Computational Statistics (2023) 38:1231–1253
https://doi.org/10.1007/s00180-022-01244-1

ORIGINAL PAPER

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

Alberto J. Hernández[1] · Maikol Solís[1]

Received: 8 January 2021 / Accepted: 23 May 2022 / Published online: 23 June 2022

Abstract
In this work, we derive a geometric goodness-of-fit index similar to R^2 using geometric data analysis techniques. We build the alpha shape complex from the data-cloud projected onto each variable and estimate the area of the complex and its domain. We create an index that measures the difference of area between the alpha shape and the smallest squared window of observation containing the data. By applying ideas similar to those found in the closest neighbor distribution and empty space distribution functions, we can establish when the characterizing geometric features of the point set emerge. This allows for a more precise application for our index. We provide some examples with anomalous patterns to show how our algorithm performs.

Keywords Goodness-of-fit · Alpha shape complex · Empty space distribution · Closest neighbor distribution · Area estimation · CSR process

Mathematics Subject Classification 62B86 · 52C35 · 55N35

1 Introduction

Data point cloud recognition is an essential task in any statistical procedure. Finding a pattern into the data exposes the inherent characteristics of the phenomena we want to model. Tools used to achieve this are described in the literature, like linear regression or clustering (Hastie et al. 2009). Other research branches use data visualization techniques to highlight features hidden in the data (Tufte 2001; Myatt and Johnson 2009; Buja et al. 2005).

✉ Maikol Solís
maikol.solis@ucr.ac.cr

Alberto J. Hernández
albertojose.hernandez@ucr.ac.cr

[1] Centro de Investigación en Matemática Pura y Aplicada (CIMPA), Escuela de Matemática, Universidad de Costa Rica, San Jose, Costa Rica

Professionals in computational modeling aim to reduce their models by choosing their problems' most relevant factors. One way to achieve this is through goodness-of-fit measures, used to find the relevance of certain chosen models to explain a variable. One classic way to determine whether some variables fit inside a model is using the determination coefficient R^2. This quantity measures the amount of variance explained by some model against the variance explained by the model formed only by a constant. It measures how preferable it would be to fit a model against fitting a constant. If R^2 is nearly one, then the model fits well into the data. Otherwise, when R^2 is closer to zero, it is better to adjust to a constant.

The R^2 method has been controversial since its origins (Barrett 1974). For example, we can increase the R^2 score by adding new variables to the model. Even if they are unnecessary to the problem the R^2 will increase. As a general rule, high R^2 does not imply causation between covariance and outcome.

Some authors have proposed some extensions and properties of this score throughout the years. The work of Barten (1962) presents a bias-reduced R^2. In Press and Zellner (1978) they conducted a Bayesian analysis. In Cramer (1987) the authors showed that in small samples, the R^2 score is higher. Even with those constraints, Barrett (1974) concludes how useful these measures are in applied research.

These goodness-of-fit measures 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 not recognize the geometric structure when the input variable is zero-sum, treating it as random noise.

Two classical methods used to describe 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 euclidean distance between points but maximizing the total variance of points. MDS extends this method to any metric space. Methods like the ISOMAP algorithm developed by Tenenbaum (2000) and expanded by Bernstein et al. (2000) and Balasubramanian (2002) unify these two concepts to allow reconstruction of a low-dimensional variety for non-linear functions. Using geodesic distance, ISOMAP identifies the corresponding manifold and searches lower dimensional spaces.

Recently, new theoretical developments have used ideas in geometry for data analysis. A well-studied tool to find patterns within the data comes from a spatial point analysis, whether it is from visualization or analytical techniques, such as in Baddeley et al. (2016).

In this work we connect the concept of goodness of fit with geometric analysis of data through a geometrical R^2 index. By doing this, it will be possible to determine what variables have structured patterns using the geometric information extracted from the data. Also, we will be able to detect those patterns even for non-correlated and noisy variables.

For this aim, in this paper we use an alpha shape construction as a proxy of the geometric information in the data. These objects are straightforward to build and we can use their area properties to build our index. Interesting applications to alpha shapes have been developed in the literature. For example, determining the complexity in 3D shapes (Gardiner et al. 2018), modeling human faces (Bouchaffra 2012), and exploring hydrological models (Guerrero et al. 2013).

The outline of this study is as follows: Section 2 deals with key notions, both in goodness-of-fit analysis and geometric tools for pattern recognition. In Sect. 2.1 we review and comment on classic methods to determine the R^2. We finish this Sect. with an example that motivated the work in this paper. Section 2.2 describes the alpha shape complex. Section 2.3 deals with spatial functions, in particular nearest neighbor and empty Space distribution functions. Section 2.4 deals with the concept of intensity of a distribution and defines the Complete Spatial Randomness Process (CSR). Section 3 explains the method used to create our sensitivity index; Section 3.1 constructs 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. It also compares our proposed method with spatial point analysis, in particular analyzes how our index relates to the empty space distribution function. Section 3.2 deals with hypothesis testing and Monte-Carlo envelopes. Section 4 describes our results, it describes the software and packages used to run our theoretical examples. Section 4.1 is a full description of each theoretical example with visual aids, such as graphics and tables describing the results. Section 5 contains our conclusions and explores scenarios for future research.

2 Preliminary aspects

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

2.1 Measuring goodness-of-fit

Let $(X_1, X_2, \dots, X_p) \in \mathbb{R}^p$ for $p \geq 1$ and $Y \in \mathbb{R}$ two random variables. Define the 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 $(X_{k1}, X_{k2} \dots X_{kp}, Y_k)$, for $k = 1, \dots, n$, 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

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

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 are

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

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

$$\hat{b}_{1i}=\frac{\sum_{k=1}^{n}\left(X_{ki}-\bar{X}_i\right)\left(Y_k-\bar{Y}\right)}{\sum_{k=1}^{n}\left(X_{ki}-\bar{X}_i\right)^2}$$
$$\hat{b}_{0i}=\bar{Y}-\hat{b}_1\bar{X}_i$$

Notice that in the particular case $p=0$ (the null model) 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) is 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 variables are similar to the residuals of a null model formed only by a constant. Otherwise, the residuals of the null model are greater than the residuals of the fitted values, meaning that the selected model does better at approximating 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 variables is the Adjusted R^2,

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

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

For example Fig. 1 presents this phenomenon for two data sets. We adjusted thin plate regression splines to each set (Wood 2006). The Ishigami model presents strong non-linearity in its second variable. The model was able capture the relevant pattern of the data and we got a $R^2 = 0.42$ (panel **(c)**). This tells us that we captured around 42% of the total variability of the data. In panels **(a)**, **(b)** and **(d)** the R^2 is near to zero. The chosen model is inflexible to the geometric pattern in the data. In particular, the "Doughnut" model requires a better understanding of the anomalous geometry of the data cloud. The next section explores how to better determine the geometric structure of the data to overcome this shortcoming.

