

Level sets of conditional Gaussian fields

T E S I S

Que para obtener el grado de
Maestro en Ciencias con especialidad en Probabilidad y Estadística

Presenta:

Victor Andres Amaya Carvajal

Directores de tesis:

Dr. Sayan Mukherjee

Dr. Víctor Manuel Pérez-Abreu Carrión

Autorización de la versión final

Abstract

Gaussian processes are extensively used for regression and optimization tasks. This thesis aims to understand the behavior of the level sets of three-dimensional conditional Gaussian processes. Given a random sample of one of these processes, we can infer information about the geometrical structure of the process that generates it. With this knowledge, we can improve our predictions or avoid local minima in the optimization case.

For our empirical analysis, we simplify the problem by conditioning a Gaussian process over a known smooth boundary that is contained inside a given square. We model the level sets topology of this conditioned Gaussian process via Vietoris-Rips complexes, for which we can use fast computer algorithms to calculate the rank of their corresponding homology groups.

Keywords: Topology of level sets, geometrical data analysis, Betti numbers, Poisson distribution, Vietoris-Rips complexes.

Acknowledgements

A mis padres, Víctor y Berta. Muchas gracias por todos sus consejos y por su apoyo durante todos estos años que no he estado en casa. Los amo mucho.

I am deeply grateful to my advisors Víctor Pérez-Abreu and Sayan Mukherjee. Víctor, thank you for all of your kind advice and guidance during these years. Sayan, thank you for your support and for believing in me.

I want to thank CIMAT for the excellent and rigorous education that I received during my undergraduate and graduate degrees. Thank you very much for transforming and changing my life.

I want to thank CONACyT for the economical support, without which these studies would not have been possible.

Introduction

This thesis aims to study and understand the level sets of conditional Gaussian processes. This type of stochastic process is widely used in many areas of probabilistic and statistical machine learning, as well as in manifold learning for tasks such as regression, classification, and prediction; as highlighted in [Rasmussen and Williams \(2005\)](#).

In a regression problem, for instance, we condition a Gaussian process over a previously observed data sample. This conditional process can then be used to produce prediction intervals, with a certain degree of confidence, for new unobserved data points. Some examples of the application of conditional Gaussian process can be found in [Frazier \(2018\)](#).

Gaussian processes (GPs) are in some cases preferred over other non-linear machine learning methods, such as neural networks, due to the vast and rich existing theory related to the Gaussian distribution. To apply a Gaussian process in a regression problem, we have to place a prior distribution over the observed data sample. In other words, we assume that the dataset comes, jointly, from a multivariate Gaussian distribution.

We are interested in studying the level sets because they can provide information about the geometrical structure of the conditional Gaussian process. If we infer something about the geometry of the object from where a random sample was taken, then we could use this knowledge to provide better estimates for new predictions. If we are using Gaussian processes to maximize a certain function, as in Bayesian optimization, then knowing something about the geometry can help us to avoid local minima.

The case of the unconditional Gaussian processes was considered by [Thoppe and Krishnan \(2018\)](#). In this work, the authors analyze the Betti numbers (rank of the homology groups, denoted by β_k for $k \geq 0$) of the level sets for the unconditional case and they obtain various results about their asymptotic distribution. In their analysis, the authors propose a theoretical novel approach to estimate the topology of the level sets via Čech complexes.

For distribution and regression functions, [Bobrowski, Mukherjee, and Taylor \(2017\)](#), [Bobrowski and Mukherjee \(2013\)](#) provide consistent estimators for the topology of their level sets that could be employed to study the conditional Gaussian process case. The theory developed in these works suggest the idea of analyzing the conditional case via an empirical simulation study.

To the best of our knowledge, there is no other work that aims to study the level sets of conditional Gaussian processes. Consequently, we propose a method that could provide some intuition about their corresponding level sets. Given a fixed threshold value L , this method simulates sample sets at level L , for which we can approximate the number of connected components and one-dimensional holes. Our proposal is computationally efficient because it does not require extensive computational resources. Additionally, it

was developed in a way that could be used in a computer cluster to run multiple simulations in parallel. This gives us the possibility of creating large sample sets, which could help to improve the inferences that we can obtain from them.

Using the proposed method, our results show that the first two Betti numbers, which count the number of connected components and the number of one-dimensional holes, respectively, distribute as a Poisson distribution whose parameters depend on the selected threshold value. To test or validate the plausibility that these empirical samples come from a Poisson distribution, we need a test that does not require independence of the sample elements because the simulated points that belong to the same level set are not statistically independent. Therefore, we employed a graphical tool proposed by [Nakamura and Pérez-Abreu \(1993\)](#).

There are two main reasons why this work is relevant: the first is practical and the second is theoretical. The practical reason is linked with its application to machine learning, in which the use of Gaussian process is common; that is, the uncertainty quantification that we could get from the study of the level sets of conditional Gaussian process. The theoretical reason is related to the numerical estimation of the parameters of the Poisson distribution that explain the first two Betti numbers β_0, β_1 ; as stated earlier, these parameters depend on the threshold value that is used to calculate the level sets.

More specifically, given a dataset of observations (training data), the question is how we can use these known data points to make a prediction over new observations if the geometric relation between the observed points and their output is not trivial.

This is a difficult and open problem in machine learning. In this work, we simplify this problem to the case where all of the known observed data points are contained in a fixed smooth boundary, within a given square on the plane, by conditioning the Gaussian process over this known boundary and then studying their level sets.

Obtaining some ideas or understanding more about the behavior of this particular conditional Gaussian process may give some insights into how these ideas could be generalized to other situations, such as when the boundary condition has a more complex geometrical structure. This could lead to more confident predictions, in the regression case, when something is known about the geometric or topological structure of the data.

This thesis is organized in five chapters. Chapter 1 is a compilation of the main analytical tools that we need, with an emphasis on functional analysis concepts, Gaussian processes, kernel methods in machine learning, and goodness of fit tests for the Poisson distribution. This chapter is also intended to introduce the reader to the notation used and it will provide some examples of the usefulness of the tools that presented here.

Chapter 2 provides an introduction on how to compute the Betti numbers of a given discrete set of points. Rather than just presenting this in a simplistic algorithmic way, we rigorously provide the definitions and theorems from algebraic topology needed to define homology and their Betti numbers. In this chapter, we also propose a computational method to efficiently calculate the Betti numbers. We do this by modifying the faster existing library by [Bauer \(2019\)](#) that is used for data analysis exploration in the area of topological data analysis.

Chapter 3 presents a detailed explanation of the proposed method to study the level sets of conditional Gaussian processes. We start with an exposition of how to simulate a conditional Gaussian process. Then, we review some of the newly developed tools that we use to approximate the topology of level sets of conditional Gaussian process. More precisely, to approximate the first two Betti numbers β_0, β_1 of their level sets.

Chapter 4 includes a compilation of some of the simulations that we have obtained and it describes the insights that we gather from them. For every level set, we calculated the sample mean and variance as a first step to test for the possibility that a Poisson distribution could explain the simulated data. We then use a graphical test to verify the plausibility that the distribution of the number connected components and one-dimensional holes at the level sets is Poisson.

Finally, Chapter 5 contains the conclusions of our simulation study and it presents some new questions that have arisen from this work.

Contents

Abstract	i
Acknowledgements	ii
Introduction	iii
List of Figures	viii
1 Preliminaries	1
1.1 Elements of functional analysis	1
1.1.1 Basic definitions	1
1.1.2 Positive definiteness and reproducing kernels	4
1.2 Regression with Gaussian processes	6
1.2.1 Gaussian processes	6
1.2.2 Gaussian processes for regression	9
1.3 Exceedances of the Poisson point process	12
1.3.1 Poisson point process in the plane	12
1.3.2 Point approximation approach	13
1.4 Goodness of fit test for Poisson distribution	13
2 Betti numbers computation	15
2.1 Topological definitions	15
2.1.1 Simplicial complexes	15
2.1.2 Chains, boundaries, and cycles	17
2.1.3 Homology groups	19
2.2 Matrix algebraic theory	19
2.2.1 Boundary matrix	20
2.2.2 Matrix reduction	21
2.3 Use of Ripser to efficiently calculate the Betti numbers	22
3 Topology of the level sets	25
3.1 Study of the supra-level sets	25
3.1.1 Proposed algorithm	25
3.1.2 Construction of a conditional Gaussian process	26
3.1.3 Level sets approximation	27
3.2 Known tools	28
3.2.1 Naive estimator	29
3.2.2 Robust homology estimator	30

4 Empirical study of level sets	32
4.1 Connected components	32
4.1.1 Method description	32
4.1.2 Simulation results	33
4.1.3 Conclusions	35
4.2 One-dimensional holes	37
4.2.1 Method description	37
4.2.2 Simulation results	38
4.2.3 Conclusions	39
4.3 Asymptotic exploration of connected components	44
5 Conclusions and pending work	47
5.1 Open questions	47
References	48

List of Figures

1.1	Sample points from MV normal distribution.	9
1.2	Example of a conditional process	11
1.3	3D Gaussian process embedded in \mathbb{R}^3	12
2.1	Čech complex to the left-hand side, and Vietoris-Rips complex to the right-hand side (Chazal et al.).	17
2.2	Example of a simplicial complex.	20
2.3	Example of a barcode. (Bauer (2019))	23
2.4	Betti curves of uniform points in the unit square.	24
3.1	Two different boundary conditions	26
3.2	VR complex approximation of a level set.	28
3.3	Discretization example of a conditional GP.	29
4.1	β_0 for the case: Above.	36
4.2	β_0 for the case: Below.	36
4.3	β_0 for the case: Double.	37
4.4	β_1 for the case: Above.	40
4.5	β_1 for the case: Below.	40
4.6	β_1 for the case: Double.	41
4.7	β_0 : Asymptotic behaviour of Theorem 3.2, $h=\text{median}(\text{dist}(\mathbf{X}))$	44
4.8	β_1 : Asymptotic behaviour of Theorem 3.2, $h=\text{median}(\text{dist}(\mathbf{X}))$	45
4.9	β_0 : Asymptotic behaviour of Theorem 3.2 (fixed h).	45
4.10	β_1 : Asymptotic behaviour of Theorem 3.2 (fixed h).	46

1 Preliminaries

Introduction

In this chapter, we will briefly introduce the analytic concepts that are needed for this work, with an emphasis on functional analysis, statistical methods, and the application of kernel methods in machine learning. In particular, we review their application to the simulation of Gaussian processes. We begin by introducing the concepts needed to rigorously define kernel functions. Then, we introduce the tools that will be used in Chapters 3 and 4 to statistically analyze the exceedances, above a fixed threshold value, of the level sets of conditional Gaussian processes. The concepts and results that are presented here are arranged in such a way that their combination becomes a natural and intuitive approach of how to simulate the process in question. Finally, we will also work out some concrete examples to try to make the concepts as accessible as possible.

1.1 Elements of functional analysis

1.1.1 Basic definitions

First, we will review and establish the notation of some basic definitions from functional analysis [Clapp \(2017\)](#).

Definition 1.1. A metric space \mathcal{M} is called complete if every Cauchy sequence of points in \mathcal{M} has a limit that is also in \mathcal{M} ; in other words, if every Cauchy sequence in \mathcal{M} converges in \mathcal{M} .

Definition 1.2. A Banach space is a vector space X over a scalar field K , which is equipped with a norm $\|\cdot\|_X$ and which is complete with respect to the distance function induced by the norm.

Definition 1.3. (Inner product space) Let \mathcal{H} be a vector space over a scalar field K . A function $\langle \cdot, \cdot \rangle_{\mathcal{H}}: \mathcal{H} \times \mathcal{H} \rightarrow K$ is said to be an inner product if, and only if

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$ (bi-linearity),
2. $\langle f, g \rangle_{\mathcal{H}} = \overline{\langle g, f \rangle_{\mathcal{H}}}$ (conjugate symmetry),
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0$, and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$ (positive-definite).

Definition 1.4. A Hilbert space \mathcal{H} is a real or complex inner product space that is also a complete metric space with respect to the distance function induced by the inner product.

On a vector space X with inner product $\langle \cdot, \cdot \rangle_X: X \times X \rightarrow K$, we can define a norm using the inner product as follows $\|f\|_X := \sqrt{\langle f, f \rangle_X}$. Using a norm we can define a distance for any $f, g \in X$ as $d(f, g) := \|f - g\|_X$.

Remark: Unless otherwise stated, when working on an inner space X , if there is no risk of confusion, then we will indistinguishably use the symbols $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle$.

Example 1.1. Let μ is a positive measure on $A \subseteq \mathbb{R}^d$, $d \geq 1$, then the space

$$L_p(A; \mu) = \left\{ f: A \rightarrow \mathbb{R} \text{ measurable} : \int_A |f(x)|^p d\mu < \infty \right\}$$

is a Banach space with norm: $\|f\|_p = (\int_A |f(x)|^p d\mu)^{1/p}$.

Theorem 1.1. The following properties of the inner product and norm hold:

1. $|\langle f, g \rangle| \leq \|f\| \cdot \|g\|$ (Cauchy-Schwartz inequality).
2. $\|f + g\|^2 + \|f - g\|^2 \leq 2\|f\|^2 + 2\|g\|^2$.
3. $4 \langle f, g \rangle = \|f + g\|^2 - \|f - g\|^2$.

Example 1.2. Let μ be a positive measure of $X \subseteq \mathbb{R}^d$, then the space $L_2(X, \mu)$ is a Hilbert space with inner product

$$\langle f, g \rangle = \int_X f(x)g(x) d\mu(x),$$

for all $f, g \in L_2(X, \mu)$, where $L_2(X, \mu)$ is the space of equivalence classes of functions that differ by at most a set of μ -measure zero.

Example 1.3. Let $a = (a_i)_{i=1}^\infty$, $b = (b_i)_{i=1}^\infty$, where $a, b \in \mathcal{H}$, $i \geq 1$, here \mathcal{H} is the vector space of all convergent sequences of real numbers. The following is an inner product on the space of infinity sequences:

$$\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i. \tag{1.1}$$

We are now ready to define a kernel function, based on lecture notes [Gretton \(2019\)](#) and [Sejdinovic and Gretton \(2014\)](#).

Definition 1.5. (*Kernel*) Let \mathbf{X} be a non-empty set. A function $k: \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ is called a *kernel* if there exists a \mathbb{R} -Hilbert space \mathcal{H} and a function $\varphi: \mathbf{X} \rightarrow \mathcal{H}$ such that $\forall x, x' \in \mathbf{X}$,

$$k(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}.$$

Example 1.4. Suppose that $\mathbf{X} = \mathbb{R}^d$, and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the usual Euclidean dot product in \mathbb{R}^d ; that is, if $x = (x_1, x_2, \dots, x_d)$, and $y = (y_1, y_2, \dots, y_d)$, the dot product defined as

$$\langle x, y \rangle = \sum_{i=1}^d x_i \cdot y_i.$$

If we take φ as the identity operator in \mathbb{R}^d , then the function k defined as follows is a kernel function

$$k(x, x') = \exp(\langle x, x' \rangle),$$

as we can use the Taylor expansion of the exponential function and use the fact that the countable sum of kernels is again a kernel.

Definition 1.6. The space l_p of p -summable sequences is defined as the set of all real sequences $a = (a_i)_{i=1}^{\infty}$ for which

$$\|a\|_{l_p} := \left(\sum_{i=1}^{\infty} |a_i|^p \right)^{\frac{1}{p}} < \infty.$$

Note that the last norm is the same as the one in example 1.1 when μ is the counting measure.

Kernels can be defined in terms of sequences in l_p , as shown below.

Lemma 1.1. Let \mathbf{X} be non-empty set. Suppose that $\varphi_i: \mathbf{X} \rightarrow \mathbb{R}$, for $i \geq 1$, and that $\varphi_i(x)$ is a sequence of functions in l_2 , then

$$k(x, x') := \sum_{i=1}^{\infty} \varphi_i(x) \varphi_i(x') \tag{1.2}$$

is a well-defined kernel function on \mathbf{X} .

Proof. Lets first see that k is indeed a kernel function. Take as the Hilbert space \mathcal{H} the set of infinite convergent real sequences in l_2 , define the function $\varphi: \mathbf{X} \rightarrow \mathcal{H}$ as the infinite sequence $\varphi(x) = (\varphi_i(x))_{i=1}^{\infty}$. For the inner product, take the one defined on equation (1.1). By definition we see that k is indeed a kernel function.

Now, using the Cauchy-Schwartz inequality, for any $x, x' \in \mathbf{X}$, we have

$$\begin{aligned} |k(x, x')| &= \left| \sum_{i=1}^{\infty} \varphi_i(x) \varphi_i(x') \right| = |\langle \varphi(x), \varphi(x') \rangle| \\ &\leq \|\varphi(x)\|_{l_2} \|\varphi(x')\|_{l_2} < \infty. \end{aligned}$$

So, the kernel in (1.2) is well defined for all $x, x' \in \mathbf{X}$. □

Problem 1.1. (Gaussian kernel) The Gaussian kernel in \mathbb{R}^d is defined as

$$k(x, x') := \exp\left(-\frac{1}{2h^2} \|x - x'\|^2\right), \tag{1.3}$$

where h is a positive constant known as *bandwidth*. Here, we are using the standard Euclidean norm in \mathbb{R}^d , $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$, for $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$.

