UNIVERSIDAD DE GUANAJUATO

División de Ciencias Naturales y Exactas

Simulation of Point Cloud Data with Various Probability Distributions on Stratified Spaces

T E S I S

Que para obtener el título de

Licenciado en Matemáticas

presenta

Yair Adán Hernández Esparza

Director de tesis:

Dr. Víctor Manuel Pérez Abreu Carrión

Guanajuato, Gto.

Marzo 2019

Agradecimientos

A mis padres, porque todo esto fue posible gracias a su apoyo incondicional. Gracias por el amor que me han dado desde que tengo memoria. Gracias por sus enseñanzas. Gracias por seguir ahí siempre a pesar de mis incontables errores.

Agradezco a mi asesor, el Dr. Víctor Pérez Abreu, por su paciencia, su guía y sus lecciones durante mis primeros semestres en la licenciatura, durante los dos años que fui ayudante de investigador, y a lo largo del desarrollo de esta tesis. En el transcurso de ese tiempo siempre conté con su apoyo y consejos.

A Gilberto Flores, compañero y amigo, con quien se desarrolló en conjunto el proyecto en el cual se basó este trabajo de tesis.

Al Dr. José Carlos Gómez Larrañaga por motivar el estudio de espacios estratificados y por sus enseñanzas durante el curso de Topología Computacional.

A Francisco Valente por facilitar el código que permite usar Ripser a través de `R`.

Agradezco al CIMAT por la formación que me brindó en un excelente ambiente para la investigación, por los apoyos de todo tipo que me fueron proporcionados a lo largo de mis estudios de licenciatura, y por la oportunidad de continuar mi formación con estudios de posgrado.

También a la Universidad de Tennessee por la oportunidad de presentar parte de este proyecto en las 47th Barrett Lectures en Knoxville en el año 2017, y por el apoyo económico proporcionado para ello.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Sistema Nacional de Investigadores (SNI) por el apoyo económico de ayudante de investigador (exp. inv. 4337, exp. ayudante 13730).

Contents

Introduction **1**

1 Preliminaries on Probability **4**

1.1 Hausdorff Measure 5

1.1.1 Carathéodory's Construction 5

1.1.2 Definition and Properties of the Hausdorff Measure 8

1.2 Area Formula 10

1.3 Uniformity 12

1.3.1 Uniformly Distributed Measure 13

2 Stratified Spaces **15**

2.1 Filtered Spaces 16

2.2 Definition and Example 17

3 Simulation Methods **22**

3.1 Uniform distribution on a parametrized manifold 22

3.1.1 Uniform Distribution on the Sphere 25

3.2 Distributions Induced on a Manifold Using the Parametrization 26

3.2.1 Independent Case 26

3.2.2 Dependent Case 27

3.3 Simulation on Stratified Spaces 32

3.3.1 Uniform Distribution on a Stratified Space 32

3.3.2 Special Cases: $\bar{X}_k \subsetneq X$ 35

4 Examples with Calculation of Persistent Homology 39

4.1 Analysis of Convergence 40

4.1.1 Polar Rose 40

4.1.2 Klein Bottle 44

4.2 Change of Persistent Homology in Mixture Models 50

A Acceptance-Rejection Method 53

A.1 The General Case 53

A.2 The Manifold Case 55

Introduction

Due to numerous applications, manifold learning has recently received considerable attention; see, for example, Ibañez et al. [19] or Zhu et al. [28]. The basic idea is as follows: given a Point Cloud of Data (PCD) sampled from a manifold in an ambient space $\mathbb{R}^d$, infer the underlying manifold [3].

The converse issue of simulating or drawing samples from a probability distribution on a manifold is also relevant in several problems in statistical and topological inference [10], in comparing algorithms used in the calculation of persistent homology [24], [26], or in evaluating MCMC methods on manifolds [8].

However, there are applications where spaces that are not manifolds arise; one example occurs in the cyclooctane energy landscape [22], which was found to have the structure of the union of a sphere with a Klein bottle intersecting in two rings. Stratified spaces make up a family with less restrictions than manifolds but with enough structure to allow relevant results to be used. Recently, Bendich et al.[2], used Topological Data Analysis (TDA) to understand musical audio data with this kind of space. In addition, from the theoretical point of view, Gómez-Larrañaga, González-Acuña and Heil [17] recently analyzed a particular case consisting of stratified surfaces.

The purpose of this undergraduate thesis is to study methods to simulate random variables with different distributions that have support on a stratified space and to provide examples of PCD using these techniques.

In 2013, Diaconis et al. [10] proposed a method to sample parametrized manifolds and included an example of the uniform distribution of the 2-Torus embedded in $\mathbb{R}^3$. Their method is based on the area formula and the Hausdorff measure [13].

We introduce simulation methods for random variables on parametrized manifolds with

probability distributions different to the uniform distribution. We consider two general cases. We call the first one the *independent* case, where we sample each parameter independently from the others. For the second case, we sample from the domain using copulas and random matrices theory, which model dependence between the parameters.

The particular stratified spaces that we work with are also rectifiable sets (for which the area formula is valid). Because they are piece-wise manifolds, we are able to provide a rich family of examples of PCD, such as data that exhibit repulsion, regions with higher concentration, among others.

As an illustration, we provide examples that calculate persistent homology and we conduct an empirical analysis of convergence using the stability theorem for persistence diagrams.

The organization of this thesis is as follows.

In Chapter 1, we present the preliminaries of measure theory, which allows us to define a probability distribution on manifolds and stratified spaces. These include a detailed study of Hausdorff measure, area formula, and different notions of uniformity.

In Chapter 2, we give a brief summary of the context of stratified spaces in the related literature, presenting the preliminaries, the concept of a stratified space, *cs*-sets, and basic examples.

In Chapter 3, we deal with simulation techniques on manifolds and stratified spaces. Section 3.1 describes the method proposed by Diaconis et al. [10] for simulating PCD with the uniform distribution on a parametrized manifold. We also review a well-known method for simulation from the uniform distribution on the n-sphere and the torus. Section 3.2 then presents the independent and dependent cases (as mentioned above) with illustrations of a simulated PCD on manifolds. The last section of this chapter will provide simulation methods on stratified spaces, following an idea suggested in Bendich et al. [3], which basically consists of a mixture model. Together with the techniques in the previous sections, this allows us to simulate PCD on stratified spaces from a wide variety of distributions.