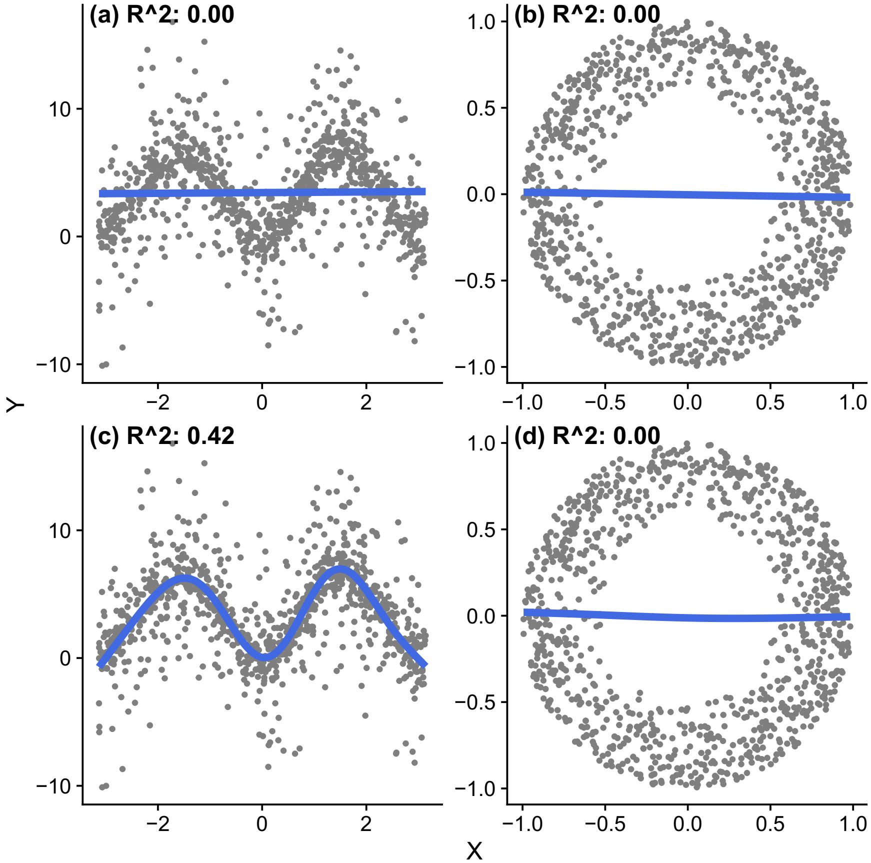

Fig. 1 *Linear regression* $Y = a_0 + a_1X + \varepsilon$ with **a** Ishigami model, and **b** Doughnut model. *Spline regression,* $Y = g(X_1) + \varepsilon$ where g is a thin-plate spline smoother. **c:** Ishigami model **d:** Doughnut model

2.2 Alpha shapes

The base of our algorithm is the construction of alpha shapes in the data. These objects are a continued version of the data-points set extending its features. In this section we will define some specific concepts necessary to develop our algorithm.

First let S be a finite set of points in $\mathbb{R}^2$. For each point s, we define the set

$$V_s = \left\{ u \in \mathbb{R}^2 \mid \|u - s\| \leq \|u - t\|, \forall t \in S \right\}$$

where $\| \cdot \|$ is the Euclidean norm. The set V_s is called the *Voronoi region* of s. The set of points that satisfy V_s are the intersection of finitely many half-planes and therefore form a convex polygon. We call the *Voronoi diagram* of S the set of Voronoi regions V_s.

Given a Voronoi diagram of $S \subset \mathbb{R}^2$, the *Delaunay triangulation* consists of connecting two points by a straight line if their respective Voronoi regions share a common boundary. Three Voronoi regions can intersect in a point. Here, there are three pairwise intersections in the Voronoi regions creating a triangle. Some regularity conditions must always have a set of triangles. They can be reviewed in Edelsbrunner (2014).

Now, for $\alpha > 0$, define as $D_s(\alpha)$ the closed disk with center s and radius α. For each site $s \in S$ we define $R_s(\alpha) = D_s(\alpha) \cap V_s$ and note that this is a convex set because it is the intersection of convex sets. Also, define $U_S(\alpha) = \bigcup_{s \in S} R_s(\alpha)$. Here, U_S is a subset of the Voronoi diagram of S. In the same way we defined the Delaunay triangulation, we now connect two points with a straight edge if they share a boundary in the $U_S(\alpha)$ set. We denote this new triangulation as $A_S(\alpha)$ or $A(\alpha)$ if the data points are understood. Two neighboring points are separated for at most 2α units. Any point beyond this value is considered *far away* and they are excluded from the triangulation. The set of triangles $A(\alpha)$ is known as the *alpha complex* of S. The union of the alpha complexes as α increases is the *alpha shape* of S.

Figure 2 presents an example of the alpha complex and its respective shape of the model of a circle with a hole at different α values. As α grows, we can pass from a shattered shape ($\alpha = 0.04$) to a unified figure without internal information ($\alpha = 0.6$).

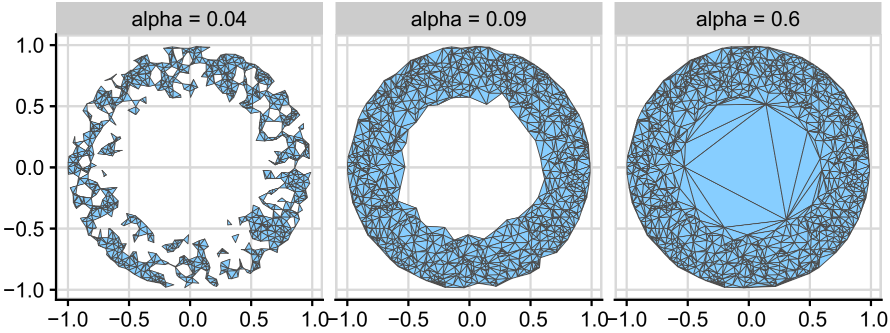

Fig. 2 Alpha shape construction as α increases

At some point, we capture the right amount of information to describe the real object ($\alpha = 0.09$). The changes in the empty space, when α changes, can allow us to detect if some variable is correlated with others.

Note that in our setting the distance r between connecting points is set to be equal to $\alpha/2$.

2.3 Spacing functions on data sets

For the current section we refer to Baddeley et al. (2016). As it is well known, a major motivation for analyzing point patterns in data is to determine whether the points are placed independently of each other or whether there is some interdependence between them. Classically, we use correlation as the statistical tool for determining dependence, it is reasonably easy to calculate, and it is a powerful tool.

The statistical correlation has its shortcomings, it is a number that summarizes statistical associations, it cannot characterize dependence or casualty, for example it does not allow discriminating between different plausible causes for clustering.

One can gather extra information about the point pattern by measuring spacing and shortest distances within a set. The mathematical theory tells us that the shortest distances in a pattern provide information complementary to the correlation structure. Assume S to be a data set on a plane, s_l elements of such set. Basic functions for such measures are the following:

1. **Pairwise distance:** $d_{lm} = ||s_l - s_m||$, between two different points s_l, s_m.
2. **Nearest neighbor distance:** $d_l = \min_{l \neq m} d_{lm}$
3. **Empty space distance:** $d(u) = \min_l ||u - s_l||$, where u is an arbitrary location within the plane that contains the dataset.

The nearest neighbor and empty space distance have several applications in field observations, such as in seed dispersion and natural distribution of trees of a given species within a forested area (Baddeley et al. 2016, Chapter 8).

Graphic techniques allow for visual exploration of data patterns. An example would be Dirichlet tiles, which are convex polygons that arise from the nearest neighbor function.

Since there is valuable information within the Nearest neighbor and empty space distance, one would like to condense or code its whole bulk within a single function, this would allow for big picture approach to the information conveyed within each function cumulatively.

Recall that a point process is called *stationary* when we view a section of the data, and its statistical properties do not depend on the section chosen.