Proof. We take as $\mathbf{X} = \mathbb{R}^d$, and the Hilbert space $\mathcal{H} = L^2(\mathbb{R}^d, \gamma_d)$, the space of square complex-integrable functions on \mathbb{R}^d with respect to the standard Gaussian measure,

$$L^2(\mathbb{R}^d, \gamma_d) := \left\{ f: \mathbb{R}^d \rightarrow \mathbb{C} \text{ s.t. } \int_{\mathbb{R}^d} |f(t)|^2 e^{-\|t\|^2/2} dt < \infty \right\}.$$

The inner product on this Hilbert space is given by

$$\langle f, g \rangle_{L^2(\mathbb{R}^n, \gamma_n)} := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} \overline{f(t)} g(t) e^{-\|t\|^2/2} dt.$$

Take the function $\varphi: \mathbb{R}^d \rightarrow L^2(\mathbb{R}^d, \gamma_d)$ as defined by $\varphi(x)(t) = e^{i\langle x, t \rangle / \sqrt{h}}$, where $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product in \mathbb{R}^d (remember that $x \in \mathbb{R}^d$ is fixed and the image of it under φ must be a function in $L^2(\mathbb{R}^d, \gamma_d)$).

It is not hard to see that φ is square integrable with respect to the standard Gaussian measure because the modulus of φ is always 1. In fact, this means that $\|\varphi(x)\|_{L^2(\mathbb{R}^d, \gamma_d)} = 1$.

Now, we need to prove that φ does what it is supposed to do. Let $x, x' \in \mathbf{X}$

$$\begin{aligned} \langle \varphi(x), \varphi(x') \rangle_{L^2(\mathbb{R}^d, \gamma_d)} &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} \overline{e^{i\langle x, t \rangle / \sqrt{h}}} e^{i\langle x', t \rangle / \sqrt{h}} e^{-\|t\|^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} e^{-i\langle x, t \rangle / \sqrt{h}} e^{i\langle x', t \rangle / \sqrt{h}} e^{-\|t\|^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} e^{i\langle \frac{x' - x}{\sqrt{h}}, t \rangle - \|t\|^2/2} dt \\ &\stackrel{\star}{=} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} e^{-\|t - i\frac{(x' - x)}{\sqrt{h}}\|^2/2} e^{-\frac{\|x' - x\|^2}{2h}} dt \\ &= e^{-\frac{\|x - x'\|^2}{2h}} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} e^{-\|t - i\frac{(x' - x)}{\sqrt{h}}\|^2/2} dt \\ &= e^{-\frac{\|x - x'\|^2}{2h}} = k(x, x'). \end{aligned}$$

Where in the equality marked with \star we have *completed the square*

$$-\frac{1}{2}\|t\|^2 + i\left\langle \frac{(y - x)}{\sqrt{h}}, t \right\rangle = -\frac{1}{2} \left[\left\| t - i\frac{(y - x)}{\sqrt{h}} \right\|^2 \right] - \frac{1}{2} \left[\left\| \frac{(y - x)}{\sqrt{h}} \right\|^2 \right].$$

The last integral is hard to see immediately, but one can use a complex-analytic argument to show that

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} e^{-\|t - i\frac{(x' - x)}{\sqrt{h}}\|^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}^d} e^{-\|t\|^2/2} dt = 1.$$

□

1.1.2 Positive definiteness and reproducing kernels

Definition 1.7. (Positive definite functions) A symmetric function $k: \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{C}$ is called *positive definite* if for $\forall n \geq 1, \forall (a_1, a_2, \dots, a_n) \in \mathbb{R}^n, \forall (x_1, x_2, \dots, x_n) \in \mathbf{X}^n$,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) \geq 0.$$

The function $k(\cdot, \cdot)$ is *strictly positive definite* if for mutually distinct x_i , the equality holds only when all the a_i are zero.

Kernels are useful because they give us a way to construct a covariance matrix for a given set of points $X = \{x_1, x_2, \dots, x_n\}$. We define the Gram matrix point-wise as $\mathbf{K} := [k(x_i, x_j)]_{i,j}$. Remember that every covariance matrix must be positive definite, which is why we need to ensure that the kernel functions always outputs matrices of this type. The next proposition states that every Gram matrix constructed with a kernel function is always positive definite.

Proposition 1.1. Let \mathcal{H} be any Hilbert space, \mathbf{X} a non-empty set and $\varphi: \mathbf{X} \rightarrow \mathcal{H}$. Then, the function $k(x, y) := \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$ is positive definite.

Proof. Let $a_i \in \mathbb{R}$, for $1 \leq i \leq n$. By the definition of a kernel function, and the bilinearity of the inner product, we have

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \langle a_i \varphi(x_i), a_j \varphi(x_j) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^n a_i \varphi(x_i), \sum_{j=1}^n a_j \varphi(x_j) \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^n a_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geq 0. \end{aligned}$$

□

Reproducing kernel Hilbert space (RKHS)

Definition 1.8. Let \mathcal{H} be a Hilbert space of functions $f: X \rightarrow \mathbb{R}$ defined on a non-empty set X . For a fixed $x \in X$, an *evaluation functional* is a linear function $F_x: \mathcal{H} \rightarrow \mathbb{R}$ that evaluates each element of the space at the given point x ; that is,

$$F_x[f] = f(x).$$

Definition 1.9. (Reproducing kernel Hilbert space) A Hilbert space \mathcal{H} of functions defined on a non-empty set X , $f: X \rightarrow \mathbb{R}$, is said to be a *Reproducing Kernel Hilbert Space (RKHS)* if the evaluation functional F_x is continuous for all $x \in X$.

Note that the last definition does not state anything related to kernels. We will now define what we meant by a *reproducing kernel*, and we will describe how it is related to the previous definition.

Definition 1.10. (Reproducing kernel) Let \mathcal{H} be a Hilbert space of \mathbb{R} -valued functions defined on a non-empty set X . A function $k: X \times X \rightarrow \mathbb{R}$ is called a *reproducing kernel* of \mathcal{H} if it satisfies

1. $\forall x \in X, k(\cdot, x) \in \mathcal{H}$,
2. $\forall x \in X, \forall f \in \mathcal{H}, \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$. (reproducing property).

In particular, note that for any $x, y \in X$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}.$$

1.2 Regression with Gaussian processes

1.2.1 Gaussian processes

Here we define one of the primary tools that we have used in this work, the *Gaussian Processes* (GPs) on the real plane, and in both the conditional and unconditional case.

This type of stochastic process is important because it is used extensively in different areas of statistical machine learning. This confirms one of the principal ingredients in Bayesian optimization, as shown in [Frazier \(2018\)](#), [Frazier and Wang \(2015\)](#). An application in finance of GPs can be found here [Gonzalvez, Lezmi, Roncalli, and Xu \(2019\)](#).

Definition 1.11. A *Gaussian process* $\{X_t\}_{t \in T}$ indexed by a set T is a family of (real-valued) random variables X_t , all defined on the same probability space, such that for any finite subset $F \subseteq T$ the random vector $\{X_t\}_{t \in F}$ has a (possibly degenerate) Gaussian distribution. If these finite-dimensional distributions are all non-degenerate, then the Gaussian process is said to be non-degenerate.

A *Gaussian process* whose index set is not \mathbb{R} or \mathbb{Z} (e.g. a topological manifold \mathcal{M} with more than one dimension) is usually called a *Gaussian random field*.

An alternative equivalent definition for a Gaussian process is given in the following.

Definition 1.12. A collection of random variables $\{X_t\}_{t \in T}$ is a Gaussian processes if, and only if, every finite linear combination $\sum_{t \in F} \alpha_t X_t$, for some real constants α_t , is either identically zero or its distributed as a (real) Gaussian distribution.

It is well-known that, by employing the Kolmogorov's Consistency Theorem, a Gaussian process is determined by its mean and variance. We can create a Gaussian process over any space \mathbf{X} , given a mean function $\mu: \mathbf{X} \rightarrow \mathbb{R}$, and a kernel function $k: \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}_+$, as follows.

Definition 1.13. A stochastic process a over domain \mathbf{X} with mean function μ and covariance kernel k is a Gaussian process if, and only if, for any $\{x_1, \dots, x_n\} \in \mathbf{X}$ and $n \in \mathbb{N}$ the distribution of $\mathbf{f} := \{f(x_1), \dots, f(x_n)\}^\top$ is

$$\mathbf{f} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_1, x_n) & \cdots & k(x_n, x_n) \end{bmatrix} \right).$$

Given that k is defined to be a kernel function, then by [Lemma 1.1.2](#) the matrix $[k(x_i, x_j)]_{1 \leq i, j \leq n}$ is positive-definite. Consequently, there exists a multivariate normal distribution with mean μ and covariance matrix $\mathbf{K} := [k(x_i, x_j)]_{1 \leq i, j \leq n}$.

Now, let us remember the following proposition, which is useful to calculate the parameters of a conditional Gaussian processes.

Proposition 1.2. Let the $p \times 1$ random vector $y \sim N_p(\mu, \Sigma)$, let A be any $p \times k$ constant matrix, with rank $k \leq p$. If b is any $k \times 1$ vector of constants, then

$$Z = Ay + b \sim N_k(A\mu + b, A\Sigma A').$$

We now establish some notation that will be used in the rest of this work. From [Rencher and Schaalje \(2008\)](#) we have the following results.

Suppose that the random vector \mathbf{v} is partitioned into two subsets of variables, which we denote by \mathbf{x} and \mathbf{y} :

$$\mathbf{v} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p, \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_q)^\top \in \mathbb{R}^{p+q}.$$

Thus, there are $p + q$ random variables in \mathbf{v} . The mean vector and covariance matrix for \mathbf{v} partitioned as above can be expressed in the following form

$$\begin{aligned} \mu &= \mathbb{E}[\mathbf{v}] = \mathbb{E} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbb{E}[\mathbf{x}] \\ \mathbb{E}[\mathbf{y}] \end{pmatrix} = \begin{pmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{pmatrix}, \\ \Sigma &= \text{cov}(\mathbf{v}) = \text{cov} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix}, \end{aligned}$$

with $\Sigma_{\mathbf{x},\mathbf{y}} = \text{cov}(\mathbf{x}, \mathbf{y}) = \mathbb{E}[(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{y} - \mu_{\mathbf{y}})]$. Note that the matrix Σ is symmetric as the matrices $\Sigma_{\mathbf{xy}}^\top = \Sigma_{\mathbf{yx}}$.

Theorem 1.2. If \mathbf{y} and \mathbf{x} are vectors with jointly multivariate normal distribution with $\Sigma_{\mathbf{yx}} \neq \mathbf{0}$, then the conditional distribution of \mathbf{y} given \mathbf{x} , $f(\mathbf{y}|\mathbf{x})$, is multivariate normal with mean vector and covariance matrix

$$\begin{aligned} \mathbb{E}[\mathbf{y}|\mathbf{x}] &= \mu_{\mathbf{y}} + \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}}), \\ \text{cov}[\mathbf{y}|\mathbf{x}] &= \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}. \end{aligned}$$

Proof. The conditional density of \mathbf{y} given \mathbf{x} is

$$f(\mathbf{y}|\mathbf{x}) = \frac{g(\mathbf{x}, \mathbf{y})}{h(\mathbf{x})},$$

where $g(\mathbf{y}, \mathbf{x})$ is the joint density of \mathbf{y} and \mathbf{x} , and $h(\mathbf{x})$ is the marginal density of \mathbf{x} . Consider the function

$$\begin{pmatrix} \mathbf{w} \\ \mathbf{u} \end{pmatrix} = A \left[\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} - \begin{pmatrix} \mu_{\mathbf{y}} \\ \mu_{\mathbf{x}} \end{pmatrix} \right], \quad (1.4)$$

where

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} I_p & -\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1} \\ 0 & I_q \end{pmatrix}.$$

Note that I_p is the $p \times p$ identity matrix, similarly I_q is the $q \times q$ identity matrix. By

simplifying equation (1.4), we obtain

$$\begin{aligned}\mathbf{w} &= \mathbf{y} - [\mu_{\mathbf{y}} + \Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}})], \\ \mathbf{u} &= \mathbf{x} - \mu_{\mathbf{x}}.\end{aligned}$$

Using the multivariate change-of-variable technique, the joint density of (\mathbf{w}, \mathbf{u}) is

$$p(\mathbf{w}, \mathbf{u}) = g(\mathbf{y}, \mathbf{x})|A^{-1}| = g(\mathbf{y}, \mathbf{x}).$$

Similarly, the marginal density of \mathbf{u} is

$$q(\mathbf{u}) = h(\mathbf{x})|I_q^{-1}| = h(\mathbf{x}).$$

By calculating the covariance of \mathbf{u}, \mathbf{w} , we get

$$\text{cov}(\mathbf{w}, \mathbf{u}) = A_1\Sigma A_2 = \Sigma_{\mathbf{y}\mathbf{x}} - \Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}\Sigma_{\mathbf{x}\mathbf{x}} = \mathbf{0}.$$

Because the vectors are normally distributed, the last equation implies that \mathbf{u} and \mathbf{w} are independent. Hence

$$p(\mathbf{w}, \mathbf{u}) = r(\mathbf{w})q(\mathbf{u}),$$

where $r(\mathbf{w})$ is the density of \mathbf{w} . Given that $p(\mathbf{w}, \mathbf{u}) = g(\mathbf{y}, \mathbf{x})$ and $q(\mathbf{u}) = h(\mathbf{x})$, we also have

$$g(\mathbf{y}, \mathbf{x}) = r(\mathbf{w})h(\mathbf{x}),$$

by (1.4)

$$r(\mathbf{w}) = \frac{g(\mathbf{y}, \mathbf{x})}{h(\mathbf{x})} = f(\mathbf{y}|\mathbf{x}).$$

Hence, we obtain $f(\mathbf{y}|\mathbf{x})$ simply by finding $r(\mathbf{w})$. By Corollary 1.2.1, we have that $r(\mathbf{w})$ is the multivariate normal density with mean and covariance matrix

$$\begin{aligned}\mu_{\mathbf{w}} &= A_1 \left[\begin{pmatrix} \mu_{\mathbf{y}} \\ \mu_{\mathbf{x}} \end{pmatrix} - \begin{pmatrix} \mu_{\mathbf{y}} \\ \mu_{\mathbf{x}} \end{pmatrix} \right] = \mathbf{0}, \\ \Sigma_{\mathbf{w}\mathbf{w}} &= A_1\Sigma A_1' \\ &= (I_p, -\Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}) \begin{pmatrix} \Sigma_{\mathbf{y}\mathbf{y}} & \Sigma_{\mathbf{y}\mathbf{x}} \\ \Sigma_{\mathbf{x}\mathbf{y}} & \Sigma_{\mathbf{y}\mathbf{y}} \end{pmatrix} \begin{pmatrix} I_p \\ -\Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1} \end{pmatrix} \\ &= \Sigma_{\mathbf{y}\mathbf{y}} - \Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}\Sigma_{\mathbf{x}\mathbf{y}}.\end{aligned}$$

Thus $r(\mathbf{w}) = r(\mathbf{y} - [\mu_{\mathbf{y}} + \Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}})])$ is of the form $N_p(0, \Sigma_{\mathbf{y}\mathbf{y}} - \Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}\Sigma_{\mathbf{x}\mathbf{y}})$. Equivalently, $\mathbf{y}|\mathbf{x} \sim N_p(\mu_{\mathbf{y}} + \Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}}), \Sigma_{\mathbf{y}\mathbf{y}} - \Sigma_{\mathbf{y}\mathbf{x}}\Sigma_{\mathbf{x}\mathbf{x}}^{-1}\Sigma_{\mathbf{x}\mathbf{y}})$

□

Corollary 1.6. Suppose that $X = (X_1, X_2)$ is jointly d -dimensional Gaussian, with $X_1 \in \mathbb{R}^n$ and $X_2 \in \mathbb{R}^m$, $n < m$, and $n + m = d$. Assume that the distribution of the mean

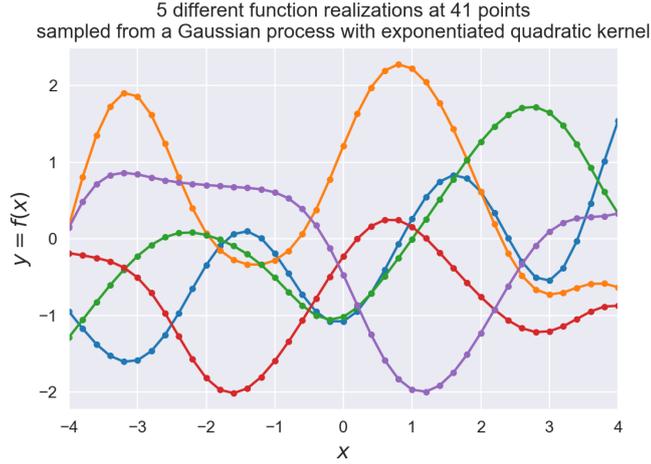


Figure 1.1: Sample points from MV normal distribution.

vector, and covariance matrix are given, respectively, by

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then, if Σ_{22}, Σ_{11} are invertible, the conditional distribution of X_1 given X_2 is normal distributed. In fact, $X_1|X_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$, with

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \quad (1.5)$$

This formula states that the terms $\mu_{1|2}$ and $\Sigma_{1|2}$ are a linear combination of the parameters of X_1 and X_2 . Thus, the parameters of the conditional distribution $X_1|X_2$ are fairly easily to calculate on a computer. The only computational bottleneck here is the inversion of a matrix which has order $\mathcal{O}(n^2)$, for a $n \times n$ matrix.