Finally, in Chapter 4 we analyze in an empirical fashion the persistent homology obtained from the Vietoris-Rips filtration on PCD using three spaces: the polar rose, the Klein bottle and an example with maximal strata of lower dimensions. For the polar rose and the Klein

bottle, we analyze aspects on convergence when the size of the point cloud increases, aided by the stability theorem and a concentration inequality proposed by Fasy et al. [12]. In the third example, we study how the persistent homology changes when varying the weighs in the mixture model. For this chapter, we assume that the reader is familiar with basic notions of Topological Data Analysis; otherwise we recommend [11].

Chapter 1

Preliminaries on Probability

The Lebesgue measure extends the concepts of length, area, and volume to a wider collection of subsets of $\mathbb{R}^d$. There are, however, limitations to this extension. For example, a surface embeded in $\mathbb{R}^3$ will typically have 3-dimensional Lebesgue measure equal to 0. Here, we are interested in a measure that allows us to capture the 2-dimensional features of this set, regardless of the ambient space, which is the motivation behind the Hausdorff measure.

A natural question is how to calculate the Hausdorff measure of a set. If the set that we want to measure happens to be an image under a function f of a subset of $\mathbb{R}^k$, then the area formula allows us, under certain assumptions, to calculate the Hausdorff measure using the Lebesgue measure of the domain. This will be possible no matter what the dimension of the ambient space is.

Section 1.1 will describe the construction of the Hausdorff measure. Section 1.2 will provide a proof of the area formula. Finally, Section 1.3 will provide some common notions of uniformity.

The content of Sections 1.1 and 1.2 is mostly based on Section 19 of Billingsley's [5] book and Section 2.10.1 of Federer's [13] book. Section 1.3 is based on Chapter 3 of [6]. This thesis will use the notation λ^d for the d-dimensional Lebesgue measure; for the special case $d = 1$, we will simply write λ.

1.1 Hausdorff Measure

The results for this construction are taken from Federer [13] and Billingsley [5].

1.1.1 Carathéodory's Construction

This general construction is used to define measures in a metric space, such as the Hausdorff measure.

Let (X, ρ) be a metric space, F be a family of subsets of X, and $\zeta : F \to [0, \infty]$ be a set function such that $\emptyset \in F$ and $\zeta(\emptyset) = 0$.

For each $0 < \delta \leq \infty$, we define a set function $\phi_\delta : 2^X \to [0, \infty]$ as

$$\phi_\delta(A) = \inf_{S \in \mathcal{G}} \sum_{n=1}^{\infty} \zeta(S), \tag{1.1}$$

where the infimium is taken over all the countable families $\mathcal{G} \subset \{S \in F : \operatorname{diam}(S) \leq \delta\}$ and $A \subset \bigcup_{S \in \mathcal{G}} S$. Recall that the infimum of an empty set is defined as ∞. If $\sigma > \delta$, then the previous infimum is taken over a smaller family, so ϕ_δ does not decrease as δ decreases. Thus, we can define (for $A \subset X$)

$$\psi(A) = \lim_{\delta \to 0^+} \phi_\delta(A) = \sup_{\delta > 0} \phi_\delta(A). \tag{1.2}$$

ψ is called the *result of Carathéodory's construction from ζ on F*, and ϕ_δ is called the *size δ approximating measure.*

We will now prove that ψ and ϕ_δ (for every $\delta > 0$) are outer measures.

Proposition 1.1. *For every $\delta > 0$, ϕ_δ is an outer measure of X.*

Proof. Fix $\delta > 0$. It is clear that $\phi_\delta(A) \in [0, \infty]$ for every $A \subset X$, $\phi_\delta(\emptyset) = 0$ and $A \subset B$ implies $\phi_\delta(A) \leq \phi_\delta(B)$. We only need to prove that ϕ_δ is countably subadditive.

Fix $\varepsilon > 0$. Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of subsets of X. If $\phi_\delta(A_n) = \infty$ for any n, then clearly $\phi_\delta\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \phi_\delta(A_n)$. Now suppose $\phi_\delta(A_n) < \infty$ for every n. For each A_n, there is a cover $\{B_{nk}\}_{k=1}^{\infty} \subset \{S \in \mathcal{F} : \operatorname{diam}(S) \leq \delta\}$ such that $\sum_{k=1}^{\infty} \zeta(B_{nk}) < \phi_\delta(A_n) + \varepsilon/2^n$. Then,

$$\phi_\delta\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n,k} \zeta(B_{nk}) < \sum_{n=1}^{\infty} \phi_\delta(A_n) + \varepsilon.$$

This is true for every $\varepsilon > 0$, so the desired inequality follows. ■

Proposition 1.2. *ψ is an outer measure of X.*

Proof. Clearly $\psi(A) \in [0, \infty]$ for every $A \subset X$. As $\phi_\delta(\emptyset) = 0$ for every $\delta > 0$, we have $\psi(\emptyset) = 0$.

Let $A \subset B$ be subsets of X. For every $\delta > 0$, we have $\phi_\delta(A) \leq \phi_\delta(B)$, so by taking the limit on both sides, we have

$$\psi(A) = \lim_{\delta \to 0^+}(A) \leq \lim_{\delta \to 0^+}(B) = \psi(B).$$

If $A_1, A_2, \ldots$ are subsets of X, we have for every $\delta > 0$ that,

$$\phi_\delta\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \phi_\delta(A_n) \leq \sum_{n=1}^{\infty} \psi(A_n),$$

so

$$\psi\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{\delta \to 0^+} \psi_\delta\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \psi(A_n).$$

Because all conditions are verified, ψ is an outer measure. ■

At this point it is worth mentioning that although "Carathéodory's construction converts an arbitrary method ζ of estimation on F to a well behaved measure ψ over X" (Federer [13]), it does not necessarily extend ζ.

Because X has the metric topology induced by ρ, we are interested in measuring Borel subsets of X. The following theorem gives a sufficient condition for $\mathcal{B}(X)$ to be measurable by an outer measure μ^*.