Assume a stationary process for a data set S, then the *cumulative distribution function of the empty space distance* is given by

$$F(r) = \mathbb{P}\{d(u) \leq r\}$$

where u is an arbitrary location within the window that contains the dataset and $r \geq 0$. The values of F are probabilities that increase as r increases. For this, such processes F are always differentiable.

Again assuming a stationary process and a dataset S the *cumulative distribution of the neighbor distance function* is given by

$$G(r) = \mathbb{P}\{d_i \leq r\}$$

The values of G are probabilities and are non-decreasing as a function of r since the process is stationary the definition is not dependent on the location on which we are measuring each time. In general G is not differentiable and hence it might not have a probability density.

2.4 Poisson processes and intensity

This section refers to Baddeley et al. (2016) chapter five. We can characterize a *Complete Spacial Randomness* process (CSR), also known as a *homogenous Poisson point process* by the *homogeneity* where the points lack spatial preference, and *Independence* where the information conveyed within one region of space has no influence over the information conveyed in other regions. The CSR processes are important for a variety of reasons, since they appear naturally as physical phenomena, such as radioactivity and signal scattering. In this sense they are a good model for pure noise.

Both conditions together, homogeneity and independence, imply that the number of points falling within a sub-region of the observation window W follows a Poisson distribution.

The homogeneity assumption means that the expected number of points that fall within a specific sub-region B of W should be a function of its area $|B|$. This means

$$\mathbb{E}[n(S \cap B)] = \lambda |B| \tag{2}$$

where λ is the intensity of the process.

The independence assumption means that there is an orderly condition, implying the probability of two points lying on top of each other is negligible.

CSR processes are usually fixed to be null models in statistics to be contrasted against.

In data analysis the inference of the intensity of the point pattern is primordial, it is a basic characteristic of the process being studied. We are interested in homogeneity: the intensity of data points is not a function of spatial location, this is expressed in Equation (2).

The intensity λ then is the expected number of points per unit of area. The empirical intensity would then be given by the following:

$$\overline{\lambda} = \frac{n(S)}{|W|} \tag{3}$$

With $n(S)$ is the number of points of our dataset S and W is the window of observation. Since for a Poisson process the variance equals the mean we have that

$$\mathrm{Var}\left(\overline{\lambda}\right) = \frac{\lambda}{|W|}.$$

3 Methodology

Recall the equation (1). The model function φ distorts the random variables $(X_1, \dots, X_p)$ into a new shape. Our aim is to measure how much each of the X_i influence this distortion, i.e. we want to determine which variables were most affected by the distortion. To this end, we will build a curve representing the space filling process inside the squared domain of the data. The curve will spotlight changes over different radii scales, allow us to quantify the effect of the data transformation.

We will use the ideas developed in Sect. 2.2 about alpha-shapes construction. The alpha-shape geometry performs the filling as the radius α grows. Besides, we can infer the geometry of the data through the behavior of this curve.

Recall that an alpha-shape is a subset of the triangulation of Delanauy. We can interpret this object as a continued version of the data-points given a fixed α. While α is small, we observe sparse patches depending on the concentration of data in the plane. As α becomes larger those patches also become larger, and will eventually unite to form one large polyhedron. During the triangulation process the geometric pattern of the data emerges.

For each construction, we will estimate the geometric correlation between the two-dimensional point clouds. This index is based on the proportion of empty space between the domain and the alpha-shape.

Finally, we estimate the geometric correlation for each α. This will generate a curve that unveils the large features of the data.

3.1 Geometrical goodness-of-fit construction

In this study, the null model is the box containing a CSR process of homogenous intensity λ equal to the empirical intensity $\overline{\lambda}$ of the process we want to analyze. We are comparing our process against random noise of the same intensity. This is how the data can be enveloped in the most basic form. The built alpha shape serves as the model itself. It gives us a representation of our data as an identifiable 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, 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 alpha shape $\mathcal{R}$ estimates the geometric structure of the data by filling the voids in-between close points. We then estimate the area of the created object. This value will not yield too much information about the influence of the variable within the model. Therefore, we must estimate the area of the minimum rectangle containing the entire object, the reader will immediately recognize that an CSR process of the same intensity λ as our pattern will tend to fill this box with some uniformity. If an input variable presents a weak correlation with the output variable, its behavior will be almost random with uniformly distributed points into the rectangular box, our null model. For the other cases, when there is some relevant correlation, it will create a pattern causing empty spaces to appear across the box.

To clarify the notation, we denote as $\mathcal{R}_{i,\alpha}$ the alpha shape generated by the pair of variables (X_i, Y) and radius α. We also denote the geometrical area of the object formed by the alpha shape $\mathcal{R}_{i,\alpha}$ by $\text{Area}_\alpha(\mathcal{R}_{i,\alpha})$.

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

$$B_i = \left[\min_{k=1,\ldots,n}(X_{ki}), \max_{k=1,\ldots,n}(X_{ki})\right] \times \left[\min_{k=1,\ldots,n}(Y_k), \max_{k=1,\ldots,n}(Y_k)\right].$$

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

We define our measure as

$$R^2_{\text{Geom},i,\alpha} = 1 - \frac{\text{Area}(\mathcal{R}_{i,\alpha})}{\text{Area}(B_i)}.$$

The index $R^2_{\text{Geom},i,\alpha}$ can be interpreted as the proportion of empty space remaining in the box containing the data as a function of α. This index is decreasing by definition because $\text{Area}(\mathcal{R}_{i,\alpha}) \leq \text{Area}(B_i)$ and $\text{Area}(\mathcal{R}_{i,\alpha})$ is increasing as $\alpha \to \infty$. If the R^2_{Geom} is close to zero, then we observe an object which shape cannot be set apart from that of the null model. Here, its structure has become noise like within the box. Otherwise, if the value is greater than zero, the object and the null model differ significantly. In this scenario we can observe emerging patterns in the data, the converse would also be true.

The estimation of R^2_{Geom} for a given α gives only partial information about the whole structure of the point cloud. By letting $\alpha \to \infty$ and observing how our index evolves, we can identify the curves defined by the map

$$f_i : \mathbb{R}^+ \to [0, 1]; \alpha \longmapsto R^2_{\text{Geom},i,\alpha} \tag{4}$$

Each curve f_i detects large geometric features in the data. In a sense, it corresponds to a discrete version of the converse to the empty space distribution function $F(\alpha)$ defined in Sect. 2.3.

$$1 - F(\alpha) = \mathcal{P}\{d(u) \geq \alpha\}$$

Our reasoning is our function unveils more information related to the geometric features within the dataset than $1 - F$ does. Nonetheless, it will be close to $1 - F$ when our pattern is an CSR. For this case, we note that

$$d(u) > \alpha \quad \textit{iff} \quad n(S \cap B(u, \alpha)) = 0$$

where $B(u, \alpha)$ is the ball centered at u with radius α.

Table 1 External features

Classification	Characteristic
E_1	The curve starts at 1 and becomes flat at some value much greater than 0 when α is large
E_2	The curve starts at 1 tends to approximately 0 when α is large

Table 2 Internal features

Classification	Characteristic
I_1	The curve has plateaus or pieces where f' is approximately zero
I_2	The curve decreases without plateaus at a slow rate
I_3	The curve decreases without plateaus at a fast rate

Since for a ball $B(u, \alpha)$ we have $|B| = \pi\alpha^2$ and for an CSR process of intensity λ the functional form of the distribution within a region B is $\exp(-\lambda|B|)$ we have that

$$1 - F_{\text{CSR}}(\alpha) = \exp(-\lambda\pi\alpha^2)$$

In this study, we identify 4 main patterns within each curve f_i, each of which point to relevant geometric features on the data. We classified each feature as either external (E) or internal (I). The main characteristics for each feature are given in Table 1 and Table 2, respectively.