Example 1.5. Suppose $\mathbf{X} = (x_1, x_2, \dots, x_{41})^\top$ are 41 ordered points sampled from the interval $[-4, 4]$. Using these points, and the quadratic kernel function (1.3), we can construct the covariance matrix $\mathbf{K} = (k(x_i, x_j))$ for $1 \leq i, j \leq 41$. By taking $\mathbf{0}_{41} = (0, 0, \dots, 0)^\top \in \mathbb{R}^{41}$ and \mathbf{K} we take five samples with distribution $\mathcal{N}(\mathbf{0}_{41}, \mathbf{K})$. We show these five samples, with five different colors, in Figure 1.1. Note that we fix the set \mathbf{X} for all the samples, and for each sample set we have a different dataset of the form $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{41}$, with $x_i \in [-4, 4]$, and $y_i = f(x_i)$ is normally distributed.

1.2.2 Gaussian processes for regression

Gaussian processes can be used to model distributions over functions, as shown in Mukherjee (2015), Rasmussen and Williams (2005). Therefore, they can be used to build regression models. Placing a multivariate normal distribution as a prior over the observed data will give a posterior distribution, which can then be used to predict the expected posterior value of the process.

Consider some dataset (training set) of $n \in \mathbb{N}$ points $D = \{(x_i, y_i)\}_{i=1}^n$ drawn from the model, where $x_1, x_2, \dots, x_n \in \mathbb{R}^d$, for some $d \geq 0$, and $y_i \in \mathbb{R}$, for $1 \leq i \leq n$. Suppose that there exists a relation, possibly non-linear, between the observations x_i and their corresponding *response* variables y_i . Suppose that the relationship is approximated by a real function f with normal noise; in other words

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2), \quad f \in \mathcal{H}_K.$$

With \mathcal{H}_K , a Hilbert space of functions can be defined over a non-empty set K .

Now imagine that we are also given $m \in \mathbb{N}$ new observations (or test data) $T = \{x_i^*\}_{i=1}^m$ each of which would have a corresponding y_i^* .

To do the estimation, we place a prior on the space of functions using a Gaussian process.

$$f \sim \mathcal{GPs}(\mu(\cdot), k(\cdot, \cdot)),$$

where $k: V \times V \rightarrow \mathbb{R}_+$ is a kernel function, and V a vector space over \mathbb{R} .

By rearranging this data in matrix form, we have

$$\mathbf{X} = \begin{bmatrix} -x_1- \\ \vdots \\ -x_n- \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} -x_1^*- \\ \vdots \\ -x_m^*- \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{Y}^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_m^* \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\varepsilon}^* = \begin{bmatrix} \varepsilon^* \\ \vdots \\ \varepsilon_m^* \end{bmatrix}, \quad f(\mathbf{X}) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad f(\mathbf{X}^*) = \begin{bmatrix} f(x_1^*) \\ \vdots \\ f(x_m^*) \end{bmatrix}.$$

Note that, $\mathbf{X} \in M_{n \times d}(\mathbb{R})$, $\mathbf{X}^* \in M_{m \times d}(\mathbb{R})$, $\mathbf{Y} \in M_{n \times 1}(\mathbb{R})$, and $\mathbf{Y}^* \in M_{m \times 1}(\mathbb{R})$. Because f is a real function, we have (entry-wise) $\mathbf{f} := f(\mathbf{X}) \in M_{n \times 1}(\mathbb{R})$, and $\mathbf{f}^* := f(\mathbf{X}^*) \in M_{m \times 1}(\mathbb{R})$.

Our ultimate goal is to specify the predictive distribution on \mathbf{Y}^* , which we know is a multivariate normal

$$\{\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{X}\} \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*).$$

Note that,

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}^* \end{bmatrix} \mid \mathbf{X}^*, \mathbf{X} = \begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I} & k(\mathbf{X}, \mathbf{X}^*) \\ k(\mathbf{X}^*, \mathbf{X}) & k(\mathbf{X}^*, \mathbf{X}^*) + \sigma^2 \mathbf{I} \end{bmatrix}\right),$$

where $k(\mathbf{X}, \mathbf{X})$ is the $n \times n$ matrix with $\mathbf{K}_{ij} = k(x_i, x_j)$ and $k(\mathbf{X}^*, \mathbf{X}^*)$ is the $m \times m$ matrix with $\mathbf{K}_{ij}^* = k(x_i^*, x_j^*)$.

To get a predictive distribution on \mathbf{Y}^* we write the conditional $\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{X}$. Given this multivariate normal distribution, we simply condition over all the other variables to get the mean and covariance for the normal posterior predictive density. Obtaining the mean

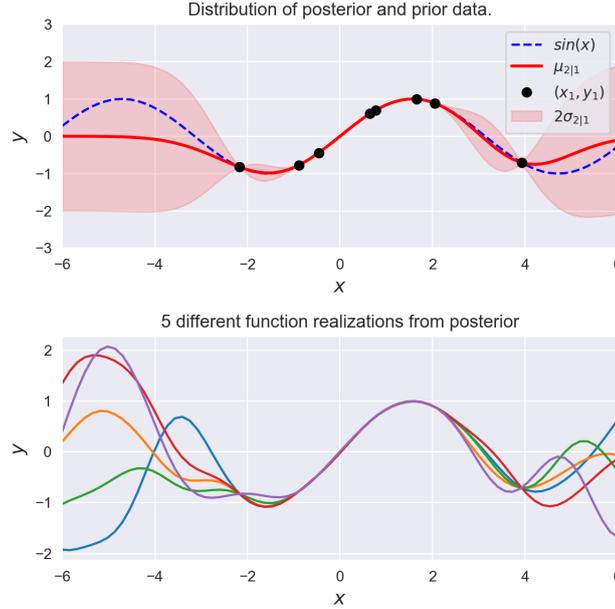


Figure 1.2: Example of a conditional process

and covariance matrix

$$\begin{aligned}\mu^* &= k(\mathbf{X}^*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} \mathbf{Y}, \\ \Sigma^* &= k(\mathbf{X}^*, \mathbf{X}^*) + \sigma^2 \mathbf{I} - k(\mathbf{X}^*, \mathbf{X})(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I})^{-1} k(\mathbf{X}, \mathbf{X}^*).\end{aligned}\tag{1.6}$$

For instance, if we want to do inference on just one new observation x^* , then we can calculate the distribution of f^* as follows

$$\begin{aligned}p(f^*|x^*, Y, X) &= \int_w p(f^*|x^*, w)p(w|y, X) dw, \\ w &\sim N(0, \Sigma).\end{aligned}$$

One estimate for f^* could be the *maximum a posteriori* (MAP) estimator.

Remark: Unless otherwise stated, we will assume that $\sigma = 0$; that is, we will not consider noise at this time.

Example 1.6. We calculated the posterior distribution based on eight observations, which comes from the sine trigonometric function. The results are plotted in Figure 1.2. The top figure shows the distribution, where the red line is the posterior mean, the light red area is the 95% prediction interval, the black dots are the observations (X_1, Y_1) . The prediction interval is computed from the standard deviation $\sigma_{2|1}$, which is the square root of the diagonal of the covariance matrix. The bottom figure shows five realizations (sampled functions) from this distribution. Note that the distribution is quite confident of the points predicted around the observations (X_1, Y_1) , and that the prediction interval gets larger the further away it is from these points. [Roelants \(2019\)](#).

Example 1.7. In Figure 1.3 we can see different plots of the same 3-dimensional Gaussian

processes from different angles. Where we sampled uniform points from the squared $[-5, 5] \times [-5, 5]$ to construct the covariance matrix \mathbf{K} .

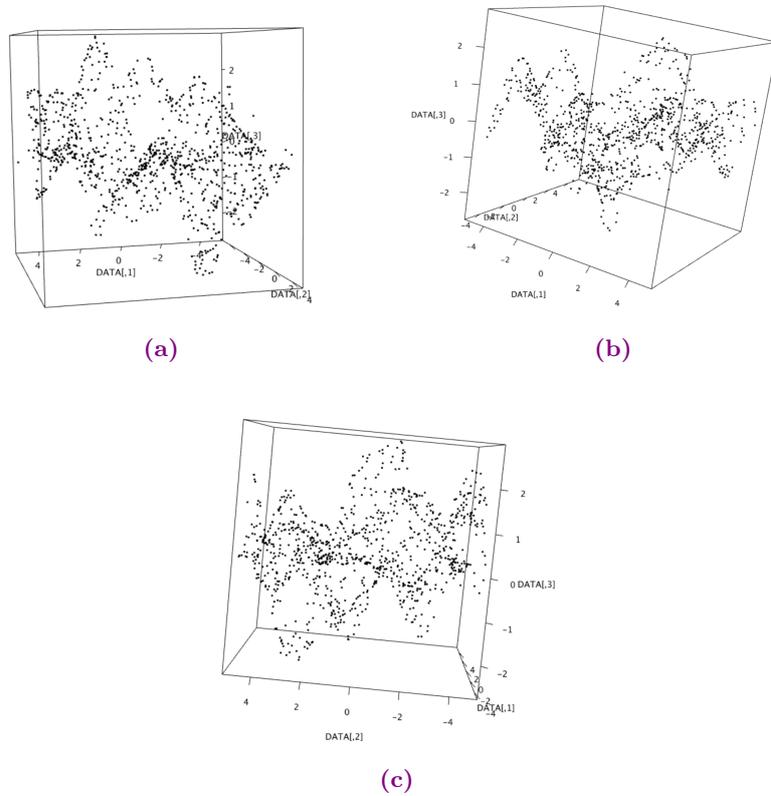


Figure 1.3: 3D Gaussian process embedded in \mathbb{R}^3 .

1.3 Exceedances of the Poisson point process

Here we introduce some tools that will allow us study the exceedances of conditional Gaussian processes, over a given threshold value.

1.3.1 Poisson point process in the plane

A counting process $\{N(t), t \geq 0\}$ is said to constitute a two-dimensional non-homogeneous Poisson process on $C \subseteq \mathbb{R}^2$ with rate $\lambda(C) \geq 0$ if:

1. The random variable $\mathcal{P}(R)$ counting the number of events in a region $R \subseteq C$ is Poisson distributed with parameter $\Lambda(R) = \int_R \lambda(V) dV$

$$\mathbb{P}(\mathcal{P}(R) = k) = e^{-\Lambda(R)} \frac{\Lambda(R)^k}{k!}, \quad k = 0, 1, 2, \dots$$

2. The number of events occurring in any finite set of non-overlapping regions are mutually independent; that is, for disjoint bounded Borel sets R_1, R_2, \dots, R_n the random variables $\mathcal{P}(R_1), \mathcal{P}(R_2), \dots, \mathcal{P}(R_n)$ are independent.

In case $\lambda(C) = \lambda$, \mathcal{P} is called a homogeneous Poisson point process with intensity λ .

Palm theory may be used later for an approximation of the sample obtained from the level sets [Coeurjolly, Møller, and Waagepetersen \(2017\)](#).

1.3.2 Point approximation approach

If we consider an unknown distribution function F of a random variable X , then we are interested in estimating the distribution function F_u , which we will define later, of the variable X above a certain threshold u . The distribution function F_u is called the conditional excess distribution function and is defined as:

$$F_u(y) = P(X - u \leq y | X > u), 0 \leq y \leq x_F - u,$$

where X is a random variable, u is a given threshold, $y = x - u$ are the excesses, and x_F is the right endpoint of F .

It is easy to verify that F_u can be written as:

$$F_u(y) = \frac{F(u + y) - F(u)}{1 - F(u)}.$$

[Pickands \(1971\)](#) posed that for a large class of underlying distribution function F the conditional excess distribution function $F_u(y)$, for u large, is well approximated by

$$F_u(y) \approx F(y; k, \sigma), \text{ as } u \rightarrow \infty,$$

where

$$F(y; k, \sigma) = \begin{cases} 1 - \left(1 - \frac{ky}{\sigma}\right)^{\frac{1}{k}}, & k \neq 0, \sigma > 0, \\ 1 - e^{-x/\sigma}, & k = 0, \sigma > 0. \end{cases}$$

For $0 \leq y \leq x_F - u$, if $k \leq 0$, and $0 \leq y \leq \sigma/k$, if $k > 0$. $F(x; k, \sigma)$ is the generalized Pareto distribution. [Smith \(1989\)](#) applied this theory to trend detection for ground-level ozone levels.

1.4 Goodness of fit test for Poisson distribution

In this section, we present a goodness of fit test that does not require independence among the sample points, only equal distribution, to test if a sample of random variables comes from a Poisson distribution. This will be useful for our simulation study because the samples that we are working with are not necessarily statistically independent.

We will employ a tool suggested by [Nakamura and Pérez-Abreu \(1993\)](#) as a graphical technique for goodness of fit test for the Poisson distribution.

Proposition 1.3. If X is a Poisson random variable with rate parameter $\lambda \geq 0$, then the probability generating function (p.g.f) of X is given by

$$\varphi_X(t) = e^{-\lambda(1-t)}, \forall t \in [-1, 1]. \quad (1.7)$$

Note that for $t \in (0, 1)$, we can define the cumulant function of X by

$$\kappa(t) := \log(\varphi_X(t)) = \lambda(1 - t).$$

Definition 1.14. (Empirical probability generating function). Given a random sample X_1, X_2, \dots, X_n , the empirical probability generating function (e.p.g.f.) is defined as

$$\hat{\varphi}(t) = \frac{1}{n} \sum_{i=1}^n t^{X_i}, \forall t \in (-1, 1). \quad (1.8)$$

Observe that if a sample X_1, X_2, \dots, X_n comes from a Poisson distribution, then the graph of the log empirical generating function

$$\kappa(t) = \log(\hat{\varphi}(t)) = \log\left(\frac{1}{n} \sum_{i=1}^n t^{X_i}\right), \forall t \in (0, 1), \quad (1.9)$$

should be similar to a straight line. Although this is not a goodness of fit for a Poisson distribution of a given parameter λ , it is a graphical test for a Poisson distribution.

Note that equation (1.8) was defined for all of the values in the interval $(-1, 1)$, whereas function (1.9) was only defined for points $x \in (0, 1)$. This happens because the log function is only defined for positive numbers (we are working only with real numbers). If we take a negative value for t , then it could be the case that the sample has more odd numbers than even numbers on the sample set X_1, \dots, X_n . Hence, we would obtain a negative value for $\hat{\varphi}(t)$, for which we cannot take its logarithm.

2 Betti numbers computation

Introduction

This chapter aims to introduce the elements and notation from algebraic topology that we need for our simulation study. We start by defining the notion of simplicial complexes, which are the generalizations of graphs in higher dimensions. By using the concept of simplicial complexes, we later define what Betti numbers are and how they can be calculated using a computer.

The Betti numbers can be interpreted as a topological summary of a certain set. Given that our objective is to understand the levels set topology of conditional Gaussian fields, we can use Betti numbers to approximate the true topological structure of such level sets. Intuitively, the k -th Betti number refers to the number of k -dimensional holes on a topological surface. This is a fundamental idea that is further developed and employed in Chapter 3.

2.1 Topological definitions

We start by establishing the notation, based on [Adler and Taylor \(2003\)](#) and [Reveles, Pérez-Abreu, Nakamura, and Biscay \(2016\)](#).

2.1.1 Simplicial complexes

One of the principal elements in the area of topological data analysis is the notion of abstract simplicial complexes. To infer topological insights from the data, we use simplicial complexes to approximate the homology of a data cloud.

A standard k -simplicial complex is a k -dimensional generalization of a graph in higher dimensions. The mathematical definition follows.

Definition 2.1. A **standard k -simplex** Δ^k is the set in \mathbb{R}^{k+1} defined as

$$\Delta^k := \left\{ (t_0, t_1, \dots, t_k) \in \mathbb{R}^{k+1} : \sum_{i=0}^k t_i = 1, t_i \geq 0, \text{ for all } i \right\}.$$

We can also define a simplex for arbitrary points in the plane, as follows:

Definition 2.2. (k -dimensional simplex) Given a set of $(k+1)$ points $X = \{x_0, x_1, \dots, x_k\}$ in \mathbb{R}^d affine independent, the simplex $\sigma = [x_0, x_1, \dots, x_k]$ generated by X , is the convex hull of X . The points in X are known as the vertices of σ , and the simplexes generated by the subsets of X are called the faces of X .

For a simplex σ generated by X , its dimension is equal to the cardinality of X minus one; that is, $\dim(\sigma) = \text{card}(X) - 1$. Simplexes of dimensions 0, 1, 2, and 3 are known as vertices, edges, triangles, and tetrahedrons, respectively.

If we want to employ simplexes to approximate the topology of a dataset, then we need to use many of them. This is the reason why we need to define a way to differentiate one from another. Next, we define a structure for a set of simplexes.

Definition 2.3. An *abstract simplicial complex* K is a set of simplexes such that:

- i. Every face of a simplex from K is also in K .
- ii. The non-empty intersection of any two simplexes $\sigma_i, \sigma_j \in K$ is a face of both simplexes σ_i and σ_j .

We call this an abstract simplicial complex because we are not thinking about its geometrical representation in the Euclidean space but as a whole with the property of being closed under the operation of taking non-empty subsets.

By the previous definition, we know that all of the different subsets of K are the simplexes of K , and that the dimension of a simplex σ is $\dim(\sigma) = \text{card}(\sigma) - 1$.

With these two definitions, we can now define the dimension of an abstract simplicial complex as a collection of simplexes.