Theorem 1.1 (Carathéodory's criterion)**.** *Let μ^* be an outer measure of a metric space (X, ρ). If $\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)$ whenever $\rho(A, B) > 0$, then every set in $\mathcal{B}(X)$ is μ^*-measurable.*

Proof. Because the collection of closed subsets of X generate $\mathcal{B}(X)$, it suffices to prove that every closed set is μ^*-measurable; that is, we must prove that for A closed and E arbitrary,

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A^c).$$

Let $B = A \cap E$, $C = A^c \cap E$, and for $n = 1, 2, \ldots$, let $C_n = \{x \in C : \rho(x, A) \geq 1/n\}$. As $\rho(B, C_n) > 0$ for every n, we have

$$\begin{aligned}
\mu^*(E) &= \mu^*(B \cup C) \\
&\geq \mu^*(B \cup C_n) \\
&= \mu^*(B) + \mu^*(C_n) \qquad \text{(Carathéodory's condition)}
\end{aligned}$$

Then, we only need to prove that $\mu^*(C_n) \to \mu^*(C)$. Because $C_n \uparrow C$, it is equivalent to prove that $\lim_n \mu^*(C_n) \geq \mu^*(C)$.

Let $D_n = C_{n+1} \setminus C_n$. We will prove that $\rho(D_{n+1}, C_n) > 0$, whenever they are nonempty. Let $x \in D_{n+1} = C_{n+2} \setminus C_{n+1}$; that is, $x \in C$ and

$$\frac{1}{n+2} \leq \rho(x, A) < \frac{1}{n+1}.$$

If y is such that $\rho(y, x) < n^{-1}(n+1)^{-1}$, we have

$$\begin{aligned}
\rho(y, a) &\leq \rho(y, x) + \rho(x, A) \\
&< \frac{1}{n}\frac{1}{n+1} + \frac{1}{n+1} \\
&= \frac{1}{n} - \frac{1}{n+1} + \frac{1}{n+1} = \frac{1}{n},
\end{aligned}$$

so $y \notin C_n$. Therefore, if $y \in C_n$, $\rho(x, y) \geq n^{-1}(n+1)^{-1}$. Because $D_{n-1} \subset C_n$, we have $n^{-1}(n+1)^{-1} \leq \rho(D_{n+1}, C_n) \leq \rho(D_{n+1}, D_{n-1})$. By the Carathéodory condition, we get by induction

$$\mu^*(C_{2n+1}) \geq \mu^*\left(\bigcup_{k=1}^{n} D_{2k}\right) = \sum_{k=1}^{n} \mu^*(D_{2k}), \tag{1.3}$$

$$\mu^*(C_{2n}) \geq \mu^*\left(\bigcup_{k=1}^{n} D_{2k-1}\right) = \sum_{k=1}^{n} \mu^*(D_{2k-1}). \tag{1.4}$$

By subadditivity,

$$\mu^*(C) \leq \mu^*(C_{2n}) + \sum_{k=n}^{\infty} \mu^*(D_{2k}) + \sum_{k=n+1}^{\infty} \mu^*(D_{2k-1}). \tag{1.5}$$

If either $\sum \mu^*(D_{2k})$ or $\sum \mu^*(D_{2k-1})$ diverge, $\mu^*(C) \leq \lim_n \mu^*(C_n)$ follows from (1.3); otherwise, it follows from (1.5).

■

1.1.2 Definition and Properties of the Hausdorff Measure

Although we will use the Hausdorff Measure for subsets of $\mathbb{R}^d$, it is defined for a general metric space.

Let (X, ρ) be a metric space and k be a positive real number. In the Carathéodory construction consider $F = 2^X$ and $\zeta : F \to [0, \infty]$ given by $\zeta(S) = c_k \operatorname{diam}(S)^k$, where c_k is a positive constant, which will be defined later. The approximating outer measures ϕ_δ and resulting outer measure ψ will be denoted by $\mathcal{H}^k_\delta$ and $\mathcal{H}^k$, respectively. $\mathcal{H}^k$ is the k-dimensional Hausdorff outer measure of X.

Specifically, for $\delta > 0$ and $A \subset X$ we have

$$\mathcal{H}^k_\delta(A) := \inf c_k \sum_n (\operatorname{diam} B_n)^k, \tag{1.6}$$

where the infimum is taken over all the countable coverings of A by sets B_n with diameters $\operatorname{diam} B_n = \sup\{\rho(x, y) : x, y \in B_n\}$ less than δ.

As stated in the previous section, we are interested in measuring Borel subsets of X. We will use the Carathéodory condition to prove that $\mathcal{H}^k$ is well defined over $\mathcal{B}(X)$. Let $A, B \subset X$ be such that $\rho(A, B) = \inf\{\rho(x, y) : x \in A, y \in B\} > 0$. Let $\varepsilon > 0$ be such that $\varepsilon < \rho(A, B)$. If $A \cup B \subset \bigcup C_n$ and $\operatorname{diam} C_n < \varepsilon$ for all n, then no C_n can intersect both A and B, so $\sum c_k(\operatorname{diam} C_n)^k$ may be split into the series of those that intersect A and B, respectively. This shows that it is at least $\mathcal{H}^k_\varepsilon(A) + \mathcal{H}^k_\varepsilon(B)$. Therefore, $\mathcal{H}^k_\varepsilon(A \cup B) \geq \mathcal{H}^k_\varepsilon(A) + \mathcal{H}^k_\varepsilon(B)$. This is true for every $0 < \delta < \varepsilon$, so taking the limit when $\delta \to 0^+$ yields

$$\mathcal{H}^k(A \cup B) \geq \mathcal{H}^k(A) + \mathcal{H}^k(B). \tag{1.7}$$

As $\mathcal{H}^k$ is an outer measure, the other inequality follows, so $\mathcal{H}^k(A \cup B) = \mathcal{H}^k(A) + \mathcal{H}^k(B)$. Thus, Carathéodory's condition is satisfied and $\mathcal{H}^k$ restricted to $\mathcal{B}(X)$ is a measure.