The external features allow us to identify whether the model is close to an CSR process or not. Any data classified as E_1 will have a large empty space around the points. To be more specific, the value where the curve f_i tends in the case E_1 is the proportion of empty space between the square domain and the convex hull of the points. For example, consider data-points spread in the shape of a circle with center in (0, 0) and radius 1. Creating the alpha shape, we will have a structure E_1 and the convergence value of f_i to $1 - \text{Area}(\mathcal{R}_{i,\infty})/\text{Area}(B_i) \approx 1 - \pi/4 \approx 0.21$. If the data points are spread in a square, then f_i tends to $1 - \text{Area}(\mathcal{R}_{i,\infty})/\text{Area}(B_i) \approx 0$.

The internal structures are more subtle to interpret. Our interpretation for the I_1 scenario is that there are geometric features in the alpha shape of the data that persist for the given intervals in which the plateaus are observed.

As an example imagine a disc with a hole in its center, at first the alpha shape will fill the area contained within the data somewhat rapidly as α grows, at some point α_1 it will stabilize and stay that way until it reaches a value α_2, big enough to allow our process to start connecting points across the hole, after that it continues to fill the whole disc, leaving no trace of the hole whatsoever. This behavior will be identified in the curve f_i as follows: First starts at 1 and rapidly descend, stabilize around a given value α_1 for a while, then, at α_2 jump down abruptly. Then stabilize again until infinite.

Clearly all the shapes constructed in the interval (α_1, α_2) for which the function plateaus will stay almost the same. Therefore, under the mentioned conditions, the

R^2_{Geom} will vary little over this interval and the plateau emerges in the curve. For reference see Figs. 4 and 8

The structures I_2 and I_3 refer to solid objects without internal features or holes within the domain. Therefore, we can expect our function f_i to be strictly decreasing. The I_3 structure has rate of change or derivative f_i' with large negative values that rapidly increase toward zero.

We can infer regularity in the distribution. This is because the triangles in the alpha shape increase their size equally along the domain. Otherwise, for the I_2 structure, there are sections where the triangles increase faster, causing a general slow rate for f_i.

3.2 Hypothesis testing and monte-carlo envelopes

As stated in (Baddeley et al. 2016, Chapter10), a Monte-Carlo test for a spatial point pattern, be it data sets, can be performed using any summary function applied to it. In the case of this study, the summary function applied to our data sets is the R^2_{Geom} function.

Since our score function is in some sense a discrete version of the function $1 - F$ we believe it makes more sense to apply a pointwise envelope test to validate its performance. To estimate a Monte-Carlo envelope we apply the following procedure:

1. Determine the empirical intensity for each variable using equation (3).
2. Determine the expected number of points $\overline{n} = \overline{\lambda} \, |W|$.
3. Generate a number $N \sim \text{Poisson}(\overline{n})$.
4. Generate N values $X \sim \text{Uniform}(0, 1)$ and $Y \sim \text{Uniform}(0, 1)$.
5. Estimate the function $f_i(r)$ for X and Y according to Sect. 3.1.
6. Repeat 40 times the steps 3–5.

For every observed data set we estimate the empirical intensity of the process and run 40 iterations of the function applied to the null hypothesis with the same intensity as the data to create our envelope.

For each value of the radius α, or for whole intervals at a time, we can infer whether our model rejects the null hypothesis. We take into account that for this genre of tests the rejection/confirmation of the null hypothesis is not global but limited to specific values of α and that the statistical significance of such outcomes is in practice much lower than the theoretical value of $\gamma = 2/40 = 0.05$.

4 Results

To measure the quality of the index described above, we work concrete examples. The software used was `R` (R Core Team 2020), along with the package `sf` (Pebesma 2018) for handling spatial objects and area estimation. Our package containing all these algorithms will be available soon in CRAN.

In all examples we sample $n = 1000$ points $\{s_l\}$ with the distribution specified in each case. To determine the radius in the alpha shape, we take a sequence of 100 values for a fixed variable i, using two defined values $r_{\min}$ and $r_{\max}$ defined for each case. We call the sequence α_k with $k = 1, \dots, 100$. We plot each curve $f_i(r_k)$ against this sequence of α values.

4.1 Theoretical examples

We will consider five unique settings for our examples, each one with different topological and geometric features. These settings are not exhaustive and there are others with interesting features. However, through this sample we do show how our method captures the geometrical correlation of the variables, where other classical methods have failed to do so.

The examples considered have some simple but outstanding geometrical features. The latter will help us interpret the shape of our function f_i and its meaning. The examples are

- **Linear:** This is a simple setting with

$$Y = 0.6X_1 + 0.3X_2 + 0.1X_3$$

  where X_3 is an independent random variable. We set $X_i \sim \text{Uniform}(-2, 2)$ for $i = 1, 2, 3$. Figure 3 presents the scatterplot of this model. It is simple enough to establish if our algorithm detects –in decreasing order– if a variable has a more geometric correlation than following.
- **Concentric circles with holes:** The model consists of three different circles all centered at $(0, 0)$. We set first $\theta \sim \text{Uniform}(0, 2\pi)$ and,

  1. Circle with radius between 3 and 4:

$$\begin{cases} C_1^X &= r_1 \cos(\theta) \\ C_1^Y &= r_1 \sin(\theta) \end{cases}$$

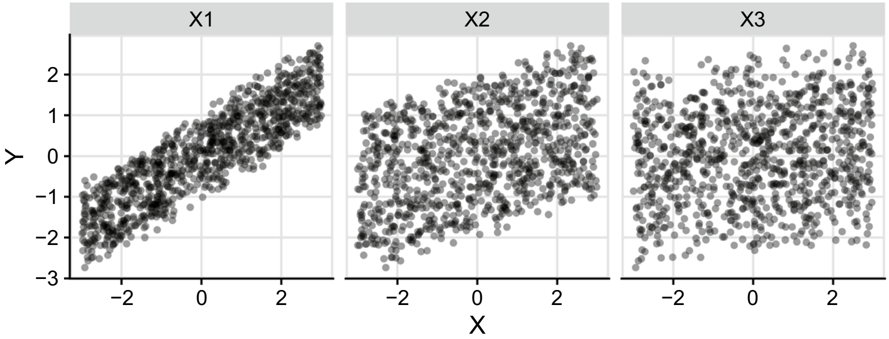

Fig. 3 Cloud of data points for linear model

where $r_1 \sim \mathrm{Uniform}(3, 4)$.

2. Circle with radius between 6 and 7:

$$\begin{cases} C_2^X & = r_2 \cos(\theta) \\ C_2^Y & = r_2 \sin(\theta) \end{cases}$$

where $r_2 \sim \mathrm{Uniform}(6, 7)$.

3. Circle with radius between 10 and 11:

$$\begin{cases} C_3^X & = r_3 \cos(\theta) \\ C_3^Y & = r_3 \sin(\theta) \end{cases}$$

where $r_3 \sim \mathrm{Uniform}(10, 11)$.