Definition 2.4. (Dimension of an abstract simplicial complex) The dimension of an abstract simplicial complex is defined as

$$\dim(K) := \max_{\sigma \in K} \dim(\sigma).$$

Observe that we can select all of the faces of a simplex σ whose faces are of a certain dimension p , for some fixed number p . These are known as p -faces or p -simplexes of σ . If σ is a simplex on a simplicial complex K , by definition, then all of the p -faces of σ are also contained in K . The union of all the p -faces is known as the p -skeleton of K . Formally, the definition is the following:

Definition 2.5. (Skeleton) Let K be an abstract simplicial complex and a non-negative $p \leq \dim(K)$. The p -Skeleton of K is defined as

$$\text{Sk}_p(K) := \bigcup_{\sigma \in \mathcal{H}_p} \sigma,$$

where $\mathcal{H}_p := \{\sigma \in K : \dim(\sigma) = p\}$.

Definition 2.6. (Čech complex) Let ρ a metric defined on \mathbb{R}^n . For $r \geq 0$, let $B_r(x) = \{y \in \mathbb{R}^n : \rho(x, y) \leq r\}$, the closed ball of radius r with center at $x \in X$. The Čech complex for a finite set $X \subseteq \mathbb{R}^n$ of radius r , is defined as

$$C(X) = \left\{ Q \subseteq X : \bigcap_{x \in Q} B_r(x) \neq \emptyset \right\}.$$

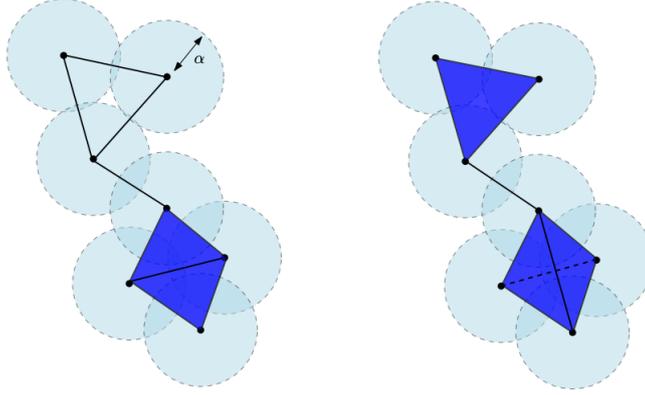


Figure 2.1: Čech complex to the left-hand side, and Vietoris-Rips complex to the right-hand side (Chazal et al.).

Computationally speaking, the Čech complex is hard to compute because we need to check all of the possible intersections. For each $x \in X$, we have to check if the intersection of the ball centered at x with the family of balls centered at $X \setminus \{x\}$ is not empty.

We can avoid this problem by modifying a bit the last definition. Instead of checking all of the possible intersections at the same time, we can do it pairwise. These are known as Vietoris-Rips complexes. The formal definition follows.

Definition 2.7. (Vietoris-Rips Complex) Let X a set of points in the metric space (M, ρ) and α a non-negative real number. A Vietoris-Rips simplicial complex, denoted by $\text{VR}(X, \alpha)$, is a set of simplexes $[x_0, x_1, \dots, x_k]$ such that $\rho(x_i, x_j) \leq \alpha$, for all values $i, j \in \{1, 2, \dots, k\}$.

It follows immediately that both Čech and VR complexes are indeed abstract simplicial complexes. Nevertheless, if X is a finite subspace of \mathbb{R}^d , then the simplicial complex $\text{VR}(X, \alpha)$ does not generally admits a geometric realization on \mathbb{R}^n ; this realization could be of dimension greater than n .

In Figure 2.1 we can see the Čech and Vietoris-Rips complex that we get using the same set of points and same radius $\alpha > 0$.

2.1.2 Chains, boundaries, and cycles

Definition 2.8. (Chain groups) Let K be a simplicial complex. A p -chain p is defined as a formal sum of p -simplexes; that is, $p = \sum_i c_i \sigma_i$, where each σ_i is a p -simplex, and $c_i \in \mathbb{F}$, for a field \mathbb{F} .

For computational simplicity, our field of election will always be $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$, which is a field because 2 is a prime number. This means that the constant values are either $c_i = 1, 0$.

We can define the sum of two p -chains as follows

$$\gamma = \gamma_1 + \gamma_2 = \sum_i c_i \sigma_i + \sum_i d_i \sigma_i = \sum_i (c_i + d_i \bmod 2) \sigma_i.$$

Definition 2.9. (p -Chain group C_p) The set of p -chains together with the addition operator defined earlier form the group of p -chains, denoted by $(C_p, +)$, or simply C_p if the sum operation is understood.

Note that this is a commutative group because the sum operator is also commutative.

Definition 2.10. The boundary of a p -simplex $\sigma = [v_0, v_1, \dots, v_p]$ is the sum of its $(p-1)$ -dimensional faces. We denote this operator by $\partial_p: C_p \rightarrow C_{p-1}$

$$\partial_p \sigma = \sum_{i=1}^p [v_0, \dots, \hat{v}_i, \dots, v_p],$$

where $[v_0, \dots, \hat{v}_i, \dots, v_p]$ denotes the simplex spanned by all vertices but v_i .

The boundary of a p -chain $c = \sum_i a_i \sigma_i$, for some $a_i \in \{0, 1\} \cong \mathbb{Z}/2\mathbb{Z}$, is defined linearly as the sum of the boundaries of its faces; that is,

$$\partial_p c = \sum_i a_i \partial_p \sigma_i.$$

We have the following convention: let K be a simplicial complex with $d = \dim K$, if $p < 0$ or $p > d$ then its chain group C_p is empty.

Lemma 2.1. (Fundamental property) The boundary operator satisfy the following: $\partial_p \partial_{p+1} \sigma = 0$, for every $(p+1)$ -simplexes $\sigma = [v_0, \dots, v_{p+1}]$

Proof. Let σ be any $(p+1)$ -simplex $\sigma = [v_0, \dots, v_{p+1}]$. Note that

$$\begin{aligned} \partial_p \partial_{p+1} \sigma &= \partial_p \sum_{i=1}^{p+1} [v_0, \dots, \hat{v}_i, \dots, v_{p+1}] = \sum_{i=1}^{p+1} \partial_p [v_0, \dots, \hat{v}_i, \dots, v_{p+1}] \\ &= \sum_{i=1}^{p+1} \sum_{j=1}^{i-1} [v_0, \dots, \hat{v}_j, \dots, \hat{v}_i, \dots, v_{p+1}] + \sum_{i=1}^{p+1} \sum_{j=i+1}^{p+1} [v_0, \dots, \hat{v}_i, \dots, \hat{v}_j, \dots, v_{p+1}] \\ &= 0. \end{aligned}$$

This is intuitively true because any $(p-1)$ -face of σ belongs to exactly two p -faces. Modulo two of the sum cancels itself and becomes zero. Hence, $\partial_p \partial_{p+1} \equiv 0$. \square

Claim 2.1. The boundary operator $\partial_p: C_p \rightarrow C_{p-1}$ is a group *homomorphism*; that is, for any p -chains $c_1, c_2 \in (C_p, +)$, we have $\partial_p(c_1 + c_2) = \partial_p(c_1) + \partial_p(c_2)$.

Definition 2.11. A p -cycle γ is a p -chain such that $\partial \gamma = 0$. The collection of all p -cycles, denoted by Z_p (which happens to be a subgroup of C_p) is called the p -th cycle group. A p -boundary is a p -chain γ that is the boundary of a $(p+1)$ -chain; that is, there exists $\gamma' \in C_{p+1}$ such that $\gamma = \partial_{p+1} \gamma'$. The collection of all p -boundaries, denoted by B_p , (which forms a subgroup of C_p) is called the p -th boundary group.

Remember that the fundamental boundary property states that any p -boundary is a p -cycle, this implies that $B_p \leq Z_p$ (B_p is a *normal* subgroup of Z_p).

Lemma 2.2. We have the following facts

1. $B_p \subseteq Z_p \subseteq C_p$.
2. For the boundary operator $\partial_p: C_p \rightarrow C_{p-1}$ have have:
 $Z_p = \ker(\partial_p)$, and $B_{p-1} = \text{im}(\partial_p)$.

2.1.3 Homology groups

The dimension of the homology groups summarizes important topological information of a data cloud, such as the k -th-dimensional holes.

Definition 2.12. (Betti numbers) The p -th homology group is $H_p := Z_p/B_p$. Under $\mathbb{Z}/2\mathbb{Z}$ -coefficients, H_p is a free commutative group, and we call its rank the p -th Betti number denoted by, $\beta_p = \text{rank}(H_p)$.

We can calculate the Betti numbers as follows

$$\beta_p := \text{rank}(H_p) = \text{rank}(Z_p) - \text{rank}(B_p). \quad (2.1)$$

Observe that equation (2.1) enables us to calculate Betti numbers on a computer if we are able to calculate the rank of the groups Z_p and B_p .

The Betti numbers have been shown to be useful in inferring insights about a manifold based solely on a random sample, as pictured in [Bobrowski et al. \(2017\)](#), [Bobrowski and Mukherjee \(2013\)](#), and [Bobrowski and Kahle \(2018\)](#).

As before, given a simplicial complex K , let C_p denote the p -chain group. We denote by n_p , z_p , and b_p the rank of the p -th chain group C_p , cycle group Z_p , and boundary group B_p , respectively. We have the following relations between these rank values.

Proposition 2.1. The following equalities holds:

1. $n_p = z_p + b_{p-1}$.
2. $\beta_p = z_p - b_p$.

Proof. Remember that for the boundary operator $\partial_p: C_p \rightarrow C_{p-1}$, we have that $Z_p = \ker(\partial_p)$, and $B_{p-1} = \text{im}(\partial_p)$. By the Rank-Nullity Theorem from linear algebra we have that

$$C_p \cong \ker(\partial_p) \oplus \text{im}(\partial_p) \quad \Rightarrow \quad C_p \cong Z_p \oplus B_{p-1},$$

which proves item 1. To prove 2, note that it follows from definition of the homology group. \square

2.2 Matrix algebraic theory

In this section, we introduce some definitions that allow us to translate the problem of the calculation of Betti numbers, the rank of the homology groups, to a linear algebra problem. This is done with the aim to use a machine to do the computations.

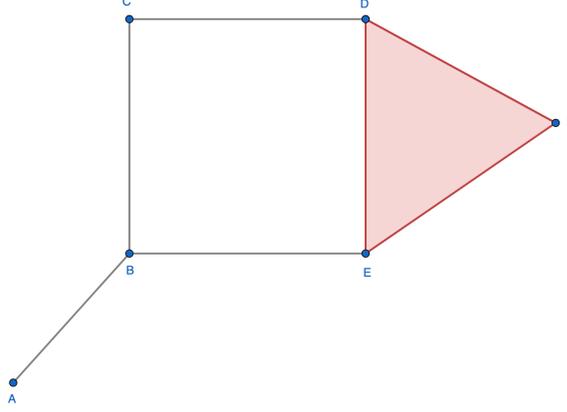


Figure 2.2: Example of a simplicial complex.

2.2.1 Boundary matrix

Definition 2.13. The p -th boundary matrix $\partial_p = [a_{i,j}]_{i,j}$ of a simplicial complex K is defined point-wise as follows

$$a_{i,j} = \begin{cases} 1, & \text{if the } i\text{-th } (p-1)\text{-simplex is a face of the } j\text{-th } p\text{-simplex;} \\ 0, & \text{otherwise.} \end{cases}$$

In other words, we order the labels of the $(p-1)$ -simplex as rows and those of the p -simplex as columns. We put a 1 if the i -th $(p-1)$ -simplex is a face of the j -th p -simplex, and zero otherwise.

Example 2.1. Suppose that $C_{n_p} = \{\alpha_1, \dots, \alpha_{n_p}\}$, and $C_{n_{p-1}} = \{\beta_1, \dots, \beta_{n_{p-1}}\}$; then, the boundary matrix $A_p = (\partial_p[i, j])_{i,j}$ has size $n_p \times n_{p-1}$. So, $\partial_p[i, j] = 1$ if and only if β_j is a face of α_i , that is $\beta_j \in \partial_p(\alpha_i)$.

Example 2.2. (Betti numbers) We can calculate the Betti numbers of a data cloud using the boundary matrix, as explained in [Edelsbrunner and Harer \(2010\)](#). In Figure 2.2, we can observe an example of a simplicial complex. Using the natural orientation, the boundary matrices are as follows:

$$\partial_2 = \begin{matrix} & \begin{matrix} DEF \end{matrix} \\ \begin{matrix} AB \\ BE \\ BC \\ CD \\ DE \\ DF \\ EF \end{matrix} & \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \end{matrix}, \quad \partial_1 = \begin{matrix} & \begin{matrix} AB & BE & BC & CD & DE & DF & EF \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{matrix}.$$

Using these matrices, we can row reduce them and then, using equation (2.1), we obtain

the following values:

$$\beta_0 = 1, \quad \beta_1 = 1.$$

Recall that the first two Betti numbers β_0, β_1 count the amount of connected components and one-dimension holes, respectively. Note that in Figure 2.2 we have one connect component and one hole, which what we got when reducing the boundary matrices.

Remark: The value of the Betti numbers is independent of the selected orientation; that is, it does not matter which orientation we choose, the Betti numbers are always going to be the same. We recommend [Reveles et al. \(2016\)](#) and [Bobrowski and Kahle \(2018\)](#) for a more detailed explanation of this fact.

2.2.2 Matrix reduction

Let K be a simplicial complex. Given a p -chain $c = \sum_i a_i \sigma_i$ its boundary ∂c can be computed by matrix multiplication

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n_p} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n_p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n_{p-1},1} & a_{n_{p-1},2} & \cdots & a_{n_{p-1},n_p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n_p} \end{bmatrix},$$

where the matrix $\partial_p = [a_{i,j}]_{i,j}$, for $1 \leq n_{p-1}$ and $1 \leq j \leq n_p$, is the p -th boundary matrix of the p -chain c . In words, a collection of columns represents a p -chain and the sum of these columns gives its boundary.

Note that the rows of ∂_p form a basis for the $(p-1)$ -st chain group C_{p-1} , and the columns form a basis of the p -st chain chain group C_p . We would like to calculate the rank of ∂_p . From linear algebra, we know that elementary row and columns operations do not change the rank of a matrix.

We can perform the following matrix operations

1. By exchanging columns k and l and adding column k to column l , we can express both by multiplying with a matrix $V = [v_{i,j}]_{i,j}$ on the right. To exchange two columns, we have $v_{l,k} = v_{k,l} = 1$ and $v_{i,i} = 1$ for all $i \neq k, l$, and all of the other entries are zero. To add column k to column l , we have $v_{k,l} = 1$, $v_{i,i} = 1$ for all i and all of the other entries are zero.
2. Similarly, we have two row operations: the first exchanges two rows and the second adds one row to another. This is represented by a left matrix multiplication by $U = [u_{i,j}]_{i,j}$. To exchange two rows, we again have $u_{k,l} = u_{l,k} = 1$, $u_{i,i} = 1$ for all $i \neq j, k$ and all the other entries zero. To add the k -th to the l -th row, we have $u_{l,k} = 1$, $u_{i,i} = 1$ for all i , and all of the other entries are zero.

To calculate the Betti number, we must reduce the boundary matrix. To do so, one can use a variety of matrix reduction algorithms, here we show one of them. First, let us remember the definition of a reduced matrix.

Algorithm 1: Matrix reduction algorithm.

```
Result: Reduce( $x$ )  
if there is  $k \geq x, l \geq x$  with  $N_p[k, l] = 1$  then  
    exchange rows  $x$  and  $k$ ; exchange columns  $x$  and  $l$ ;  
    for  $i = x + 1$  to  $n_{p-1}$  do  
        if  $N_p[i, x] = 1$  then  
            | add row  $x$  to row  $i$   
        end  
    end  
    for  $j = x + 1$  to  $n_p$  do  
        if  $N_p[x, j] = 1$  then  
            | add column  $x$  to column  $j$   
        end  
    end  
    Reduce( $x + 1$ )  
end
```

Definition 2.14. (Matrix reduction) We say that a matrix M is in *reduced* form if all of the columns are linear independents. We say that a M is in Smith normal form, for modulo two arithmetic, if it has the following structure: all initial segments of the diagonal are 1 and everything else is 0; for instance, it has to look like N_p :

$$N_p = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & 0 \\ \vdots & & & 1 & \vdots \\ & & & & 0 \\ & & & & \ddots \\ 0 & \cdots & & & 0 \end{pmatrix}.$$

To reduce a matrix M , we can follow Algorithm 1. To get the bases of the boundary and cycle groups, we keep track of the matrix products that represent the row and column operations. Writing U_{p-1} and V_p for the left and right products, we get the normal form as $N_p = U_{p-1} \partial_p V_p$. We start the reduction by initializing the matrix to $N_p[i, j] = [a_{i,j}]_{i,j}$ for all i and j . By calling the function for $x = 1$, we get the position of the considered diagonal element.

Claim 2.2 The procedure of reduction terminates in $\mathcal{O}(n_p^2 n_{p-1})$ time, and its output matrix M is in reduced form. Furthermore, the set of non-zero columns in M form a basis for B_{p-1} , and the set of columns $\{\beta_j : \text{col}_M[j] = 0\}$ form a basis for Z_p .

2.3 Use of Ripser to efficiently calculate the Betti numbers

In the summer of 2020, *Ripser* [Bauer \(2019\)](#) is the fastest library that computes topological barcodes. Ripser was written in the C/C++ programming language.