We now only need to choose an adequate constant c_k. The motivation for defining $\mathcal{H}^k$ is to extend the Lebesgue measure, so it is reasonable to require that $\mathcal{H}^k$ and λ^k agree on $\mathcal{B}(\mathbb{R}^k)$. Let V_k be the volume of a ball of radius 1; that is, $V_1 = 2$ and

$$V_{2i-1} = \frac{2(2\pi)^{i-1}}{1 \cdot 3 \cdot \ldots \cdot (2i-1)}, \quad V_{2i} = \frac{(2\pi)^i}{2 \cdot 4 \cdot \ldots \cdot (2i)}.$$

More generally, we have

$$V_k = \frac{\Gamma\left(\frac{1}{2}\right)^k}{\Gamma\left(\frac{k}{2}+1\right)},$$

allowing k to be any positive real number, although we will only consider integer values of k. We take $c_k = V_k/2^k$ the volume of a ball of diameter 1. For this choice of c_k, we have the following result.

Theorem 1.2. *If $A \in \mathcal{B}(\mathbb{R}^k)$, then $\mathcal{H}^k(A) = \lambda^k(A)$.*

Before proving this, we make some remarks. The unit cube $C = [0,1]^k$ in $\mathbb{R}^k$ can be covered by n^k cubes of side n^{-1} and diameter $\sqrt{k}n^{-1}$. If $n > \sqrt{k}\varepsilon^{-1}$, then $\mathcal{H}^k_\varepsilon(C) \leq c_k n^k (k^{1/2}) = c_k\sqrt{k^k}$, showing that $\mathcal{H}^k(C) < \infty$. If $C \subset \bigcup_n B_n$ and $\operatorname{diam} B_n = d_n$, enclose B_n in a closed ball S_n of radius d_n. Then $C \subset \bigcup_n S_n$ and so $1 = \lambda^k(C) \leq \sum_n \lambda^k(S_n) = \sum_n V_k d_n^k$. Thus, $\mathcal{H}^k(C) \geq c_k/V_k$ so $\mathcal{H}^k(C)$ is nonzero.

Now, if $\mathcal{H}^k(C) = K$, because $\lambda^k(C) = 1$, then we have $\mathcal{H}^k(C) = K\lambda^k(C)$. Consider the mapping $x \mapsto \theta x$ for a fixed $\theta > 0$. Because this mapping is linear and nonsingular, $A \in \mathcal{B}(\mathbb{R}^k)$ implies $\theta A \in \mathcal{B}(\mathbb{R}^k)$ and $\lambda^k(\theta A) = \theta^k \lambda^k(A)$. We also have that $\operatorname{diam}(\theta A) = \theta^k \operatorname{diam} A$, so it follows from (1.6) that

$$\mathcal{H}^k(\theta A) = \theta^k \mathcal{H}^k(A). \tag{1.8}$$

Then $\mathcal{H}^k(A) = K\lambda^k(A)$ holds for every cube A, and by additivity it holds for rectangles whose vertices have rational coordinates. Because the family of this kind of rectangles form a π-system that generates $\mathcal{B}(\mathbb{R}^k)$, $\mathcal{H}^k(A) = K\lambda^k(A)$ holds for every $A \in \mathcal{B}(\mathbb{R}^k)$.

Then, to prove theorem 1.2, we only need to prove $K = 1$. To do this, we will use two lemmas.

Lemma 1.1. *Suppose that G is a bounded open set in $\mathbb{R}^k$ and that $\varepsilon > 0$. Then, there exists in G a disjoint sequence $S_1, S_2, \ldots$ of closed balls such that $\lambda^k(G \setminus \bigcup_n S_n) = 0$ and $0 < \operatorname{diam} S_n < \varepsilon$.*

Lemma 1.2. *If $A \in \mathcal{B}(\mathbb{R}^k)$, then $\lambda^k(A) \leq c_k(\operatorname{diam} A)^k$.*

Proof of theorem 1.2. Let $C = [0,1]^k$ be the unit cube in $\mathbb{R}^k$. We need to prove that $K := \mathcal{H}^k(C) = 1$. Let $\varepsilon > 0$. By lemma 1.1, we can cover the interior of C with a sequence of disjoint closed balls $S_1, S_2, \ldots$ such that $\operatorname{diam} S_n < \varepsilon$ and

$$\mathcal{H}^k_\varepsilon(S \setminus \bigcup_n S_n) \leq \mathcal{H}^k(C \setminus \bigcup_n S_n) = K\lambda^k(C \setminus \bigcup_n S_n) = 0.$$

Because $c_k = V_k/2^k$, we have

$$\begin{aligned}\mathcal{H}^k_\varepsilon(C) = \mathcal{H}^k_\varepsilon(\bigcup_n S_n) &\leq c_k \sum_n (\operatorname{diam} S_n)^k \\ &= c_k \sum_n \tfrac{1}{c_k}\lambda^k(S_n) \leq \lambda^k(C).\end{aligned}$$

This is true for every $\varepsilon > 0$, so $\mathcal{H}^k(C) \leq 1$.

If $C \subset \bigcup_n B_n$, then by lemma 1.2 we have $1 \leq \sum_n \lambda^k(B_n) \leq \sum_n c_k(\operatorname{diam} B_n)^k$, so $\mathcal{H}^k(C) \geq 1$. ■

Corolary 1.1. *If $A \subset \mathbb{R}^m$ and $B \subset \mathbb{R}^k$ is an isometric copy of A, then $\mathcal{H}^m(B) = \mathcal{H}^m(A) = \lambda^m(A)$.*

1.2 Area Formula

This section is based mainly on Section 2.1 in Diaconis et al.'s [10] paper. For a discussion at length on these topics, see, for example, the book by Federer [13].

Given the k-dimensional Hausdorff measure, we would now like to use it on an ambient space. With this goal in mind, we present some definitions and results. The first ones that we will need are those of a *Lipschitz function* and a *rectifiable set.*

Definition 1.1. A function $f : \mathbb{R}^k \to \mathbb{R}^d$ is *Lipschitz* if there is a positive constant $c > 0$ such that $\|f(x) - f(y)\| \leq c\,\|x - y\|$ for all $x, y \in \mathbb{R}^k$ (with $\|\cdot\|$ the usual Euclidean norm).

A set in $\mathbb{R}^d$ is rectifiable if it is the image of a bounded subset in $\mathbb{R}^k$ under a Lipschitz function.