Now define $X_1 = (C_1^X, C_2^X, C_3^X)$ and $Y = (C_1^Y, C_2^Y, C_3^Y)$ as the concatenation of three circles and $X_2 \sim \mathrm{Uniform}(-10, 10)$. The cloud of data point is presented in Fig. 4. This example will allow us to show how we can capture all the geometric feature of the data as r increase.

- **Two circles and one ellipse with holes:** In this case we have two different circles and one ellipse. We set first $\theta \sim \mathrm{Uniform}(0, 2\pi)$ and,

1. Circle centered at (1.25, 5) with radius between 1 and 1.25:

$$\begin{cases} C_1^X & = r_1 \cos(\theta) + 1.25 \\ C_1^Y & = r_1 \sin(\theta) + 5 \end{cases}$$

where $r_1 \sim \mathrm{Uniform}(1, 1.25)$.

2. Circle centered at (3, 1.5) with radius between 0.75 and 1.5:

$$\begin{cases} C_2^X & = r_2 \cos(\theta) + 3 \\ C_2^Y & = r_2 \sin(\theta) + 1.5 \end{cases}$$

where $r_2 \sim \mathrm{Uniform}(0.75, 1.5)$.

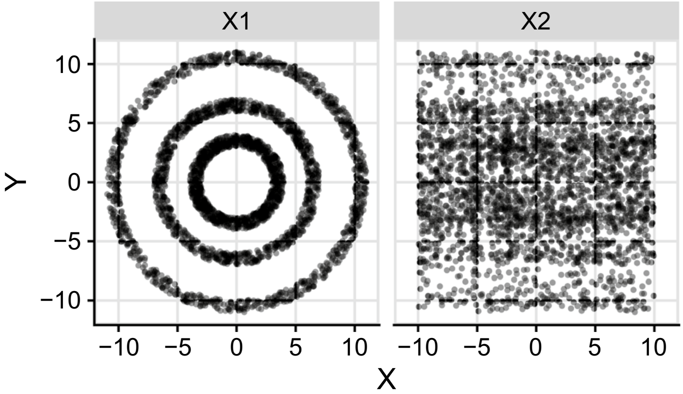

Fig. 4 Cloud of data points for Concentric circles with holes model

3. Ellipse centered at (6.25, 4) with radius between 1 and 1.5:

$$\begin{cases} C_3^X &= 0.5 \times r_3 \cos(\theta) - 6.25 \\ C_3^Y &= 2 \times r_3 \sin(\theta) + 4 \end{cases}$$

where $r_3 \sim \mathrm{Uniform}(1,4)$.

Now define $X_1 = (C_1^X, C_2^X, C_3^X)$ and $Y = (C_1^Y, C_2^Y, C_3^Y)$ as the concatenation of three circles and $X_2 \sim \mathrm{Uniform}(0,7)$. The cloud of data point is presented in Fig. 5. Our aim is to detect multiple geometric features at different resolutions. As the irregular pattern is filling the space, the curve f will bump downward with the same behavior. Also, notice how X_2 does not have a CSR pattern due to the different spatial point densities.

- **Ishigami:** The final model is

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

where $X_i \sim \mathrm{Uniform}(-\pi, \pi)$ for $i = 1, 2, 3$, $a = 7$ and $b = 0.1$. This model is represented in Fig. 6. This is a popular model in sensitivity analysis because it presents a strong non-linearity and non-monotonicity with interactions in X_3. Using 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 further explanation of this function, we refer the reader to Sobol and Levitan (1999).

4.2 Numerical results

To show the effects in the space filling function f_i we estimate its discrete derivative. We define

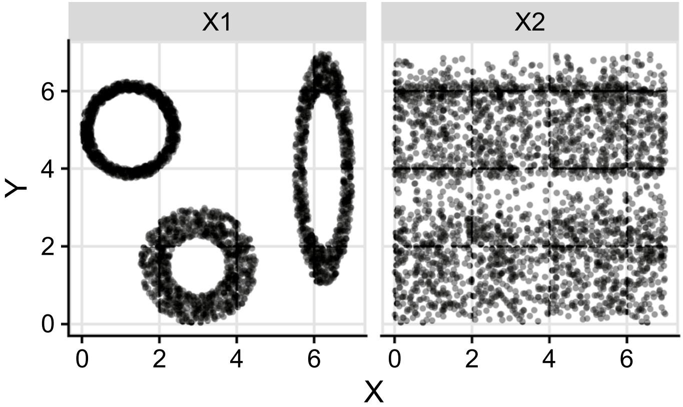

Fig. 5 Cloud of data points for the two circles and ellipse with holes model

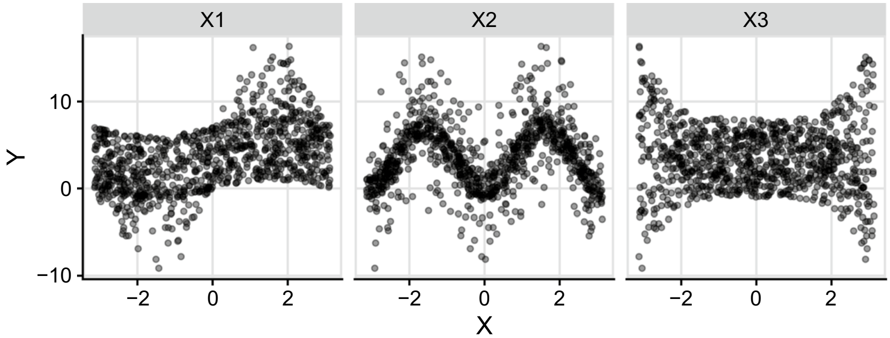

Fig. 6 Cloud of data points for Ishigami model

$$f_i'(r) = \frac{f_i(r+h) - f_i(r-h)}{2h},$$

where h is a small value. This function represents the discrete rate of change of the function f_i as α increases. As visual help to the reader, we estimate a cubic interpolation through the values of f_i' (represented with a solid line).

The figures presented in this section represent each theoretical example from different perspectives. They consist of three panels:

Upper panel: In solid black the function f_i in terms of the radius $r = \alpha/2$. In dashed red the function $1 - F(r) = \exp\left(-\lambda\pi r^2\right)$ describing the CSR process. Furthermore, in light gray the envelopes generated by Monte-Carlo simulation.

Middle panel: The dots are the pointwise derivative of f_i explained before. In solid black a spline interpolation just for visual aid.

Bottom panel: The function $1 - F(r)$ estimated with the package `spatstat` (Baddeley et al. 2016). In solid black we present the Kaplan-Meier estimation. In dashed red, the CSR process.

For all cases we define a maximum radius r and all the panels are estimated using this value.

The linear model in Fig. 7 is simple enough to allow us to directly see that variable X_1 decreases rapidly but stabilizes at value near 0.6. We classify this behavior as E_1. The same behavior occurs with the second variable. Here the stabilization value is just above 0.25. Finally, the third variable decreased and stabilized to values nearly 0. For X_3, we have and E_1 structure but depending on the tolerance level we can define as E_2. Our scale lacks a proper normalization, nonetheless it helps us describe the behavior of our objects.