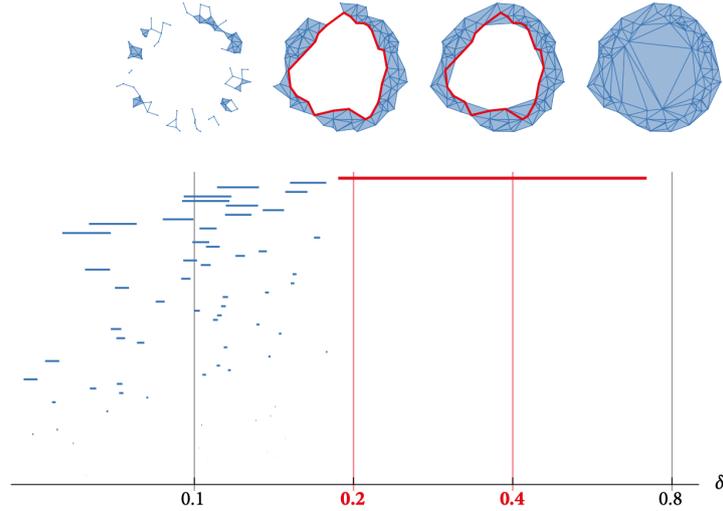


Figure 2.3: Example of a barcode. (Bauer (2019))

Algorithm 2: Betti number calculation.

Result: Betti number: radius r

Calculate $n_1 \leftarrow \#\{\text{Births} < 2r\}$, and $n_2 \leftarrow \#\{\text{Deaths} < 2r\}$

return $n - n' + 1$.

A barcode provides a way to visualize the topological features of a given data cloud for different r values. Because we can calculate the Vietoris-Rips simplicial complex of any dataset for different radius r , it is not clear which one should be used or which one will help us extract its *true* topological structure. One option is to calculate these complexes for all possible radius values. Then, we can plot them and see how the inherent topological features *persist* as we change the radius.

Example 2.3. Generate n points on the circumference S^1 . Assume that this sample of points is noisy; that is, there is some of error in the exact location of the points. Given this noisy data cloud, we can to infer from where these points where sampled. In Figure 2.3, we have the output given by the Ripser algorithm. We observe that only one of the bars persists, which suggests that this noisy data cloud comes from a topological space with one connected component.

Proposed algorithm to compute Betti numbers

The Ripser library does not directly give the Betti numbers. Therefore, we have created a sub-routine to extract them from a given data cloud with a fixed radius value $r > 0$.

The Ripser algorithm gives the birth and death values of a dataset at different radius. We can use this information to calculate the Betti numbers for a particular fixed radius value $r > 0$. First, we count the number of births and deaths that we have until a distance $d = 2r$. The difference between these two values is the number of components that have died within the interval $[0, 2r]$ minus one. Algorithm 2 shows the pseudo-code for this sub-routine.

With this new sub-routine, we can calculate the *Betti curves*, which are plots that visually provide a representation of the changes that β_k experience as we change the radius.

Example 2.4. We have generated a set of uniform points within the unit square. In Figure 2.4, we show the the Betti curves of its first two Betti numbers; that is, we have calculated the Betti number of the same dataset using different radius. We observe that the number of connected components converge to one as we increase the radius. For the number of holes, we observe a maximum at about 0.025.

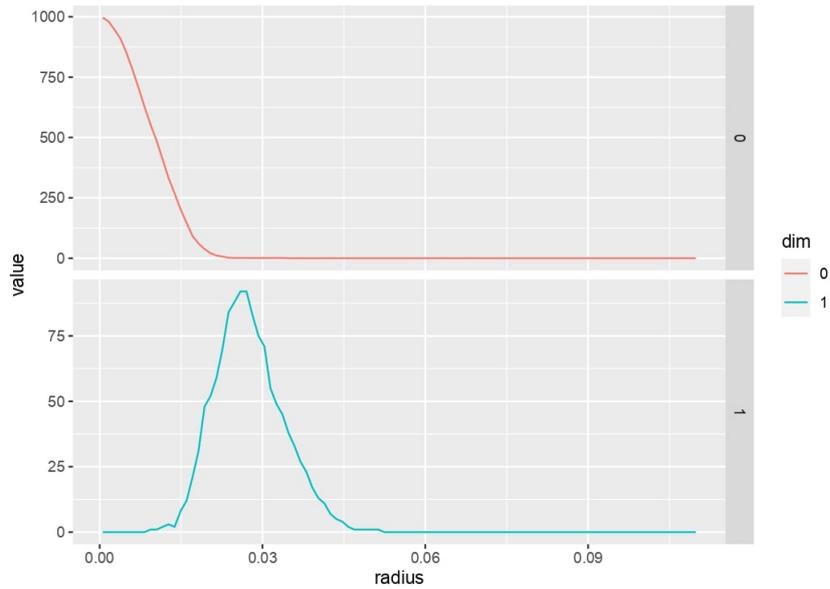


Figure 2.4: Betti curves of uniform points in the unit square.

3 Topology of the level sets

Introduction

In this chapter, we propose a computational method to analyze the level sets of conditional Gaussian processes via simulations. This method combines a variety of tools from statistics, algebraic topology, and computational algorithms to try to understand the statistical distribution of the first two Betti numbers. We start by providing an algorithm that simulates conditional Gaussian processes and we will then analyze an approximation to its level sets. We are not aware of any other work that aims to study the level sets of conditional Gaussian processes. Therefore, a computational study will be useful to obtain intuition about the statistical distribution of the number of connected components and the numbers of holes in the level sets. In Chapter 4, we present a simulation study of a particular conditional Gaussian process that employs this algorithm.

Recently, [Thoppe and Krishnan \(2018\)](#) studied the topology of the level sets of non-conditional Gaussian fields using the novel approach of estimating their topology via Čech complexes. Consequently, we employ a similar strategy to approximate the topology of the level sets of a conditional Gaussian field through the use of Vietoris-Rips complexes. We have decided to use these complexes instead of the Čech complexes for computational efficiency given that we aim to calculate them on a computer.

3.1 Study of the supra-level sets

In this section, we present the methodology that we used to simulate a conditional Gaussian process and to approximate its level sets topology. We must keep in mind that this procedure must be computationally feasible. An advantage of our method is that it can be implemented on a computer cluster, which enables us to use parallel computing.

The understanding of the level sets of a conditional Gaussian process is relevant because they can be used to measure or quantify the uncertainty of the predictions given by the model, which is important if the aim is to predict based on previous observations.

3.1.1 Proposed algorithm

To study the behavior of the (supra-)level sets of a conditional Gaussian process, we will use the following algorithm, referring to Figure 3.1.

1. Sample n points $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ uniformly on the square $[-k, k] \times [-k, k]$, for some $k > 0$ natural number.
2. Sample n' points uniformly that belongs to a known boundary \mathbf{B} .

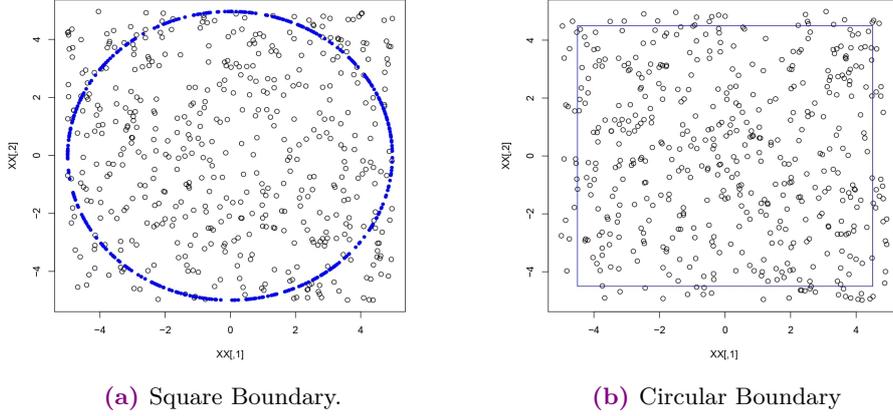


Figure 3.1: Two different boundary conditions

3. Calculate the covariance matrix \mathbf{K} using the points in \mathbf{X} and the boundary \mathbf{B} with a kernel function K .
4. Let \mathbf{Y} be the response vector for the observations in the vector \mathbf{X} conditioned over $\mathbf{B}, f(\mathbf{B})$. We know, for our normality prior assumption, that $\{\mathbf{Y} | \mathbf{B}, f(\mathbf{B}), \mathbf{X}\} \sim \mathcal{N}(\mu', \Sigma')$. Where μ', Σ' are the conditional vector mean and covariance matrix, respectively, given by equations (1.6).
5. Generate a posterior Gaussian process with parameters μ', Σ' . Note that this new process will be conditioned on the values in the boundary $\mathbf{B}, f(\mathbf{B})$ (blue points in Figure 3.1), and \mathbf{X} (black points in Figure 3.1).
6. Study the level sets of this newly generated conditional Gaussian process. For this new process, fix a threshold value $L \in \mathbb{R}$ and approximate the level set topology of the points whose response value is greater or equal than L . That is, we want to study the topology of the set

$$D_L := \{x \in [-k, k]^2 : f(x) \geq L\},$$

where $f(x)$ is the density of our conditional Gaussian process.

3.1.2 Construction of a conditional Gaussian process

To simulate an **unconditional** Gaussian process, we employ the (fixed continuous) *radial basis function kernel* to construct the covariance matrix:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2h^2}\right), \quad (3.1)$$

with bandwidth value $h^2 > 0$, which is a fixed positive real number.

Remember that we sampled n points from the square, see Figure 3.1(a); that is, $\mathbf{X} = \{x_1, x_2, \dots, x_n\} \in [-k, k] \times [-k, k]$, with x_i a 2-dimensional vector, and k a natural number.

To construct the covariance matrix, we use the kernel function (3.1) as follows

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1}, \quad \mathbf{K} = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_1, x_n) & \cdots & K(x_n, x_n) \end{bmatrix}.$$

Here, we have fixed value of $h = \text{median}(\text{dist}(\mathbf{X}))$, the median of the distances between the uniform points in the square. We take this value for the bandwidth heuristically.

Now, given that this kernel is continuous, the conditional Gaussian process is also continuous. Therefore, we fix the response value of the boundary points to be zero; that is, $f(\mathbf{B}) = 0$. For continuity, the process $\{\mathbf{Y} | \mathbf{B}, f(\mathbf{B}), \mathbf{X}\}$ is going to be zero *near* the boundary.

In this work, we take as boundary \mathbf{B} the circle centered at zero with radius $R - \varepsilon$, for some given $0 < \varepsilon < R$. Another possible boundary is the square with end points A, B, C , and D , which is defined as (see Figure 1.3(b))

$$A = (-R + \varepsilon, R - \varepsilon), B = (R, R), C = (R - \varepsilon, -R + \varepsilon), D = (-R, R).$$

We work with a circular boundary because of the mathematical facility and the better behavior of the simulations¹. Then, sample n' points uniformly distributed on the boundary $\mathbf{B} := \{x \in [-k, k]^2 : x^2 = (R - \varepsilon)^2\}$ (blue points in Figure 3.1). The boundary points are going to be our *known* points, whose response vector is $f(\mathbf{B}) = \mathbf{0}_{n \times 1}$. Note that fixing the response vector to be zero is not a wild assumption. We simply require that $f(\mathbf{B})$ to be known. Then, we can do a re-scaling of the dataset by subtracting this vector. Geometrically, this is a translation of the data points on the ambient space.

We want to conditioned a Gaussian process over the given boundary points \mathbf{B} and their corresponding response variable $f(\mathbf{B})$. In other words, we want the distribution of $\{\mathbf{Y} | \mathbf{B}, f(\mathbf{B}), \mathbf{X}\}$, where \mathbf{Y} is the response vector of \mathbf{X} ; see Figure 3.3.

As we saw in Chapter 1, we know that $\{\mathbf{Y} | \mathbf{B}, f(\mathbf{B}), \mathbf{X}\} \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$, and then we can sample from this new conditioned process. We wish to study the level sets of this new process.

3.1.3 Level sets approximation

By making use of the radial basis function kernel (3.1), and the procedure described earlier, we can generate a conditional Gaussian process. To study the behavior of its level sets, we fix different threshold values L , and the same radius $r = 0.4$ for the Vietoris-Rips complexes. We calculate the Betti numbers of the points whose response variable surpasses the threshold value L . Figure 3.2 shows an example of what this looks like.

Given a threshold value $L \in \mathbb{R}$ (fixed), we follow the next procedure to simulate the topology of the level sets of the conditional Gaussian process.

¹We have tried both boundary conditions.

Points above threshold 0 is: 239

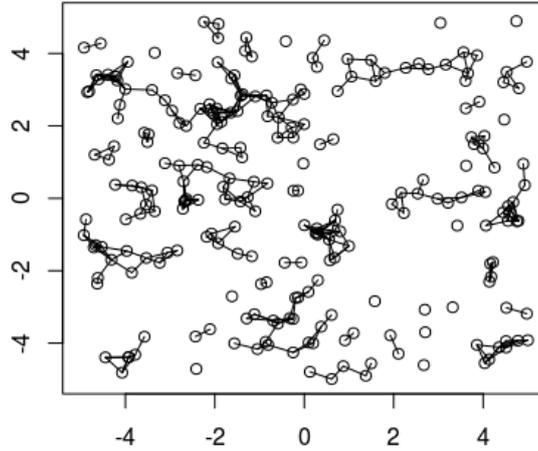


Figure 3.2: VR complex approximation of a level set.

1. Among all the points in \mathbf{X} , select only those points $\mathbf{X}' \subseteq \mathbf{X}$ whose response or height value are greater or equal than the given threshold value L .
2. Construct a Vietoris-Rips (VR) complex using the points in \mathbf{X}' . In other words, given a fixed value $r \geq 0$, the VR complex is given by

$$\text{VR}(\mathbf{X}', r) = \{\sigma \subset \mathbf{X}' : \text{diam}\{\sigma\} \leq 2r\},$$

where

$$\text{diam}(\sigma) = \sup_{x_i, x_j \in \sigma} \{d(x_i, x_j)\},$$

for some arbitrary distance $d: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}_+$. In our case, we will use the standard Euclidean distance $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$, for 2-dimensional real vectors $x = (x_1, x_2)$, $y = (y_1, y_2)$, and $x_i, y_i \in \mathbb{R}$.

3. Find the first two Betti numbers of \mathbf{X}' using the Ripser algorithm.

To calculate the Betti numbers of a simplicial complex, we have a variety of computational algorithms that were developed in the area of topological data analysis via matrix reductions. Note that our procedure turns the problem of studying the level sets of Gaussian process conditioned over a fixed boundary to a linear algebra problem. This new problem is one that is computationally realizable on a computer, which was our goal.

3.2 Known tools

In this section, we present some of the recently discovered tools that we use in this simulation study. These tools are robust estimators of the Betti numbers of a given data cloud embedded in a finite-dimensional space.

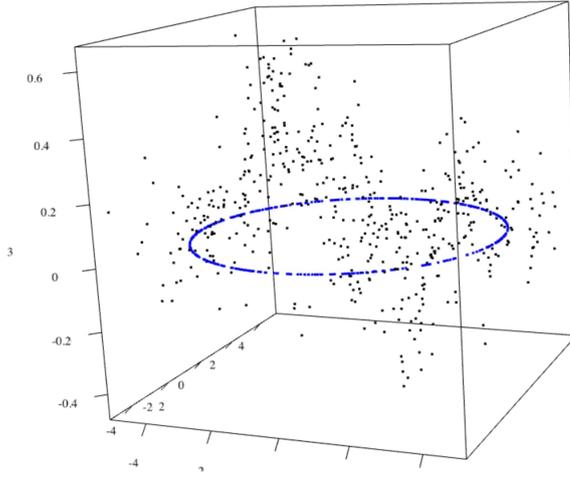


Figure 3.3: Discretization example of a conditional GP.

3.2.1 Naive estimator

Given a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$, the objects that we analyze here are the level sets of f ,

$$D_L := \{x \in \mathbb{R}^d: f(x) \geq L\}. \quad (3.2)$$

A common way to estimate the homology of an unknown space \mathcal{S} from a random sample $\mathcal{X} \subseteq \mathcal{S}$ is to compute the homology of a union of closed balls around the sample points

$$U(\mathcal{X}, r) := \bigcup_{x \in \mathcal{X}} B_r(x),$$

for some fixed $r \in \mathbb{R}_+$. Where $B_r(x)$ denotes the closed d -dimensional ball with radius r and centered at the point x . This estimator is known as the *naive estimator*. A similar idea can be employed to estimate D_L . This is summarized in the following procedure:

1. Use the entire data to construct an estimator \hat{f} .
2. Using this estimator \hat{f} , define:

$$\mathcal{X}^L := \{X_i: \hat{f}(X_i) \geq L\},$$

We can use kernel estimators to get a \hat{f} in both the regression and density estimation case.

3. Consider the set $U(\mathcal{X}^L, r)$ as an estimate of D_L . So we could use $H_*(U(\mathcal{X}^L, r))$ as an estimator for $H_*(D_L)$.

The drawback of this estimator is that it is very sensitive to small changes in \mathcal{X} . Note that because homology is a discrete descriptor, even tiny errors in the filtering step can introduce large errors in the homology estimates. For example, even a single point incorrectly included in the level set assignment can form an extra connected component, and increase the zeroth Betti number by one.

3.2.2 Robust homology estimator

In [Bobrowski et al. \(2017\)](#), the authors provided a consistent method for recovering the homology of the level sets D_L of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$, where f is either a probability density function or a regression function. They suggest a robust homology estimator for the level sets of both density and regression functions that overcome the drawbacks of the naive estimator. Showing that instead of using D_L as an estimate, one should consider the inclusion map between the nested pairs $D_{L+\varepsilon} \subseteq D_{L-\varepsilon}$ (for a properly chosen $\varepsilon > 0$). The key object of interest is then the following induced map between the homology groups of the two level sets:

$$\iota_* : H_*(D_{L+\varepsilon}) \rightarrow H_*(D_{L-\varepsilon}),$$

where $*$ is a standard notation for an arbitrary degree.