Remark. (Federer[13]) In (1.6), the coverings can be restricted to balls or cubes if A is rectifiable.

$Df(x)$ will denote the matrix associated to the derivative of f at x in the typical sense, when it exists. A theorem by Rademacher establishes the existence λ^k-almost everywhere of $Df(x)$ when f is a Lipschitz function. In this case we can define the i-th dimensional Jacobian, which is denoted by $J_i f(x)$.

Definition 1.2. Let $f : \mathbb{R}^k \to \mathbb{R}^d$, be differentiable at $x \in \mathbb{R}^d$. We define the *i-dimensional Jacobian* of f at x, $J_i f(x)$, as the *maximum i-dimensional volume* of the image of a i-dimensional unit cube under $Df(x)$; that is,

$$J_i f(x) = \max_C \text{vol}(Df(x)(C)),$$

where C is a i-dimensional unit cube.

When $i = k$ we have (Diaconis et al. [10])

$$J_k f(x) = \sqrt{det(Df(x)^T Df(x))}.$$

We now present the area formula theorem. This theorem allows us to compute an integral with respect to $\mathcal{H}^k$ of a function on a k-dimensional manifold by computing instead of an integral with respect to the Lebesgue measure of a function over $\mathbb{R}^k$.

Theorem 1.3 (Area Formula)**.** *Let $f : \mathbb{R}^k \to \mathbb{R}^d$ be Lipschitz, with $k \leq d$. We define*

$$N(f|_A, y) = \#\{x \in A \,:\, f(x) = y\}.$$

Then:

1. *If A is λ^k-measurable,*

$$\int_A J_k f(x) \lambda^k(dx) = \int_{\mathbb{R}^d} N(f|_A, y) \mathcal{H}^d(dy).$$

2. *Furthermore, if $g : \mathbb{R}^k \longrightarrow \mathbb{R}$ is any Borel function,*

$$\begin{aligned} \int_A g(f(x)) J_k f(x) \lambda^k(dx) &= \int_{\mathbb{R}^d} g(y) N(f|_A, y) \mathcal{H}^k(dy) \\ &= \int_{\mathbb{R}^d} \sum_{x \in f^{-1}(y)} g(x) \mathcal{H}^k(dy). \end{aligned}$$

We will be interested in taking $f : A \to \mathbb{R}^d$ a parametrization of a manifold. However, the theorem requires the function to be Lipschitz over all $\mathbb{R}^k$. The following theorem allows us to use the Area Formula more generally without any issue. The proof can be read, for example, in [13].

Theorem 1.4 (Kirszbraun)**.** *If $A \subset \mathbb{R}^k$ and $f : A \to \mathbb{R}^d$ is Lipschitz, then there is a Lipschitz function $g : \mathbb{R}^k \to \mathbb{R}^d$ such that $g|_A = f$ and both have the same Liptschitz constant.*

1.3 Uniformity

The term *uniform* can be found throughout the literature of topological data analysis. The purpose of this section is to define with precision the meanings this term can take. These definitions will be used in Section 4. In [6] this topic is discussed in more detail.

Usually, the first notion of *uniform distribution* is that of a probability measure on the power set of a finite set $\{a_1, \ldots, a_k\}$: that is, $\mathbb{P}(a_i) = 1/k$ for $i = 1, \ldots, k$.

The next natural notion of this concept is the uniform measure on an interval $[a, b]$. The uniform distribution (probability) on $\mathcal{B}([a, b])$ is that which to every set $A \in \mathcal{B}(\mathbb{R})$ assigns the probability

$$\mathbb{P}[A] = \frac{\lambda(A \cap [a, b])}{b - a}.$$

An immediate extension of this case is that of the uniform distribution on a compact set $K \subset \mathbb{R}^d$, where the uniform distribution on K is defined as

$$\mathbb{P}[A] = \frac{\lambda^d(A \cap K)}{\lambda^d(K)} = \int_A \frac{\mathbf{1}_K}{\lambda^d(K)} d\lambda^d.$$

In the last example, the uniform distribution is obtained from a previously constructed measure, which in this case is λ^d, and the new measure relies on the fact that $0 < \mu(K) < \infty$.

With this idea in mind, we give a general definition of a uniform distribution.

Definition 1.3 ([6])**.** Let μ be a measure space on a metric space $(\mathcal{M}, \rho)$, and $K \in \mathcal{B}(\mathcal{M})$ such that $0 < \mu(K) < \infty$. Let $\mathcal{B}(K) = \mathcal{B}(\mathcal{M}) \cap K$. The probability measure μ_K on $\mathcal{B}(K)$ defined by

$$\mu_K(A) = \frac{\mu(A)}{\mu(K)}, \quad A \in \mathcal{B}(K).$$

is μ-uniform: that is, $\mu_K(A) = \mu_K(B)$ if and only if $\mu(A) = \mu(B)$. K is usually a compact set and μ a Radon measure.

Using the Hausdorff measure on a manifold $\mathcal{M}$, in Section 3.1 we will be able to define the uniform distribution on a manifold.

A second notion of uniformity is for measures (not necessarily probability measures) defined on metric spaces. A measure μ defined on a metric space $(\mathbb{M}, \mathcal{B}(\mathbb{M}))$ is uniform if for every $\varepsilon > 0$, $\mu(B_\varepsilon(x_1)) = \mu(B_\varepsilon(x_2))$ for all $x_1, x_2 \in \mathbb{M}$; that is, all balls with the same radius have the same measure.

One last notion of uniformity arises in the context of random vectors/matrices: the uniform distribution is invariant under left orthogonal (or unitary) transformations in the *Stiefel manifold*, which is defined as follows. Let $\mathbb{R}^{d\times p}$ be the vector space of $d \times p$ matrices with real entries and norm given by

$$\|S\|^2 = \frac{1}{d}\mathrm{Tr}(S^T S),\ S \in \mathbb{R}^{d\times p}.$$

Then, the Stiefel manifold $\mathbb{L}_p^d$ is defined as

$$\mathbb{L}_p^d = \{T \in \mathbb{R}^{d\times p} :\ T^T T = \mathrm{I}_p\}.$$

For a brief discussion and references on this topic, see Section 3.3.1 of [6].