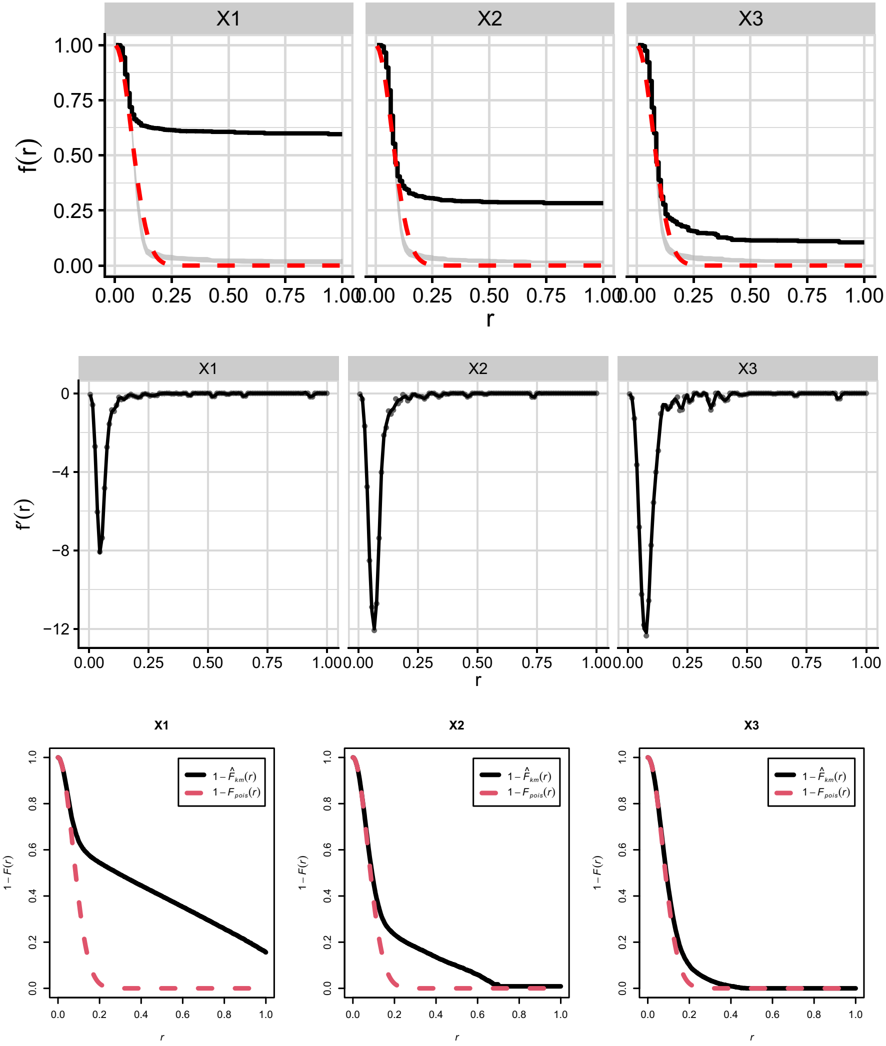

Fig. 7 Results for the linear model. *Upper panel:* Function f_i of the alpha shape filled empty space. *Middle panel:* Derivative of f_i. *Lower panel:* The solid black line is the Kaplan-Meier estimator of the empty space distribution function $1 - F(r)$. The red dashed line is the distribution of a Poisson (random) process

The internal structure shows for the three variables a fast-paced rate of decrease. The middle panels show the first variable is slower than the others. In terms of correlation, we can say that the three variables have I_3 structures. The function f' for variable X_2 tells the cloud of points is more uniformly distributed than X_3. Note that X_3 presents a series of micro-bumps in the stabilization phase. The behavior indicates small geometric persistence. However, curve approaches to the CSR process indicating a pure noise pattern. The generated envelopes showed that X_1 and X_2 diverge from the CSR process while X_3 approaches to it.

A particular case of the model of concentric circles with holes Fig. 8 was discussed in the preliminaries. Recall that in this case, both variables have R^2 scores nearly zero for our model, even if the geometric shape showed two different behaviors. In the external structures, we have E_1 and E_2 for X_1 and, X_2 respectively.

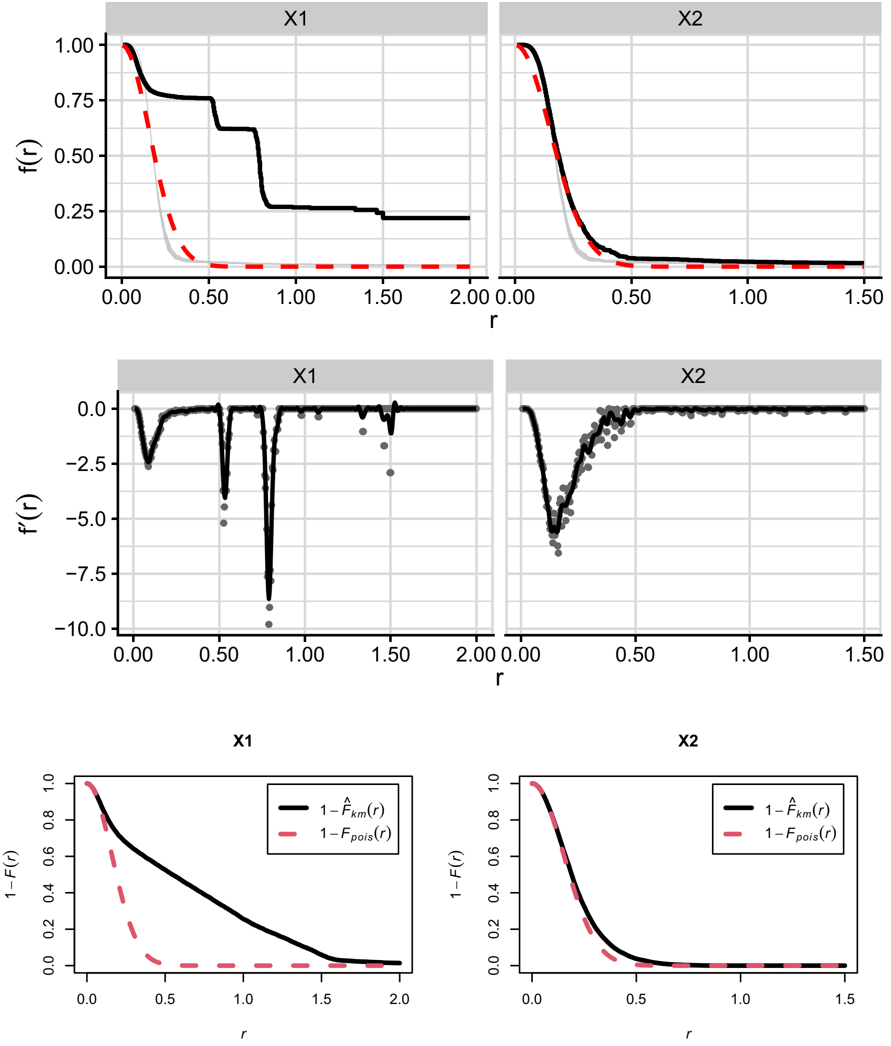

Fig. 8 Results for the circle with one hole model. *Upper panel:* Function f_i of the alpha shape filled empty space. *Middle panel:* Derivative of f_i. *Lower panel:* The solid black line is the Kaplan-Meier estimator of the empty space distribution function $1 - F(r)$. The red dashed line is the distribution of a Poisson (random) process

For the internal structures, notice the first variable has two larger bumps near $r = 0.5$ and $r = 0.75$. Around $r = 1.5$ we observe a small jump which complete fully the space. This jump is smaller than the others because the previous radii have capture all the geometric information of the data. The function $f'(r)$, is zero at large portions just before the bumps. We see a persistence in data in those segments. We establish an I_1 structure for X_1.

