Topological Sensitivity Analysis using R

Maikol Solís

@maikol_solis

maikol-solis

joint work with:

Alberto Hernández & Ronald Zúñiga

UNIVERSIDAD DE COSTA RICA

@UniversidadCR

CIMPA Centro de Investigación en Matemática Pura y Aplicada

@CIMPAUCR

ESCUELA DE MATEMÁTICA UNIVERSIDAD DE COSTA RICA

@EscuelaMateUCR

LATINR

Buenos Aires, Argentina

September 5th, 2018

Outline

Motivation

Assume that $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^p$ produces the output $Y \in \mathbb{R}$ linked by the model

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

The function m could be known or unknown.

Generally is a complex function.

Questions

- ¿How sensitive is each input with respect to Y?
- ¿How much changes the output Y if there exist a perturbation in the inputs?

Classic methods to study sensitivity

- Screening
- Regression and Correlation Analysis
- Sobol indices (ANOVA)
- Moment independent indices (based in monotonic invariant transformations)
- among others.

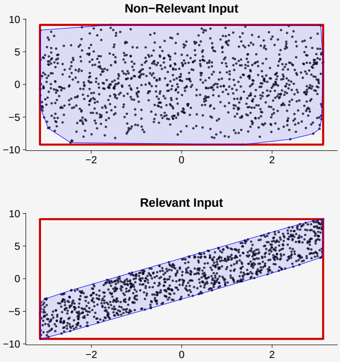
Comprehensive review

P. Wei, Z. Lu, and J. Song. 2015. Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety* 142:399–432

Our first hypothesis

If X_i is non-relevant with respect to Y then blue area is similar to the red box

Otherwise, the blue area does not cover all the domain.



The sensitivity index is $S_i^{\text{Area}} = 1 - \dfrac{\text{Object Area for variable i}}{\text{Box Area for variable i}}$.

Simplicials

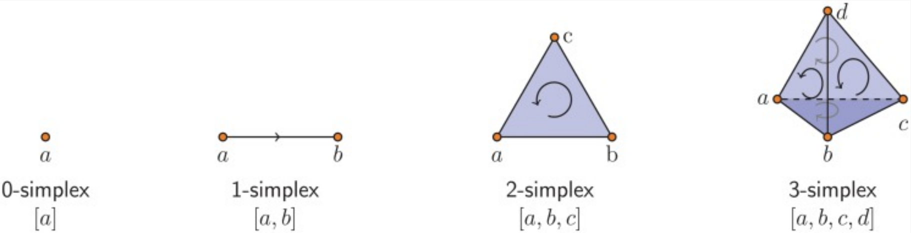


Image taken from: C. M. Topaz, L. Ziegelmeier, and T. Halverson. 2015. Topological Data Analysis of Biological Aggregation Models:1–26

Vietoris-Rip complex

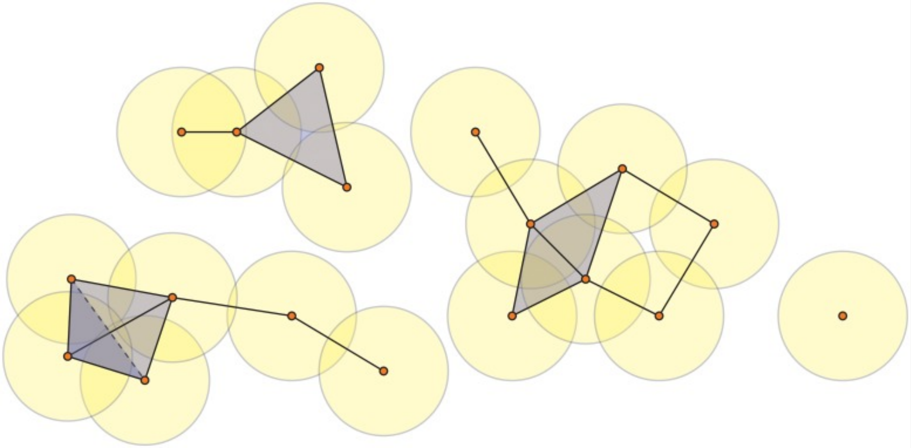


Image taken from: C. M. Topaz, L. Ziegelmeier, and T. Halverson. 2015. Topological Data Analysis of Biological Aggregation Models:1–26

Outline

The 0-simplex

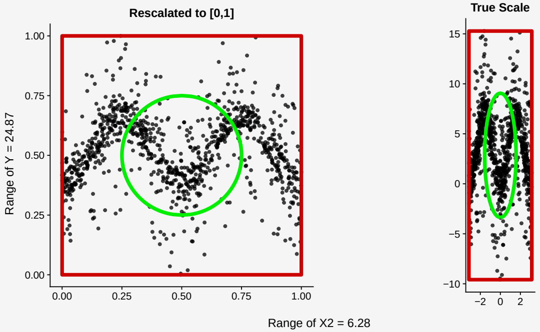
Suppose that we have a sample (X_{ki}, Y_k) where $k = 1, \ldots, n$ (observations) and $i = 1 \ldots, p$ (variables). **From now, assume that the variable i is fixed.**

0-simplex: The data points with coordinates (X_{ki}, Y_k).

The 1-simplex

To preserve the dimensions of the x and y axis, we rescaled all the data-points to the square $[0,1] \times [0,1]$.