1.3.1 Uniformly Distributed Measure

In this section, we will consider measures defined on the Borel σ-algebra of a polish metric space; that is, measures on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, where $(\mathcal{M}, \rho)$ is a metric space.

Definition 1.4. A measure defined over $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ is a *Radon* Measure if it satisfies the following two conditions:

- μ is Borel: this is, for every $x \in \mathcal{M}$ there is $r \in (0, \infty)$ such that $\mu(B_r(X)) < \infty$.
- μ is inner regular: for every $A \in \mathcal{B}(\mathcal{M})$,

$$\mu(A) = \sup\{\mu(K) : K \subset A, K \text{ is compact}\}.$$

Probability measures on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ and Lebesgue-Stieltjes measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are examples of Radon measures.

Definition 1.5. The *support* of a Radon measure μ on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ is defined as the set

$$\operatorname{supp}(\mu) = \bigcap_{\substack{\mu(C^c)=0,\\ C \text{ is closed}}} C$$

It is well-defined, as $\mathcal{M}$ is a closed subset, and its complement is a null set.

Definition 1.6. A Radon measure μ on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ is *uniformly distributed* if

$$\mu(B_r(x)) = \mu(B_r(y)), \quad \forall x, y \in \operatorname{supp}(\mu), r \in (0, \infty)$$

were $B_r(x) = \{y \in \mathcal{M} : \rho(x, y) < r\}$.

Theorem 1.5 (Christensen)**.** *If μ_1, μ_2 are two uniformly distributed measures on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, then there is a constant $0 < c < \infty$ such that $\mu_1 = c\mu_2$.*

The previous theorem means that up to a multiplicative constant, the uniform measure in a polish metric space is unique.

The Lebesgue measure and the counting measure are examples of uniformly distributed measures.

Remark 1. Since a compact set K is bounded, the uniform distribution described on definition 1.3 is not a uniformly distributed measure.

Chapter 2

Stratified Spaces

There are topological spaces that are not manifolds but which can be separated into pieces that are manifolds and fit together 'nicely'. This is the basic idea behind a stratified space. These spaces may or may not be subsets of $\mathbb{R}^d$. In [16], for example, one of such spaces is constructed with equivalence relations on a surface, and it is not immediately seen as a subset of an Euclidean space (although they may be homeomorphic to one). For our purposes, however, we will only consider topological spaces that are embeded in some Euclidean space $\mathbb{R}^d$.

The requirements on how the pieces that are manifolds need to fit together vary across the literature where they appear. Hughes and Weinberger's [18] paper gives an extensive review on different requirements for a stratified space that have come up across related literature. The study of these spaces has received special attention in areas like homology/cohomology and topological data analysis (which is strongly based on the first). Bendich et al. [3], for example, present a method for clustering data points into different strata. Bhattacharya et al. in [4] present one version of the Central Limit Theorem for *random objects* supported on a stratified space. In a recent paper titled "Challenges in Topological Object Data Analysis" by Patrangenaru et al. [25], the objects of interest are "points on some manifold or *stratified space*".

In our context, we are interested in working with so called *cs*-spaces. This chapter is mostly based on the definitions given in the lecture notes by Friedman [15], where the goal is *"to provide a single coherent exposition of the basic piecewise linear and singular chain*

intersection homology theory as it has come to exist today". There, a succession of classes of spaces is introduced, beginning with general filtered topological spaces and introducing each time more requirements. We are interested in this development of the definition of a *cs*-space because some of the preliminary spaces come up in some of the topological data analysis-related literature, such as in [1].

In this chapter we aim to give a brief introduction to notions of stratified spaces needed in Chapter 3. In Section 2.1, we give preliminaries and we then define stratified spaces and *cs*-sets, and give a basic example.

2.1 Filtered Spaces

We assume all topological spaces to be Hausdorff. We begin with the definition of a *filtered topological space.*

Definition 2.1. A *filtered space* is a Hausdorff topological subspace X together with a sequence of closed subsets

$$\emptyset = X^{-1} \subset X^0 \subset X^1 \subset \ldots \subset X^n = X,$$

for some integer $n \geq -1$, which will be the formal dimension of X. The space X^i is called the *i-skeleton*, and it has *formal dimension i.*

Remarks.

- The smallest integer is always -1 and X^{-1} is always the empty set, so it will not be mentioned explicitly.
- It is possible to have $X^i = X^{i-1}$.
- We will usually be working with X a subset of $\mathbb{R}^n$. In these cases, each subspace in the filtration must be a closed set in the topology of X, although some of them may not be closed sets in the topology of $\mathbb{R}^n$.
- For this class of spaces, i is the formal dimension of the i-skeleton. This formal dimension does not necessarily relate with other concepts of dimension.

Definition 2.2. For a filtered space X of formal dimension n we define $X_i = X^i \setminus X^{i-1}$. The connected components of X_i are called the *strata* of X of formal dimension i or formal codimension $n - i$. The strata of all dimensions partition X.

The strata of formal dimension n are the *regular* strata and all the others are the *singular* strata.

Example 2.1. Consider $X = \{(x, y) \in \mathbb{R}^2 : y > 0\} \cup Y$, with $Y = \{(x, y) : x = 0\}$ the vertical axis. X is a filtered space with the filtration

$$\begin{aligned} \emptyset \subset X^0 &= \{(x, y) \in \mathbb{R}^2 : x \leq 0, y > 0\} \cup \{(0, 0)\} \\ &\subset X^1 = X. \end{aligned}$$

Note X^0 is not a closed subset of $\mathbb{R}^2$, but it is a closed subset of X. X is also a filtered space with the filtration given by $\emptyset \subset Y \subset X$.

Example 2.2. Let X be a finite-dimensional simplicial complex. Its simplicial skeleta induce a filtration, where the strata are the open simplices.

This example is important because one of our motivations is to work with datasets in Topological Data Analysis, where triangulable spaces are of special importance.

To avoid some of the possible pathologies in how the strata can fit together, we introduce the *frontier condition*.

2.2 Definition and Example

Definition 2.3. We say the filtered space X satisfies the *frontier condition* if for any two strata S, T,

$$S \cap \bar{T} \neq \emptyset \Rightarrow S \subset \bar{T}, \tag{2.1}$$

where $\bar{T}$ is the closure of T (in X).

