_2 \sim \text{Uniform}(0, 1)$ as a noise input. We want to explore the model $Y = m(X_1, X_2)$.

As we expect, the empty spaces for variable X_1 represent 51% meaning more influence into the model against the 18% for X_2 (see table below).



Variable	Manifold Area (Blue)	Square Area (Red)	S_i^{Area}
X_1	1.85	3.77	0.51
X_2	3.12	3.80	0.18

We build a package in R that estimates these indices using the persistent homology of the proximity graph of a given cloud of points.

References

1. Wei, P., Lu, Z., Song, J.: Variable importance analysis: A comprehensive review. Reliability Engineering and System Safety **142**, 399–432 (oct 2015)
2. Zomorodian, A.: Fast construction of the Vietoris-Rips complex. Computers & Graphics **34**(3), 263–271 (jun 2010)