For X_2, as expected, we observe a I_3 structure consisting of an almost noisy variable. The envelopes confirm the pattern from X_1. For X_2 is clear that our estimation, the CSR process and the envelopes has the same tendency.

Our index extracted the correct structure of the data. If we compare the result with the function $1 - F(r)$, the results are also consistent. The function $1 - F(r)$ detects the correct spatial pattern. However, notices how the Kaplan-Meier estimator decreases smoothly ignoring the geometric hole in the data.

To test our algorithm further, we present the model of "Two circles and one ellipse with holes" in Fig. 9. Here we created two circles and one ellipse at different scales and positions. We captured the most relevant feature for each projection. The first variable again has an E_1 structure because we observe some stabilization after $r = 1$. Variable X_2 in contrast tends to have an E_2 structure.

Our geometric correlation index detected the geometric pattern. However, due to the irregularity, the bump are smaller than the case before. In $r = 0.25$ we notice the first circle captured. The second one is near $r = 0.5$. Given the relative size of the ellipse, in $r = 0.75$ we observe the bigger bump of the curve. The function $f'(r)$ confirm this presenting large negative values at the mentioned points before. We can infer a combination of I_1 and I_2 structures. In terms of $1 - F(r)$, we noticed a correct identification of the spatial point pattern. Our method, as in the previous case, detected the intervals in the radius where the geometric features of the data existed.

The final model we consider is the ishigami model. Figure 10 presents the results. The pattern in this example is more complex, given the spread of points in the domain. The first variable has a compact geometry, being straightforward to identify the pattern. Notice how the function f_1 decreases slowly until stabilization. The variable has a E_1 and I_2 structures. For the second variable, the pattern is a “M” shape figure with a noisy background. The function f' reveals the slowest rate of decreasing among the three variables. One aspect to note is the abrupt changes in the derivative function. This means that the spatial pattern contains irregularities filled at different rates. In any case, the curve describes the rich structure of the data.

The third variable is the most interesting due to it is uncorrelated nature with respect to Y. However, our curve fills the space slowly until starting a big bump around 1.25 and finish it in 1.5. From here, the whole geometry is filled. The structures are E_1 and I_2 differing with the statistical behavior in the Kaplan-Meier estimator. We can appreciate it has a rich geometry and the method captures it. In the three variables the envelopes negate that the presence of CSR pattern in the data.

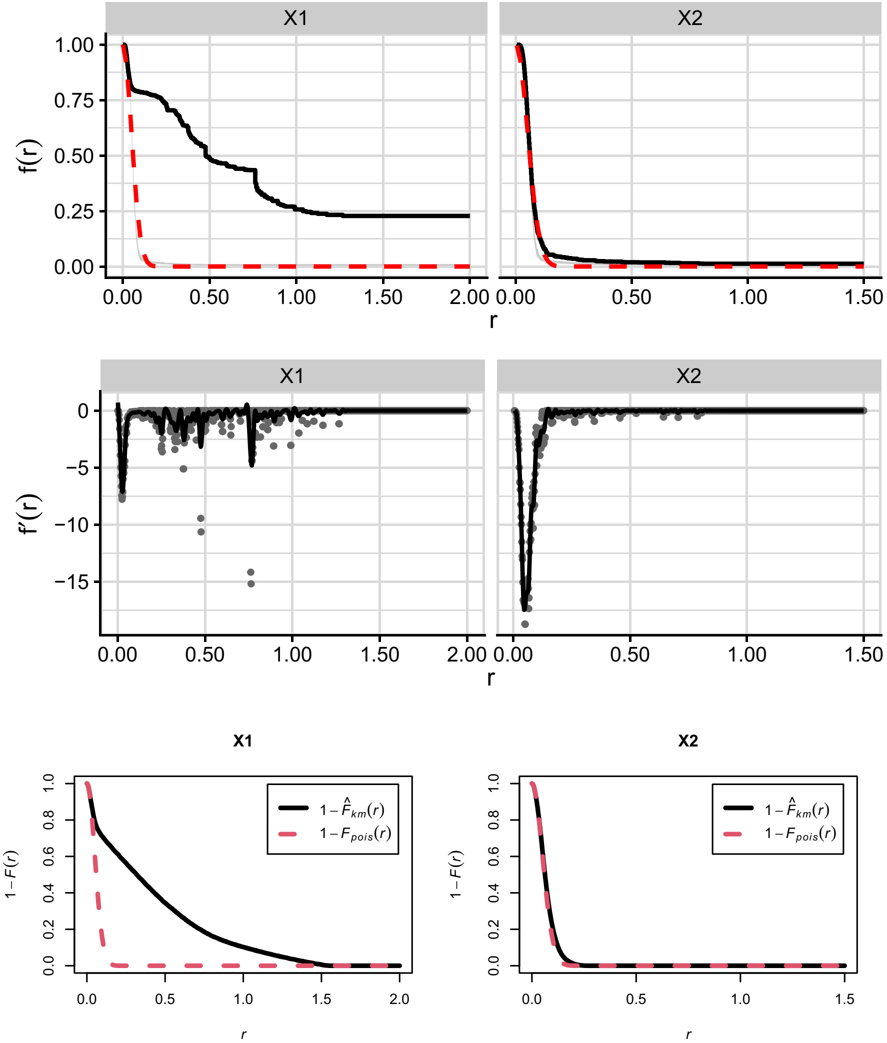

Fig. 9 Results for the two circles and ellipse with hole model. *Upper panel:* Function f_i of the alpha shape filled empty space. *Middle panel:* Derivative of f_i. *Lower panel:* The solid black line is the Kaplan-Meier estimator of the empty space distribution function $1 - F(r)$. The red dashed line is the distribution of a Poisson (random) process

5 Conclusions and further research

As mentioned above, we built a goodness-of-fit index relying solely on the geometric features of a data-cloud. Purely analytic or statistical methods cannot recognize the structure when we project certain variables, primarily when the input is of zero-sum, which might be artificial noise. In such cases those projections, or