Now, denote the Vietoris-Rips complex constructed from the filtered sample as the set $R_L(n, r) := VR(X_n^L, r)$ and define the following inclusion map for any $\varepsilon \in (0, L/2)$

$$\iota : R_{L+\varepsilon}(n, r) \hookrightarrow R_{L-\varepsilon}(n, r).$$

This inclusion induces a map in homology

$$\iota_* : H_*(R_{L+\varepsilon}(n, r)) \rightarrow H_*(R_{L-\varepsilon}(n, r)),$$

and we denote by

$$\hat{H}_*^R(L, \varepsilon; n) := \text{im}(\iota_*).$$

Note that the Nerve lemma [Munkres \(2000\)](#) applies only to the Čech complex and not the Vietoris-Rips complex. Nevertheless, the following theorem states that we can also compute the homology of D_L using the Vietoris-Rips complex. Providing a consistent estimator for $H_*(D_L)$ that uses the Vietoris-Rips complex is important because of its computational efficiency.

The following theorems are very important because they are the theoretical sustain for this simulation study.

Theorem 3.1. (4.3 in [Bobrowski et al. \(2017\)](#)) Let $L > 0$ and $\varepsilon \in (0, L/2)$ be such that the function $f(x)$ has no critical values in the range $[L-2\varepsilon, L+2\varepsilon]$. If $r \rightarrow 0$ and $nr^d \rightarrow \infty$, then for n large enough we have

$$\mathbb{P}(\hat{H}_*^R(L, \varepsilon; n) \cong H_*(D_L)) \geq 1 - 6ne^{-C_{\varepsilon/2}^* nr^d}.$$

In particular, if $nr^d \geq D \log(n)$, with $D > (C_{\varepsilon/2}^*)^{-1}$, then

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{H}_*^R(L, \varepsilon; n) \cong H_*(D_L)) = 1.$$

Remark: On our simulations we approximate the topology of the set D_L directly via VR complexes, NOT using $\text{im}(\iota_*)$. As $\text{im}(\iota_*) \subseteq D_{L-\varepsilon}$, we can take $\varepsilon = 10^{-100}$ and work with the set D_L itself as it is.

On the articles [Bobrowski et al. \(2017\)](#), [Bobrowski and Mukherjee \(2013\)](#) the authors prove the theorems here presented using Čech complexes. We want to test if we can use the fact that $VR(X, r) \subseteq C(X, \sqrt{2}r) \subseteq VR(X, \sqrt{2}r)$ to approximate those results employing Vietoris-Rips complexes instead, which are easier to calculate on a computer.

For the rest of this chapter, we assume that $\mathcal{M} \subseteq \mathbb{R}^d$ is a closed, smooth, compact and without boundary m -dimensional manifold, with $m < d$.

Theorem 3.2. (4.1 in [Bobrowski and Mukherjee \(2013\)](#)) If $nr_n^m \rightarrow 0$, then for all values $1 \leq k \leq m - 1$,

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}(\beta_{k,n})}{n^{k+2}r_n^{m(k+1)}} = \frac{\text{Var}(\beta_{k,n})}{n^{k+2}r_n^{m(k+1)}} = \mu_k^b,$$

and

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}(\beta_{0,n})}{n} = 1.$$

Where

$$\mu_k^b = \frac{1}{(k+2)!} \int_{\mathcal{M}} f^{k+2}(x) dx \int_{(\mathbb{R}^m)^{k+1}} h_1^b(0, \mathbf{y}) d\mathbf{y}.$$

The function h_ε^b is an indicator function on subsets \mathcal{Y} of size $k+2$, testing that a subset forms a non-trivial k -cycle; that is,

$$h_\varepsilon^b(\mathcal{Y}) := 1\{\beta_k(U(\mathcal{Y}, \varepsilon)) = 1\}. \quad (3.3)$$

Theorem 3.3. (4.2 in [Bobrowski and Mukherjee \(2013\)](#)) Let $nr_n^m \rightarrow 0$, and $1 \leq k \leq m$,

1. If $\lim_{n \rightarrow \infty} n^{k+1}r_n^k = 0$, then

$$\beta_{k,n} \xrightarrow{L^2} 0.$$

If, in addition, $\sum_{n=1}^{\infty} n^{k+1}r_n^{mk} < \infty$, then

$$\beta_{k,n} \xrightarrow{a.s.} 0.$$

2. If $\lim_{n \rightarrow \infty} n^{k+2}r_n^{m(k+1)} = \alpha \in (0, \infty)$, then

$$\beta_{k,n} \xrightarrow{\mathcal{L}} \text{Poisson}(\alpha\mu_k^b).$$

3. If $\lim_{n \rightarrow \infty} n^{k+2}r_n^{m(k+1)} = \infty$, then

$$\frac{\beta_{k,n} - \mathbb{E}(\beta_{k,n})}{(n^{k+1}r_n^{mk})^{1/2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mu_k^b).$$

4 Empirical study of level sets

Introduction

In this chapter, we report the results of our analysis of the statistical distribution of the first two Betti numbers β_0, β_1 of the level sets of a particular conditional Gaussian process, via an empirical study. The method that we used for this analysis is the one that was described in Chapter 3. Using our method, we have observed that the distribution of β_0 and β_1 is similar to a Poisson distribution, whose parameters depend on the threshold value considered. We tested this asseveration using the graphical goodness of fit test that was presented in Chapter 1. We also present some results about an asymptotic exploration of β_0 , and β_1 . We also employ the outcomes obtained by our method to approximate to the theoretical results presented in [Bobrowski et al. \(2017\)](#), whose conclusions were obtained for unconditional regression and density functions.

4.1 Connected components

4.1.1 Method description

Definition (Discretization of the plane). We define $\mathcal{D}_{n,k}$ as the collection of n uniform random points in the square $[-k, k] \times [-k, k]$, for $n \in \mathbb{N}$, and k some natural number. We take k to be natural to have a square of odd length, as the total amount of integer points from k to $-k$ is $2k + 1$.

To analyze the distribution of β_0, β_1 of the conditional Gaussian process of interest, we follow the next procedure:

1. Fix a threshold value $L \in \mathbb{R}$. Then, proceed to calculate β_0 , the number of connected components, of the level set

$$D_L := \{x \in \mathcal{D}_{n,k} : f(x) \geq L\},$$

where $f(\cdot)$ is the density of our conditional multivariate normal distribution.

2. Generate m simulations of D_L and calculate for each of them the number of connected components β_0 ; call S_k^L the k -th simulation generated by our method with threshold value L . We write S_k when there is no risk of confusion.
3. For each threshold value L , let $X_k^L = \beta_0(S_k^L)$; that is, X_k^L is the number of connected components in the k -th simulation, among the m simulations of D_L , for every $1 \leq k \leq m$.

4. If the dataset $\mathcal{B}_0 := \{X_1^L, X_2^L, \dots, X_m^L\}$ comes from a Poisson distribution distribution with parameter μ , then we can approximate it using the maximum likelihood estimator (MLE) of the sample: $\hat{\mu} = \frac{1}{m} \sum_{i=1}^m X_i^L$.

One particular characteristic of a Poisson distributed random variable is that its theoretical mean and variance coincide. The first test to check if a dataset comes from a Poisson distribution is to examine if its sample mean and variance are numerically similar. Tables 4.1, 4.2, and 4.3 contain a collection of sample mean and variance for different threshold values. For every threshold value L , we generate $m \in \mathbb{N}$ independent simulations of a conditional Gaussian process.

After observing the geometrical structure of our simulations, we have decided to explore three different cases individually, which we call: Above, Below, and Double. Whose definitions follow, for some fixed real value $L \geq 0$:

1. Above: The threshold value is a positive number,

$$\{x \in \mathcal{D}_{n,k} : f(x) \geq L\}.$$

2. Below: The threshold value is a negative number,

$$\{x \in \mathcal{D}_{n,k} : f(x) \leq -L\}.$$

3. Double: We take, at the same time, both positive and negative threshold values. This means we are taking the union of the Above and Below sets; that is,

$$\{x \in \mathcal{D}_{n,k} : f(x) \geq L\} \cup \{x \in \mathcal{D}_{n,k} : f(x) \leq -L\}.$$

4.1.2 Simulation results

Tables

In the following tables, we have included the variable: *Cnts*, which represents the number of simulations, among the m that we have done in total, whose height surpasses the fixed threshold value L . For those simulations that did not surpass the threshold value L , we agree that $\beta_0 = \beta_1 = 0$ because we did not see any point above the given threshold value.

Individually for each threshold value L —say for example $L = 0.6$ —, we generate m simulations of the set D_L . For each of these m simulations, we calculate the number of connected components on the set D_L , obtaining the sample $\mathcal{B}_0 = \{X_1^L, X_2^L, \dots, X_m^L\}$, with X_i^L the random variable that counts the number of connected components of the i -th simulation of the set D_L , for $1 \leq i \leq m$. We denote by $\hat{\mu}$, and $\hat{\sigma}^2$ the empirical mean and variance, respectively, of the sample set \mathcal{B}_0 . For all of the simulations in this chapter, we take the values of $m = 1000$, and $k = 5$.

β_0 : Above. $n = 500$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.24	0.35	183
0.62	0.21	0.27	169
0.63	0.16	0.23	124
0.64	0.15	0.18	128
0.66	0.11	0.15	91
0.68	0.08	0.09	72
0.70	0.06	0.08	55
0.72	0.06	0.07	51

Table 4.1: Positive threshold value.

β_0 : Below. $n = 500$			
$-L$	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.245	0.319	198
0.62	0.223	0.324	172
0.63	0.173	0.267	133
0.64	0.188	0.319	139
0.66	0.115	0.169	92
0.68	0.087	0.115	73
0.70	0.062	0.082	53
0.72	0.052	0.063	46

Table 4.2: Negative threshold value.

β_0 : Double. $n = 500$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.555	0.848	376
0.62	0.409	0.568	297
0.63	0.381	0.590	264
0.64	0.319	0.498	236
0.66	0.266	0.414	193
0.68	0.186	0.277	150
0.70	0.151	0.188	127
0.72	0.104	0.135	88

Table 4.3: Double threshold value.

β_0 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.341	0.457	257
0.62	0.259	0.398	188
0.63	0.238	0.334	185
0.64	0.201	0.251	167
0.66	0.170	0.203	142
0.68	0.134	0.184	110
0.70	0.105	0.152	84
0.72	0.102	0.139	84

Table 4.4: Positive threshold value.

β_0 : Below. $n = 1,000$			
$-L$	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.299	0.353	242
0.62	0.248	0.262	214
0.63	0.210	0.264	171
0.64	0.206	0.283	164
0.66	0.162	0.199	135
0.68	0.131	0.166	111
0.70	0.140	0.208	110
0.72	0.068	0.079	61

Table 4.5: Negative threshold value.

β_0 : Double. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.637	0.876	427
0.62	0.538	0.749	380
0.63	0.458	0.618	331
0.64	0.430	0.583	314
0.66	0.343	0.457	255
0.68	0.242	0.323	190
0.70	0.164	0.195	139
0.72	0.158	0.221	126

Table 4.6: Double threshold value.

More data points

We want to consistently simulate n data points inside the square $[-5, 5]^2$, and n' of them on the boundary \mathbf{B} . In addition, we want to maintain the same ratio of points in the square and the boundary as the values of n , and n' are increased. To do this, we follow the next rule for the number of points:

$n :=$ is the number of points inside the square $[-5, 5]^2$;

$n' := 0.8 \times n$ be the number of points sampled from the boundary \mathbf{B} ;

$r := 0.4$ is the constant radius value for all the Vietoris-Rips simplicial complexes.

β_0 : Above. $n = 1,500$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.346	0.444	262
0.62	0.313	0.389	241
0.63	0.238	0.297	191
0.64	0.230	0.281	189
0.66	0.180	0.243	141
0.68	0.148	0.182	123
0.70	0.116	0.132	103
0.72	0.070	0.080	62

Table 4.7: Positive threshold value.

β_0 : Above. $n = 2,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.410	0.558	306
0.62	0.293	0.361	229
0.63	0.281	0.374	212
0.64	0.246	0.283	203
0.66	0.180	0.203	153
0.68	0.141	0.185	117
0.70	0.139	0.171	118
0.72	0.090	0.108	81

Table 4.8: Positive threshold value.

For the values of $n = 500$, $n' = 400$, and $r = 0.4$, we show in Tables 4.1, 4.2, and 4.3 the sample mean and variance for different threshold values.

Similarly, we have in Tables 4.4, 4.5, and 4.6 the respective sample mean and variance for $n = 1000$, $n' = 800$, and $r = 0.4$, for different threshold values.

In Tables 4.7, and 4.8 we show sample mean and variance for $n = 1500$, $n' = 1200$, and $n = 2000$, $n' = 1600$ points, respectively.

Remark: For all of the n uniformly sampled points inside the square $[-5, 5]^2$, we have generated $m = 1,000$ simulations for every set D_L . The variable *Cnts* counts the number of simulations, among the $m = 1,000$ done in total, with at least one point whose height surpasses the threshold value L .

Empirical graphical tool

In Section 1.4, we presented a test for a Poisson distribution that does not require independence among the observations, only equal distribution. In Figures 4.1, 4.2, and 4.3 we show a set of such plots. We can see that they all suggest that we are in the presence of a straight line, which supports our believe that the counting sets of connected components may come from a Poisson distribution. Figures 4.1, 4.2, and 4.3 were created by sampling a number of $n = 1000$ points inside the square and $n' = 800$ on a circular boundary.

4.1.3 Conclusions

Observation 1: For the case $n = 500$, we see that sample mean and variance are fairly similar. A Poisson behavior is plausible. Note that the parameter (remember that for a Poisson distribution an estimator for this value in the sample mean) is different for each threshold value, which suggests that the process is not homogeneous.

Observation 2: By increasing the number of points proportionally for $n = 1000$, something similar happens as in the last case. Note that the variance is slightly higher in some cases, see the remark below for a possible interpretation of this. We can proceed and test if it is plausible that these points come from a Poisson distribution.

Observation 3: For $n = 1500, 2000$, we again see that mean and variance are fairly

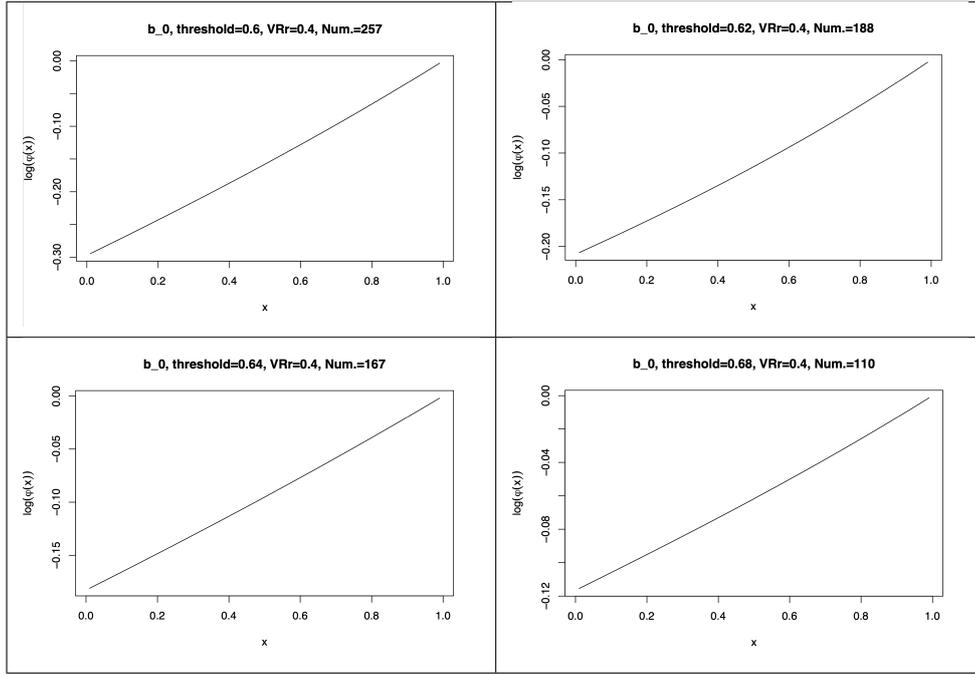


Figure 4.1: β_0 for the case: Above.