1-simplex: With the rescaled variables ($[0,1] \times [0,1]$), construct pairwise the edges of the data points with distance less to r.



Setting the 0-simplex and 1-simplex

For some radius r given.

```r
Xr <- scales::rescale(X, to = c(0, 1))
Yr <- scales::rescale(Y, to = c(0, 1))
neigborhood.distance <- dist(cbind(Xr[, i], Yr))

adjacency.matrix <- (as.matrix(neigborhood.distance) <= r)
diag(adjacency.matrix) <- 0

graphBase <- igraph::graph.adjacency(adjacency.matrix,
                                     mode = "undirected",
                                     weighted = NULL)
```

The 2-simplex

Using the framework of Zomorodian[1], we estimate the 2-simplex as the cliques of dimension 3 of a graph using the package igraph,

```
clq <- igraph::cliques(graphBase, min = 3, max = 3)
TwoSimplex <- do.call("rbind", clq)
```

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

Transforming the 2-complex to Polygons

Each 2-simplex is a Triangle which is stored into a `Polygon` of the package `sp`:

```
p <- sp::Polygon(Triangle, hole = FALSE)
```

With all the Polygons we form a list `l` and then create a `SpatialPolygons` object:

```
sps <- sp::SpatialPolygons(list(sp::Polygons(l,1)))
```

Index estimation

Using the packages rgeos we estimate the Area of the all 2-simplex and the box contained all the points.

```
ObjArea <- rgeos::gArea(sps)
bb <- sp::bbox(sps)
SqArea <- prod(diff(t(bb)))
```

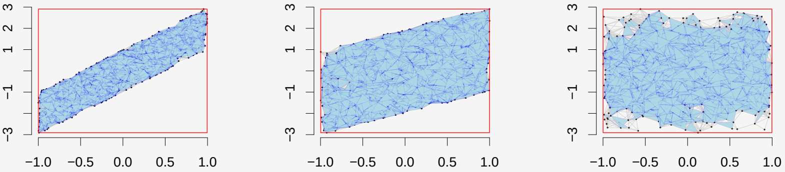
Finally, the sensitivity index is:

```
index = 1 - ObjArea / SqArea
```

Outline

Linear

The model is $Y = 2X_1 + X_2$ and X_3, X_4, X_5 variables of pure noise.



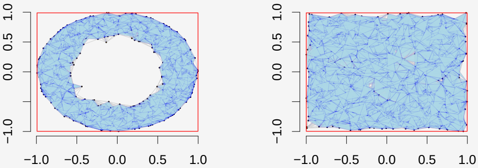
Variable	Radius	Obj Area	Square Area	Index
X_1	0.08	3.78	11.61	0.67
X_2	0.11	7.67	11.61	0.34
X_3	0.12	10.05	11.61	0.13
X_4	0.12	10.19	11.59	0.12
X_5	0.12	10.24	11.62	0.12

Circle with hole

The model is

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

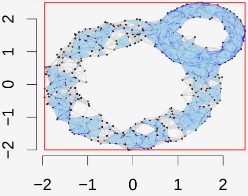
with r and θ random. The variable X_2 is pure noise.

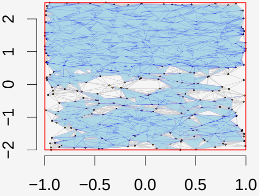


Variable	Radius	Obj Area	Square Area	Index
X_1	0.10	2.05	3.95	0.48
X_2	0.13	3.73	3.97	0.06

Circle with two holes

Similar to the last one with two circles.

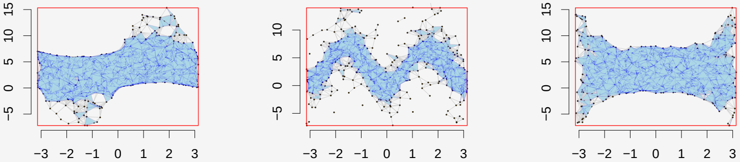




Variable	Radius	Obj Area	Square Area	Index
X_1	0.08	8.33	19.98	0.58
X_2	0.08	8.27	8.97	0.08

Ishigami

The model is determined by $Y = \sin X_1 + 7\ \sin^2 X_2 + 0.1\ X_3^4 \sin X_1$ where $X_i \sim \mathrm{Uniform}(-\pi, \pi)$ for $i = 1, 2, 3$, $a = 7$ and $b = 0.1$.



Variable	Radius	Obj Area	Square Area	Index
X_1	0.09	64.06	141.44	0.55
X_2	0.07	54.34	133.08	0.59
X_3	0.09	69.31	141.42	0.51

Summary and future work

- We could capture geometric structures of the projection of each variable.
- Our method match the classic theories in certain cases.
- When the data has a *zero-sum* pattern our method fails.
- We conjecture the method could recognize structured and non-structured noisy variables.

Future work:

- Estimate efficiently the neighborhood graph (less than $\mathcal{O}(n^2)$).
- Improve the algorithm to recognize relevant variables and structured noise variables.
- Improve the radius choice.
- Submit this package to CRAN.

Maikol Solís
@maikol_solis
maikol-solis

joint work with:

Alberto Hernández & Ronald Zúñiga

Preprint

A. Hernández, M. Solís, and R. Zúñiga. 2018. Sensitivity Analysis with Manifolds. *ArXiv e-prints, 1809.00669*