Definition 2.4. A *stratified space* is a filtered space that satisfies the frontier condition.

Remark. If S, T are strata of a stratified space X, we will write $S \prec T$ if $S \subset \bar{T}$. We have that $\prec$ is a partial order. We also have that the closure of any stratum is a union of strata of lower dimension: $\bar{T} = \bigcup_{S \prec T} S$.

At this point, we introduce the definition of a manifold because the next type of spaces will have manifolds as strata.

Definition 2.5. A (topological) n-manifold $\mathcal{M}$ is a space that is locally homeomorphic to $\mathbb{R}^n$; that is, there exists a covering $\mathcal{U} = \{U_\alpha\}$ of $\mathcal{M}$ along with homeomorphisms

$$\phi_\alpha : U_\alpha \to \mathbb{R}^n.$$

Definition 2.6. A *manifold stratified space* is a stratified space if every i-dimensional stratum is a (possibly empty) i-manifold.

Remark. It can be easily verified that a finite dimensional simplicial complex X filtered by its simplicial skeleta satisfies the frontier condition (it actually follows from the definition). We also have that the strata are the open simplices, so X is a manifold stratified space.

For our purposes, the definition of a manifold stratified space is enough. However, even though they have a nice structure, it is still a general setting and more requirements are needed for the results. Thus, we present two more definitions that are commonly used in the TDA-related literature.

First, we recall the definition of the *cone* of a topological space X.

Definition 2.7. The (closed) cone $\bar{C}X$ of a compact topological space X is defined as the quotient space $(X \times [0,1])/(X \times \{0\})$. Intuitively speaking, it is obtained from collapsing one end of the cylinders $X \times [0,1]$ into a point. The open cone CX is the topological space obtained from the quotient

$$X \times [0,1) \Big/ X \times \{0\}.$$

Alternatively, it may be seen as the space obtained from removing 'the other end' from the closed cone: $CX = \bar{C}X \setminus (X \times \{1\})$.

Definition 2.8. A filtered space X of formal dimension n is *locally cone-like* if for each i and $x \in X_i$ there is a neighborhood U of $x \in X_i$, a neighborhood N of x in X, a (possibly empty) compact filtered space L, and a homeomorphism $h : U \times cL \to N$ such that

$$h(U \times C(L^k)) = X^{i+k+1} \cap N.$$

L is called a *link* of x and N is called a *distinguished neighborhood* of x.

A *cs*-set is a locally cone-like filtered space whose i-dimensional strata are i-dimensional manifolds. In this case, the previous requirement is equivalent to the following: for every $x \in X_i$, there is a neighborhood N of x and a filtration-preserving homeomorphism $N \cong \mathbb{R}^i \times CL$, where L is a compact $(n-i-1)$-dimensional stratified space.

The definition of a *cs*-set is used, for example, in Bendich et al. [3]. We next give an example of a *cs*-set.

Example 2.3. Consider the space $X \subset \mathbb{R}^3$ consisting of a 2-sphere and a disc in its equator, both of which are punctured by a closed curve as shown:

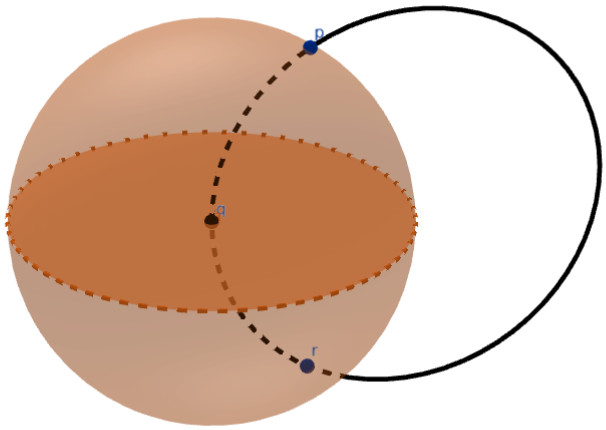


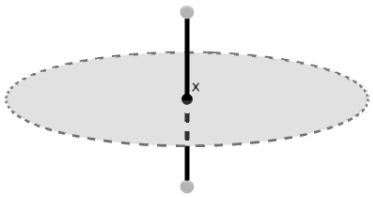
Figure 2.1: Example 2.3

We will denote the sphere by Γ_2, the disc by Γ_1, the equator by β, the closed curve by α, and the intersection points of α by p, q, r as shown. We propose the following filtration:

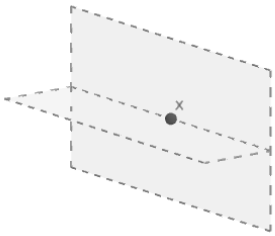
$$X^2 = X \supseteq X^1 := \alpha \cup \beta \supseteq X^0 := \{p, q, r\} \supseteq = \emptyset.$$

Now let us see that this filtration meets the required condition.

(a) $X_0 = \{p, q, r\}$. We see that if $x \in S_0$, x has an open neighborhood homeomorphic to $\mathbb{R}^0 \times CL_x \cong CL_x$, where L_x is a space consisting of a circle and two external points (it can be easily verified that this is a stratified space of dimension $2 - 0 - 1 = 1$).

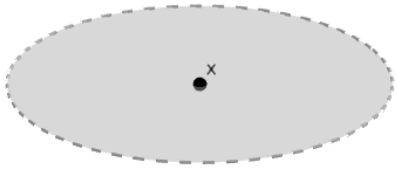


(b) $X_1 = (\alpha \cup \beta) \setminus \{p, q, r\}$. If $x \in X_1$ lies in β, then x has an open neighborhood homeomorphic to $\mathbb{R}^1 \times CL_x$, where L_x is a space consisting of three points (which is a stratified space of dimension $2 - 1 - 1 = 0$).



(c) If $x \in X_1$ lies in α, then x has an open neighborhood that is homeomorphic to $\mathbb{R}^1 \times CL_x$, where L_x is the empty set (recall that the cone of the empty set consists of only one point).