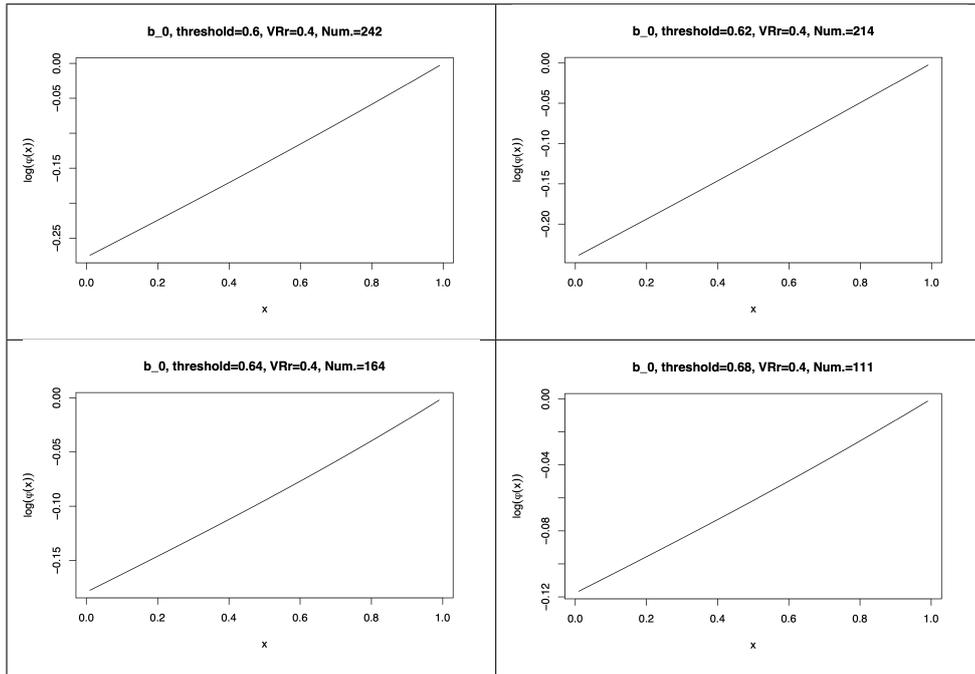


Figure 4.2: β_0 for the case: Below.

similar in both tables. Note that for the cases where there is a difference, the variance is always the one that is slightly greater than the mean. This could be caused by the Binomial approximation to the Poisson distribution (see the following remark).

Remark: We can approximate a random variable with Poisson distribution by taking the limit on the size number of a Binomial distributed random variable. For the last one, the theoretical variance is greater than the mean. Note also that for the Binomial distribution,

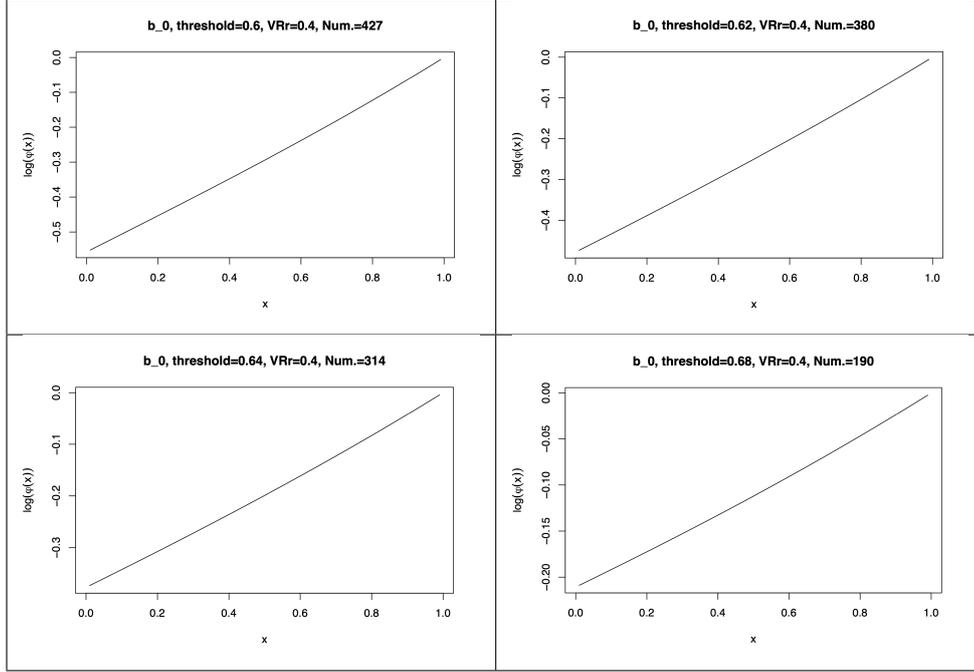


Figure 4.3: β_0 for the case: Double.

the graphs of the empirical probability generating function are concave upwards.

4.2 One-dimensional holes

4.2.1 Method description

Similarly, using the 4-steps method that we introduced for the number of connected components, defined as $Y_k^L = \beta_1(S_k^L)$, where S_k^L is the k -th simulations of the set D_L for the conditional Gaussian process of interest. In other words, Y_k^L is the random variable that counts the number of one-dimensional holes of the k -th simulation of the set D_L , for a fixed given threshold value L , with $1 \leq k \leq m$.

Let \mathcal{B}_1 be the collection of the number of holes in every simulation; that is, $\mathcal{B}_1 := \{Y_1^L, Y_2^L, \dots, Y_m^L\}$. If this set comes from a Poisson distribution with parameter μ , then we know that the MLE is $\hat{\mu} = \frac{1}{m} \sum_{i=1}^m Y_i^L$. As shown in the last section for the number of connected component, in Tables 4.9, 4.10, and 4.11 we show a collection of sample variances and means for different threshold values. For every threshold value L , we generate m (independent) simulations of a conditional Gaussian process.

Tables 4.9, 4.10, and 4.11 show sample mean and variance for our β_1 's for the case when the number of points inside the square is $n = 500$. We see that $\hat{\mu}$ and $\hat{\sigma}^2$ are numerically similar, as we saw previously for β_0 . Note that a zero value for both mean and variance implies that we have not seen holes in all the simulations above the given (and fixed) threshold value L .

Figures 4.4, 4.5, and 4.6 show the empirical generating function for some of our threshold values when $n = 1000$. Note that most of the plots suggest that we are in the presence of

a straight line. This means that a Poisson distribution behavior is reasonable for our data. A horizontal line means that the value for mean and variance are both zero, which tells us that we have not seen any holes in any of the simulations above that given threshold value. For consistency, we will present the empirical probability generating functions (e.p.g.f) with the same threshold values for both, β_0 and β_1 , including those non-informative ones with horizontal lines.

4.2.2 Simulation results

Tables

In this section, we present the corresponding tables for β_1 and the number of holes.

β_1 : Above. $n = 500$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.005	0.011	183
0.62	0.001	0.001	169
0.63	0.003	0.003	124
0.64	0.002	0.002	128
0.66	0.003	0.003	91
0.68	0.000	0.000	72
0.70	0.000	0.000	55
0.72	0.000	0.000	51

Table 4.9: Positive threshold value

β_1 : Below. $n = 500$			
$-L$	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.005	0.007	198
0.62	0.001	0.001	172
0.63	0.002	0.002	133
0.64	0.004	0.006	139
0.66	0.004	0.010	92
0.68	0.001	0.001	73
0.70	0.000	0.000	53
0.72	0.000	0.000	46

Table 4.10: Negative threshold value

β_1 : Double. $n = 500$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.008	0.008	376
0.62	0.011	0.024	297
0.63	0.002	0.002	264
0.64	0.001	0.001	236
0.66	0.003	0.005	193
0.68	0.000	0.000	150
0.70	0.000	0.000	127
0.72	0.000	0.000	88

Table 4.11: Double threshold value

β_1 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.004	0.009	257
0.62	0.003	0.004	188
0.63	0.003	0.004	185
0.64	0.003	0.003	167
0.66	0.002	0.002	142
0.68	0.001	0.001	110
0.70	0.000	0.000	84
0.72	0.001	0.001	84

Table 4.12: Positive threshold value.

β_1 : Below. $n = 1,000$			
$-L$	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.003	0.003	242
0.62	0.003	0.003	214
0.63	0.003	0.003	171
0.64	0.002	0.002	164
0.66	0.000	0.000	135
0.68	0.001	0.001	111
0.70	0.000	0.000	110
0.72	0.000	0.000	61

Table 4.13: Negative threshold value.

β_1 : Double. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.012	0.019	427
0.62	0.007	0.008	380
0.63	0.001	0.001	331
0.64	0.007	0.006	314
0.66	0.001	0.001	255
0.68	0.003	0.004	190
0.70	0.000	0.000	139
0.72	0.002	0.002	126

Table 4.14: Double threshold value.

β_1 : Above. $n = 1,500$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.002	0.002	262
0.62	0.002	0.002	241
0.63	0.000	0.000	191
0.64	0.000	0.000	189
0.66	0.001	0.001	141
0.68	0.000	0.000	123
0.70	0.001	0.001	103
0.72	0.002	0.004	62

Table 4.15: Positive threshold value.

β_1 : Above. $n = 2,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
0.60	0.002	0.002	306
0.62	0.004	0.005	229
0.63	0.000	0.000	212
0.64	0.002	0.004	203
0.66	0.002	0.002	153
0.68	0.000	0.000	117
0.70	0.000	0.000	118
0.72	0.000	0.000	81

Table 4.16: Positive threshold value.

In Tables 4.9, 4.10, and 4.11 we show the sample mean and variance for the number of holes. Observe that mean and variance are both zero for some of the threshold values, which implies that we have not observed any one-dimensional holes for those respective threshold values.

Tables 4.12, 4.13, and 4.14 contain the information for the case where $n = 1,000$.

Tables 4.15, and 4.16 show, for the Above case, the sample mean and variance for the case where $n = 1,500$, and $n = 2,000$, respectively.

Remark: For all of the n uniformly sampled points inside the square $[-5, 5]^2$, we have generated $m = 1,000$ simulations for every set D_L . The variable Cnts counts the number of simulations, among the $m = 1,000$ done in total, with at least one point whose height surpasses the threshold value L .

Empirical graphical tool

The following graphs, Figures 4.4, 4.5, and 4.6 were created by sampling a number of $n = 1000$ uniform points inside the $[-5, 5]^2$ square.

4.2.3 Conclusions

Observation 4: For β_1 , the statistical distribution is similar as for β_0 when we increase the number of points. Therefore, a Poisson behaviour may be plausible.

Observation 5: Sample and mean variance are very similar in both cases, $n = 1500, 2000$. We can proceed and test if these points come form a Poisson distribution.

Observation 6: In Figure 4.4 we see that the first two are not straight lines at all but they get *more* straight as we increase the threshold. Figure 4.5 suggest that they are all straight lines. Figure 4.6 suggest that the majority are also straight lines. This implies that a Poisson distribution may explain these datasets.

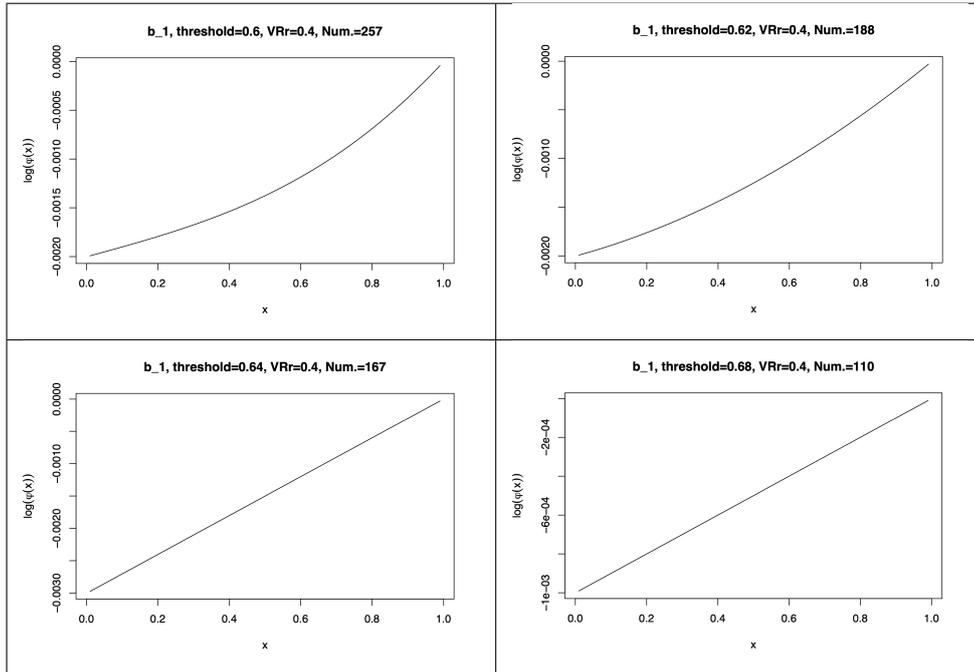


Figure 4.4: β_1 for the case: Above.

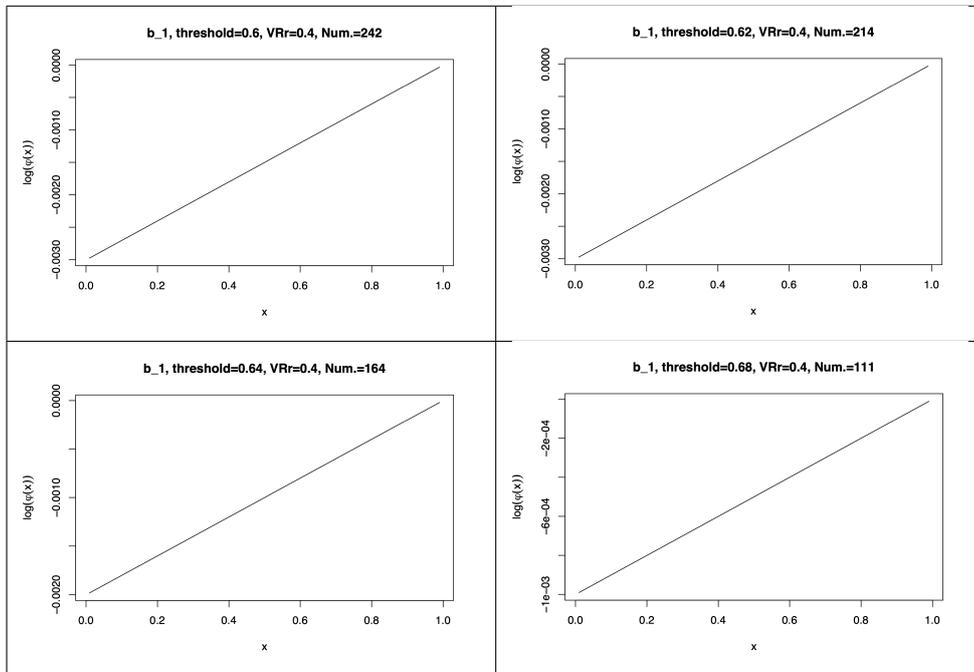


Figure 4.5: β_1 for the case: Below.

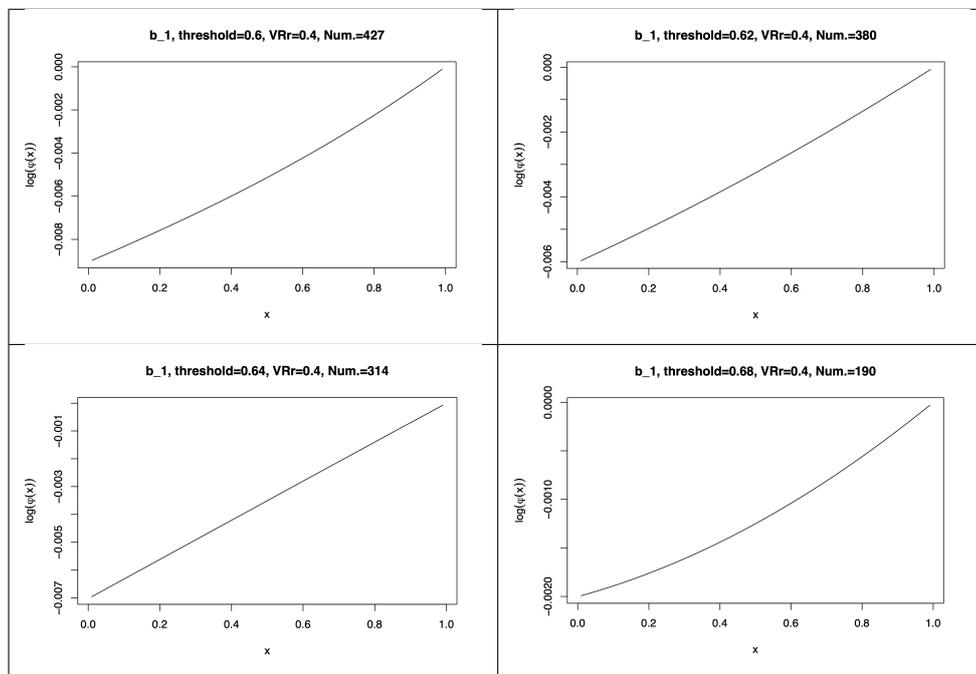


Figure 4.6: β_1 for the case: Double.

Fixed bandwidth

In this section, we have fixed the bandwidth to be $h^2 = 0.5$ for all of the simulations. Here we only take into consideration the case *Above* because the constant bandwidth value makes the simulations more symmetric with respect to the plane $z = 0$, thus there is no longer a need to consider different cases. The threshold has been modified, the new tables that are obtained follow.

β_0 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
1.38	11.326	8.938	1000
1.40	10.944	8.587	1000
1.42	10.765	8.035	1000
1.44	10.314	9.480	1000
1.46	10.104	8.567	1000
1.48	9.820	7.857	1000
1.50	9.522	8.584	1000
1.52	9.372	8.343	1000

Table 4.17: Positive threshold value.

β_0 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
2.00	4.072	4.451	981
2.02	3.700	4.402	970
2.04	3.547	3.907	965
2.06	3.545	4.376	958
2.08	3.231	4.293	953
2.10	3.185	3.912	955
2.12	2.980	3.629	943
2.14	2.994	3.675	947

Table 4.18: Positive threshold value.

β_0 : Positive. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
2.20	2.495	3.111	912
2.22	2.381	3.146	878
2.24	2.262	2.914	884
2.26	2.230	2.920	879
2.28	2.131	2.700	866
2.30	2.029	2.608	843
2.32	1.961	2.495	849
2.34	1.786	2.216	819

Table 4.19: Positive threshold value.