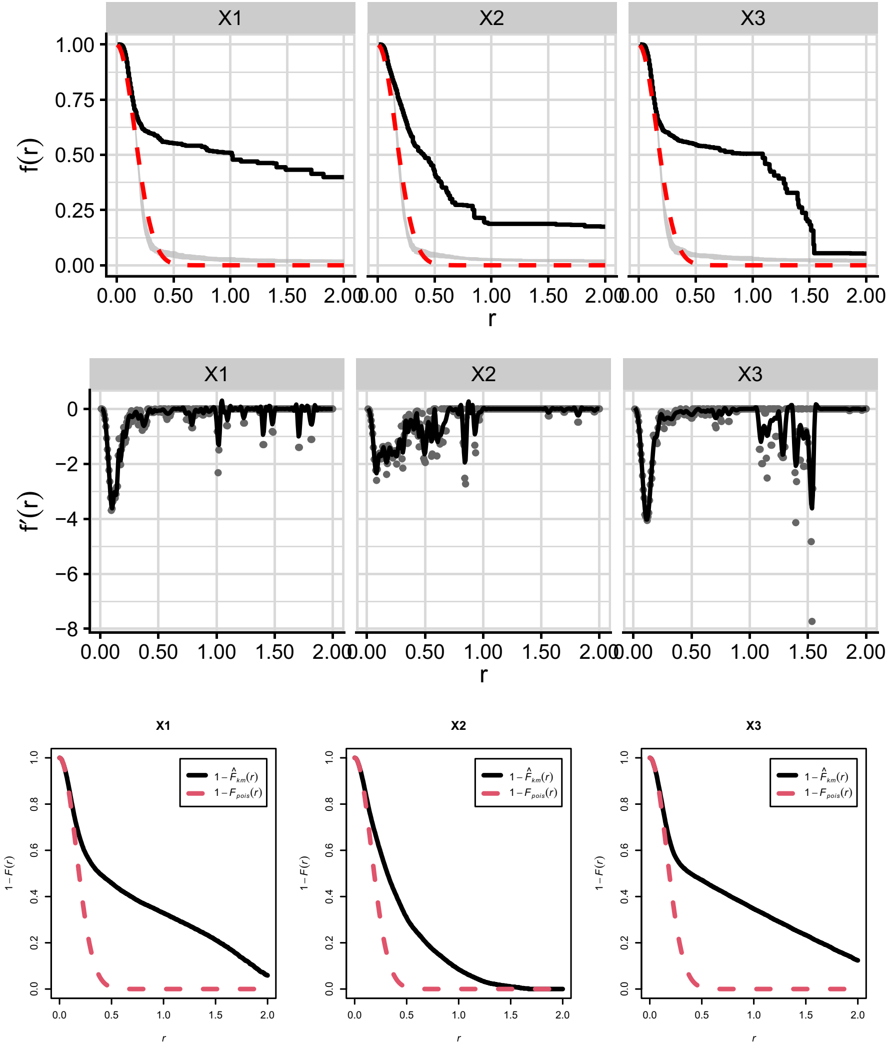

Fig. 10 Results for the Ishigami model. *Upper panel:* Function f_i of the alpha shape filled empty space. *Middle panel:* Derivative of f_i. *Lower panel:* The solid black line is the Kaplan-Meier estimator of the empty space distribution function $1 - F(r)$. The red dashed line is the distribution of a Poisson (random) process

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 variability of the data when the variable is of zero-sum, differentiating between pure random noise and well-structured inputs. Where the model presents pure noise, our index fully coincides with other methods indexes. It also detects relevant structured inputs, in the other cases our index shows the structure in the variables, which was the notion we wanted to explore. We use a subjective scale (E_1, E_2, I_1, I_2, I_3) to describe those structures. The scale is incomplete and requires more work to be formalized. Nevertheless, it helped us describe possible patterns in the data given the trend of the filled space function f_i.

The study of the *filled space* function with alpha shapes was insightful to discover real patterns in the data. Also, by estimating its derivative, we can describe more precisely its behavior. We used a reconstruction of the curve using cubic splines. Even if the fit was sufficient, it can be improved.

As a continuation of this research we propose to study and describe the functional form of f_i. With this, we can estimate precisely where the bumps occur and estimate the derivative with more accuracy. Higher order derivatives can also help to reveal more patterns.

Our major goal for a future project is to assess whether our model is useful in determining the relevance of a variable within a model.

Acknowledgements We acknowledge Santiago Gallón and Ronald A. Zuñiga for enlightening discussions about the subject, whose help has been valuable. The authors acknowledge the financial and administrative support from Vicerrectoría de Investigación, Centro de Investigaciones en Matemática Pura y Aplicada (CIMPA) and Escuela de Matemática from Universidad de Costa Rica through the project 821–B8–221.

References

Baddeley Ad, Rubak E, and Turner R (2016) Spatial point patterns: methodology and applications with R. eng. Chapman & Hall/CRC inter- disciplinary statistics series. Boca Raton London New York: CRC Press. ISBN: 978-1-4822-1021-7 978-1-4822-1020-0

Balasubramanian M (2002) The isomap algorithm and topological stability. Science. https://doi.org/10.1126/science.295.5552.7a (**ISSN: 00368075**)

Barrett James P (1974) The coefficient of determination and some limitations. Am Stat 28(1):19–20. https://doi.org/10.1080/00031305.1974.10479056 (**ISSN: 0003-1305, 1537-2731**)

Barten AP (1962) Note on unbiased estimation of the squared mul- tiple correlation coefficient. Stati Neer 16(2):151–164 (**ISSN: 0039-0402, 1467-9574**)

Bellman R (1957). Dynamic programming. Princeton university press. ISBN: 978-0-486-42809-3

Bellman R (1961). Adaptive control processes: A guided tour. 4, Princeton University Press

Bernstein M et al (2000). Graph approximations to geodesics on embedded manifolds. In: Igarss 2014 01.1, 1–5. ISSN: 0717-6163. https://doi.org/10.1007/s13398-014-0173-7.2.arXiv: 1011.1669v3

Bouchaffra D (2012) Mapping dynamic Bayesian networks to α- shapes: application to human faces identification across ages. IEEE Trans Neural Netw Learn Syst 23(8):1229–1241 (**ISSN: 2162-2388**)

Buja A et al. (2005). Computational methods for high-dimensional rotations in data visualization. In: Handbook of statistics. Ed. by CR Rao, EJ Wegman, and JL Solka. 24 Data mining and data visualization. Elsevier, 391–413. https://doi.org/10.1016/S0169-7161(04)24014-7

Cramer JS (1987) Mean and variance of R2 in small and moderate samples. J Econom 35(2–3):253–266 (**ISSN: 03044076**)

Edelsbrunner H (2014). A short course in computational geometry and topology. English. 1 Springer Briefs in Applied Sciences and Tech-nology. Cham: Springer International Publishing. ISBN: 978-3-319-05956-3 978-3-319-05957-0. https://doi.org/10.1007/978-3-319-05957-0.

Gardiner James D, Julia B, Charlotte AB (2018) Alpha shapes: determining 3D shape complexity across morphologically diverse structures. BMC Evolutionary Biology 18(1):184 (**ISSN: 1471- 2148**)

Guerrero José-Luis et al (2013) Exploring the hydrological robustness of model-parameter values with alpha shapes. Water Res Research 49(10):6700–6715 (**ISSN: 1944-7973.**)

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, Inference, and Prediction. Springer, New York. 978-0-387-84857-0. arXiv: 1011.1669v3. https://doi.org/10.1007/978-0-387-84858-7

Myatt GJ, Johnson WP (2009) Making Sense of Data II: a practical guide to data visualization, advanced data mining methods, and applications. Hoboken, N.J, Inglés. 978-0-470-22280-5

Pebesma E (2018) Simple features for R: standardized support for spatial vector data. R Journal 10(1):439 (**ISSN: 2073-4859**)

Press SJ, Zellner A (1978) Posterior distribution for the multiple correlation coefficient with fixed regressors. J Econom 8(3):307–321 (**ISSN: 03044076**)

R Core Team (2020). R: a language and environment for statistical computing

Sobol IM, Levitan YuL (1999) On the use of variance reducing multipliers in monte carlo computations of a global sensitivity index. Comput Phys Commun 117(1–2):52–61 (**ISSN: 00104655**)

Tenenbaum JB (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323 (**ISSN: 00368075**)

Tufte ER (2001) The visual display of quantitative information. Cheshire, Conn, Inglés. 978-1-930824-13-3

Wood S (2006) Generalized additive models: an Introduction with r. 1, CRC Press. ISBN: 978-1-58488-474-3

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.