β_0 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
2.50	1.190	1.494	687
2.60	0.894	0.999	599
2.70	0.677	0.807	471
2.80	0.550	0.602	416
2.90	0.356	0.359	307
3.00	0.265	0.271	235
3.10	0.205	0.217	183
3.20	0.158	0.167	145

Table 4.20: Positive threshold value.

Observation 7: For Tables 4.17, 4.21 we still do not have a Poisson behaviour. Consequently, we need to increase the threshold value to start seeing a Poisson distribution behaviour. Currently, the amount of connected components is $X_k \geq 1$ for all $k \in \{1, 2, \dots, 1000\}$.

Observation 8: As the threshold is increased, see Tables 4.19 and 4.20, the sample mean and variance start to look more similar, especially the last one.

Observation 9: For β_1 , we see that Tables 4.22 and 4.23 show the more similar sample mean and variance. In table 4.24, zero mean and variance implies that none of the simulated conditional Gaussian processes surpasses the given threshold value; that is, we did not observe a one-dimensional hole.

Remark: For all of the n uniformly sampled points inside the square $[-5, 5]^2$, we have generated $m = 1,000$ simulations for every set D_L . The variable *Cnts* counts the number

β_1 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
1.38	0.177	0.225	1000
1.40	0.177	0.251	1000
1.42	0.147	0.189	1000
1.44	0.147	0.175	1000
1.46	0.128	0.169	1000
1.48	0.135	0.182	1000
1.50	0.108	0.124	1000
1.52	0.101	0.128	1000

Table 4.21: Positive threshold value.

β_1 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
2.00	0.023	0.024	981
2.02	0.011	0.016	970
2.04	0.012	0.013	965
2.06	0.013	0.013	958
2.08	0.008	0.008	953
2.10	0.008	0.008	955
2.12	0.010	0.010	943
2.14	0.013	0.014	947

Table 4.22: Positive threshold value.

β_1 : Positive. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
2.20	0.003	0.003	912
2.22	0.013	0.014	878
2.24	0.008	0.008	884
2.26	0.004	0.004	879
2.28	0.007	0.007	866
2.30	0.008	0.009	843
2.32	0.004	0.004	849
2.34	0.003	0.003	819

Table 4.23: Positive threshold value.

β_1 : Above. $n = 1,000$			
L	$\hat{\mu}$	$\hat{\sigma}^2$	Cnts
2.50	0.000	0.000	687
2.60	0.002	0.004	599
2.70	0.001	0.001	471
2.80	0.000	0.000	416
2.90	0.000	0.000	307
3.00	0.000	0.000	235
3.10	0.000	0.000	183
3.20	0.000	0.000	145

Table 4.24: Positive threshold value.

of simulations, among the $m = 1,000$ done in total, with at least one point whose height surpasses the threshold value L .

Remark: To avoid any critical point, we have not fixed the threshold value $L \geq 0$. One of the assumptions in Theorem 3.1 is that there are no critical points in the range $[L - 2\varepsilon, L + 2\varepsilon]$.

4.3 Asymptotic exploration of connected components

In this section, we study the asymptotic behavior of Theorem 3.2 using the simulations generated by our method. Theorem 3.2 employs Čech complexes for its results, while our method utilizes Vietoris-Rips complexes. With this asymptotic study, we aim to test if we can approximate the theorem's results with Vietoris-Rips complexes. Our motivation to do this comes from the fact that for any radius value $r > 0$, we can approximate a Čech complex from above and below using the following fact from algebraic topology [Silva and Ghrist \(2007\)](#):

$$VR(X, r) \subseteq C(X, \sqrt{2}r) \subseteq VR(X, \sqrt{2}r).$$

If our method is successful, then we will be able to approximate the value of μ_k^b , for the cases $k = 0, 1$, from Theorem 3.2, using Vietoris-Rips complexes instead of Čech complexes. This is convenient because the calculation of Čech complexes is not computationally tractable in an efficient way.

For all of the graphs that we will present next, on the vertical axis we have $y = \frac{\mathbb{E}(\beta_k)}{n^{k+2}r_n^{z(k+1)}}$, while on the horizontal axis we have the number of uniformly sampled points $n = 500, 1000, 1500$, and 2000 from the square $[-5, 5]^2$. For these simulations, we have fixed the constant radius $r_n = 0.4$, for every n . We take $k = 0, 1$, and z represents the dimension where the points are embedded. Because we are simulating points in the plane, we have that $z = 2$.

Median of the distances as bandwidth

In this subsection, we present the plot for the *Above* case, with bandwidth $h = \text{median}(\text{dist}(\mathbf{X}))$. We only study this case because calculating these graphs for different sample sizes is computationally expensive.

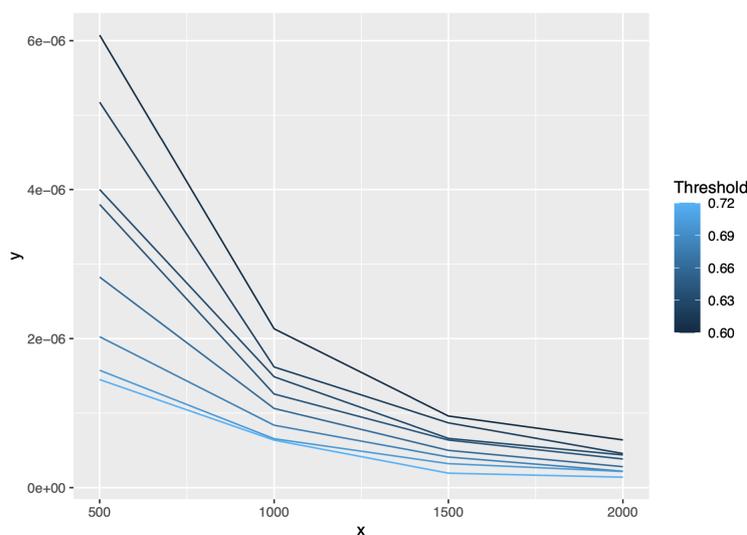


Figure 4.7: β_0 : Asymptotic behaviour of Theorem 3.2, $h = \text{median}(\text{dist}(\mathbf{X}))$.

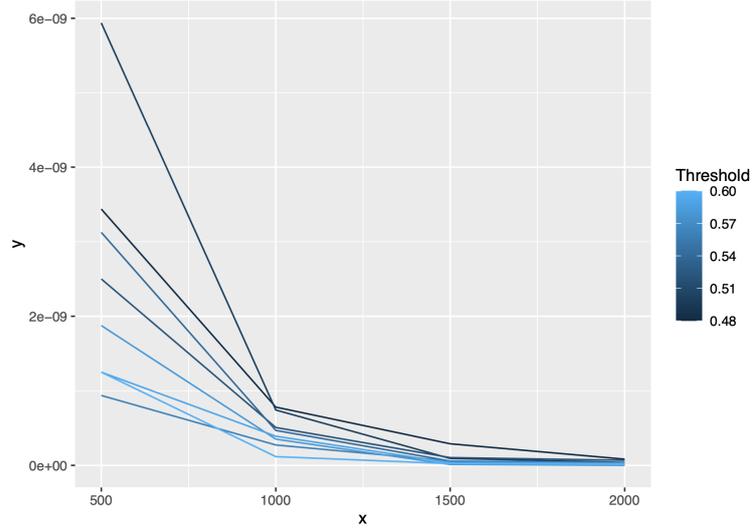


Figure 4.8: β_1 : Asymptotic behaviour of Theorem 3.2, $h = \text{median}(\text{dist}(\mathbf{X}))$.

In Figures 4.7, and 4.8 we have the asymptotic behaviour of $\mathbb{E}(\beta_0)/n^2 r_n^2$, and $\mathbb{E}(\beta_1)/n^3 r_n^4$, respectively, for the sample size $n = 500, 1000, 1500, 2000$. Note that all of the graphs tend to be contained in the same interval. However, it may be possible to use Vietoris-Rips complexes to approximate μ_0^b instead of Čech complexes.

Remark: Note the graphs decay exponentially as the number of points n , and the threshold value L increase (independently from each other); that is, $y \approx e^{-\theta \cdot n}$, where $\theta = \theta(L)$ (we believe) only depends on the threshold value L .

Constant bandwidth

Now, taking a constant value for the bandwidth on the kernel function 3.1, with $h^2 = 0.5$, we see in Figures in Figures 4.9, and 4.10 a similar behaviour as for the non-constant bandwidth value for h that we took before.

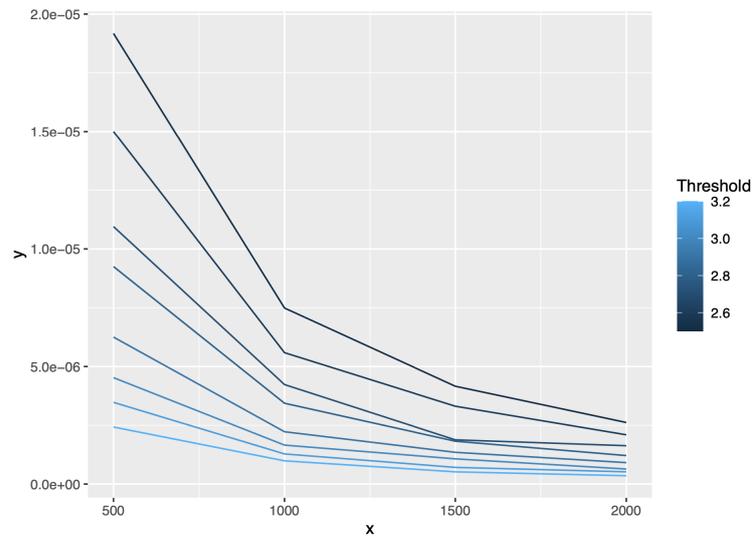


Figure 4.9: β_0 : Asymptotic behaviour of Theorem 3.2 (fixed h).

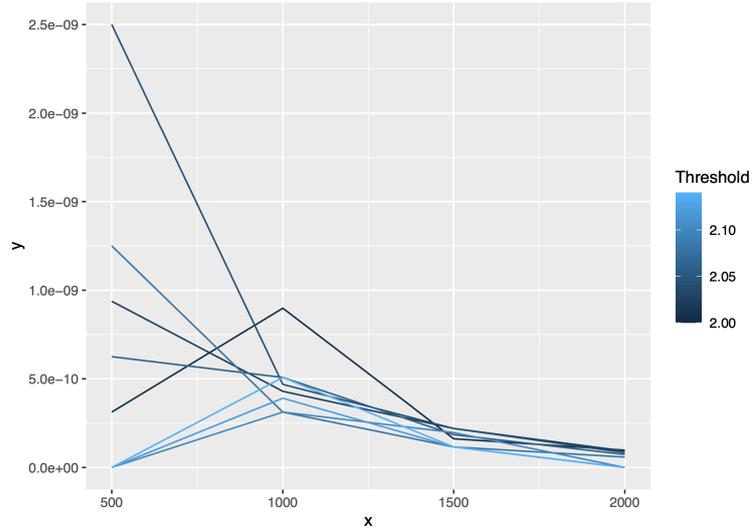


Figure 4.10: β_1 : Asymptotic behaviour of Theorem 3.2 (fixed h).

Figure 4.9 shows the asymptotic behaviour of $\mathbb{E}(\beta_0)/n^3 r_n^4$ for the different sample values. Similarly, in Figure 4.10 we show the same for $\mathbb{E}(\beta_1)/n^3 r_n^4$.

Remark: Note that it seems that the convergence is faster for β_1 than for β_0 in both cases. Take into account that for β_1 we are dividing by $n^3 r_n^4$ and for β_0 we divide by $n^2 r_n^2$. It may be possible to use Vietoris-Rips complexes instead of Čech complexes to approximate the value of μ_k^b from Theorem 4.3 using our simulation method.

5 Conclusions and pending work

We have observed that the first two Betti numbers of the particular conditional Gaussian process that we are considering in this work seem to be distributed as Poisson, whose parameters depend on the threshold value selected.

We have seen that it may be possible to approximate the results in Theorems 3.2 and 3.3 by using Vietoris-Rips complexes instead of the Čech complexes. This opens up the possibility to study the asymptotic behavior of these theorems via an empirical study through simulation. However, proving these results using Vietoris-Rips complexes is still pending.

5.1 Open questions

Some open questions have arisen from this simulation study, as follows:

1. Fix a threshold value $L \in \mathbb{R}$. Given a conditional Gaussian process of interest $\{\mathbf{Y} | \mathbf{B}, f(\mathbf{B}), \mathbf{X}\}$, then take all the points whose response variable value is greater or equal to L . However, what is the behavior of its Betti numbers β_0, β_1 above the threshold? This is the question that is studied in this thesis.
2. Continuous kernel.
 - a. How *far* from the boundary do we have to go to see an unconditional Gaussian process in the interior (or something similar)? Here, by interior we mean the Gaussian process that is in the middle of the square, such as in Figure 3.1. We believe that the conditional process is *close* to zero near the boundary, while it should behave similarly to an unconditional Gaussian process in the interior.
 - b. How *near or far* from the boundary do we have to go to observe this?
3. In [Thoppe and Krishnan \(2018\)](#), the authors take the length of the sides of the square $k = 2n + 1$, for some $n \in \mathbb{N}$. They do this because they want the length of the square to be an even number around the origin. They study the relationship of k and L because they let these values tend to infinity ($L, k \rightarrow \infty$). Can we find or prove something similar in the conditional case?
4. Until now we have used a circle as the conditional boundary curve. However, what would happen if we change this curve to another smooth one? At first we used a square, but the behavior of the random field was more chaotic than if we used a circular boundary.

References

- Adler, R., & Taylor, J. E. (2003). *Random fields and their geometry*. Springer.
- Bauer, U. (2019). *The ripser algorithm*. Retrieved from <https://ulrich-bauer.org/riper-talk.pdf>
- Bobrowski, O., & Kahle, M. (2018). Topology of random geometric complexes: a survey. *Journal of Applied and Computational Topology*, 1, 331-364.
- Bobrowski, O., & Mukherjee, S. (2013). The topology of probability distributions on manifolds. *Probability Theory and Related Fields*, 161, 651-686.
- Bobrowski, O., Mukherjee, S., & Taylor, J. E. (2017, 02). Topological consistency via kernel estimation. *Bernoulli*, 23(1), 288–328. Retrieved from <https://doi.org/10.3150/15-BEJ744> doi: 10.3150/15-BEJ744
- Clapp, M. (2017). *Análisis matemático* (2nd ed.). LIBRUNAM.
- Coeurjolly, J. F., Møller, J., & Waagepetersen, R. (2017). A tutorial on palm distributions for spatial point processes. *International Statistical Review*, 85(3), 404-420. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12205> doi: 10.1111/insr.12205
- Edelsbrunner, H., & Harer, J. (2010). *Computational topology: An introduction*. American Mathematical Society.
- Frazier, P. (2018). A tutorial on bayesian optimization. *arXiv preprint, arXiv:1807.02811v1*. Retrieved from <https://arxiv.org/abs/1807.02811>
- Frazier, P., & Wang, J. (2015). Bayesian optimization for materials design. *arXiv preprint, arXiv:1506.01349v1*. Retrieved from <https://arxiv.org/abs/1506.01349>
- Gonzalvez, J., Lezmi, E., Roncalli, T., & Xu, J. (2019). Financial applications of gaussian processes and bayesian optimization. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3344332
- Gretton, A. (2019). *Introduction to rkhs, and some simple kernel algorithms*. Retrieved from http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf
- Mukherjee, S. (2015). *Probabilistic machine learning*. Duke University. Retrieved from http://www2.stat.duke.edu/~sayan/561/2015/stat_ml.pdf
- Munkres, J. (2000). *Topology (2nd edition)*. Prentice Hall, Inc.
- Nakamura, M., & Pérez-Abreu, V. (1993). Exploratory data analysis for counts using the empirical probability generating function. *Communications in Statistics: Theory and Methods*, 22:3, 827-842. Retrieved from <https://doi.org/10.1080/03610929308831059> doi: 10.1080/03610929308831059
- Pickands, J. (1971). The two-dimensional poisson process and extremal processes. *Journal of Applied Probability*, 8, 745–756. Retrieved from <https://doi.org/10.1214/ss/>

- 1177012400 doi: 10.2307/3212238
- Rasmussen, C. R., & Williams, C. K. I. (2005). *Gaussian processes for machine learning*. The MIT Press.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics, second edition* (Vol. 76) (No. 3). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2008.00062_10.x
- Reveles, F., Pérez-Abreu, V., Nakamura, M., & Biscay, R. (2016). *Notas en persistencia, probabilidad e inferencia estadística para análisis topológico de datos*. CIMAT, AC. Retrieved from https://atd.cimat.mx/sites/ATD/files/dwn/PPyETDA_Notas_y_Proyectos.pdf
- Roelants, P. (2019). *Understanding gaussian processes*. Retrieved from <https://peterroelants.github.io/posts/gaussian-process-tutorial/>
- Sejdinovic, D., & Gretton, A. (2014). *What is an rkhs?* Retrieved from www.stats.ox.ac.uk/sejdinov/teaching/atml14/Theory2014.pdf
- Silva, V., & Ghrist, R. (2007, 04). Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7. Retrieved from <https://msp.org/agt/2007/7-1/p16.xhtml> doi: 10.2140/agt.2007.7.339
- Smith, R. L. (1989, 11). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4(4), 367–377. Retrieved from <https://doi.org/10.1214/ss/1177012400> doi: 10.1214/ss/1177012400
- Thoppe, G., & Krishnan, S. R. (2018). Betti numbers of gaussian excursions in the sparse regime. *arXiv preprint, arXiv:1807.11018*. Retrieved from <https://arxiv.org/abs/1807.11018>