(d) We have that $X_2 = \Gamma_1 \cup \Gamma_2 \setminus (\alpha \cup \beta)$, which is a manifold consisting of three connected components where each component is homeomorphic to an open disc. Thus, every $x \in X_2$ has an open neighborhood homeomorphic to $\mathbb{R}^2 \times C\emptyset$.


x

Chapter 3

Simulation Methods

In this chapter, we will deal with simulation techniques on manifolds and stratified spaces. Section 3.1 is based on Diaconis et al. [10], and we recall a method by Marsaglia [21] to simulate from the uniform distribution on the 2-sphere, and an extension of this result. Section 3.2 introduces two general methods to simulate point cloud data on parametrized manifolds: the first considering independent parameters, and the second with models for dependence using copulas of multivariate distributions and the circular law in random matrices theory. In Section 3.3, we provide simulation methods on stratified spaces.

3.1 Uniform distribution on a parametrized manifold

In this section, we present an algorithm to simulate $\mathcal{H}^k$-uniformly distributed points on a particular on a k-dimensional parametrized manifold embedded in $\mathbb{R}^d$ ($k < d$). The content presented in this section is based on Diaconis et al.'s [10] paper, where only the example of the torus is presented.

The Hausdorff measure allows us to define the concept of volume for a wider class of sets, and the area formula

$$\int_A g(f(x))J_k f(x)\lambda^k(dx) = \int_{\mathbb{R}^d} g(y)N(f|_A, y)\mathcal{H}^k(dy) \tag{3.1}$$

relates this measure with the Lebesgue measure on $\mathbb{R}^k$. Thus, we can get a sample with a particular distribution with respect to the Hausdorff measure from a distribution on the

domain. In the context of this work, f is a parametrization of a manifold $\mathcal{M}$ and A is a Lebesgue measurable subset in the domain of f. Note that for $y \notin f(A)$, $N(f|_A, y) = 0$, so the right-hand integral of (3.1) is calculated just on $\mathcal{M}$. Hence, the goal is to get a sample of points from the density $J_m f / \operatorname{vol}(\mathcal{M})$ (on the domain of f); noting that by making $g \equiv 1/\operatorname{vol}(\mathcal{M})$, we obtain from the area formula

$$\int_A \frac{1}{\operatorname{vol}(M)} J_k f(x) \lambda^k(dx) = \int_{\mathbb{R}^d} \frac{1}{\operatorname{vol}(M)} N(f|_A, y) \mathcal{H}^k(dy).$$

In the examples presented later $N(f|_A, y)$ is equal to 1 or 0, $\mathcal{H}^k$-a.s., and therefore the right-hand integral is equal to the $\mathcal{H}^k$-uniform measure of $f(A)$ on $\mathcal{M}$.

Given that the density function $J_k f / \operatorname{vol}(\mathcal{M})$ may not be a known distribution, the acceptance-rejection (A-R) method is useful to get the desired samples. This method (and how to use it in this context) is discussed in detail in Appendix A. The domain is usually of the form $D = \prod_{i=1}^k [a_i, b_i]$; so to apply the A-R method, we may take g as the λ^k-uniform distribution over D.

There are times at which the domain may take a more general form. In this case, it is also possible to use the A-R method to get a sample from the uniform distribution on the parametrization's domain: get a uniform sample from a k-dimensional rectangle (a set of the form of D in the previous paragraph) and then reject those points outside of the domain of f.

Example 3.1. The 2-torus $\mathbb{T}^2$ embedded in $\mathbb{R}^3$ can be parametrized using

$$\begin{aligned} x &= (R + r\cos(\theta))\cos(\phi), \\ y &= (R + r\cos(\theta))\sin(\phi), \\ z &= r\sin(\theta), \end{aligned}$$

for $0 \leq \theta, \phi < 2\pi$, where R is the distance from the center of the torus to the center of the tube (major radius) and r is the radius of the tube (minor radius), and $R > r > 0$.

A quick calculation yields

$$J_2^2 f(\theta, \phi) = r^2 (R + r\cos(\theta))^2,$$

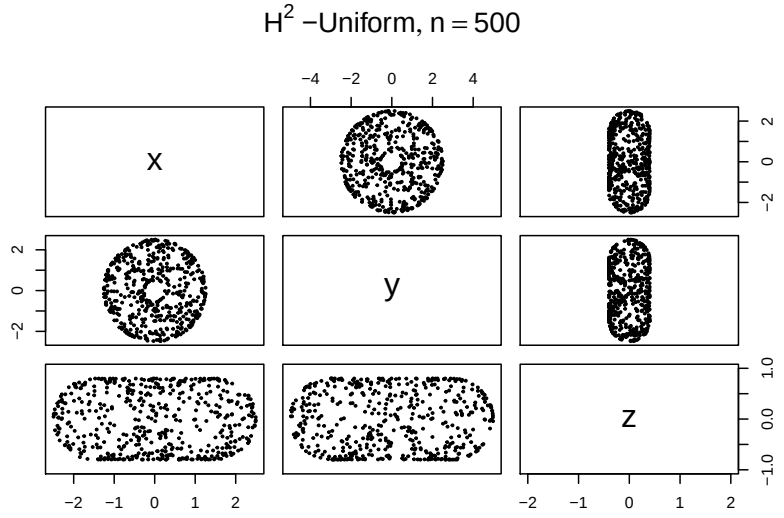
so

$$\frac{1 + (r/R)\cos(\theta)}{4\pi^2} \tag{3.2}$$

is the density on $[0, 2\pi) \times [0, 2\pi)$ which induces the $\mathcal{H}^2$-uniform measure on $\mathbb{T}^2$ via the area formula. Note that (3.2) can be written as the product of the uniform density on $[0, 2\pi)$ and a density on $[0, 2\pi)$ which depends only of θ:

$$\frac{1}{2\pi}\frac{1 + (r/R)\cos(\theta)}{2\pi}.$$

Therefore, sampling from the density $J_2 f / \operatorname{vol}(\mathbb{T}^2)$ is equivalent to sampling two independent random variables on $[0, 2\pi)$ from the mentioned densities. For the case of the density $(1 + (r/R)\cos(\theta))/(2\pi)$, we can use the acceptance-rejection method. The following image shows the contrast between sampling from the $\mathcal{H}^2$-uniform distribution on the torus and sampling from the λ^2-uniform distribution on $[0, 2\pi) \times [0, 2\pi)$ (and mapping those points to the torus). Using the second method, the sample is concentrated at the central part of the torus.



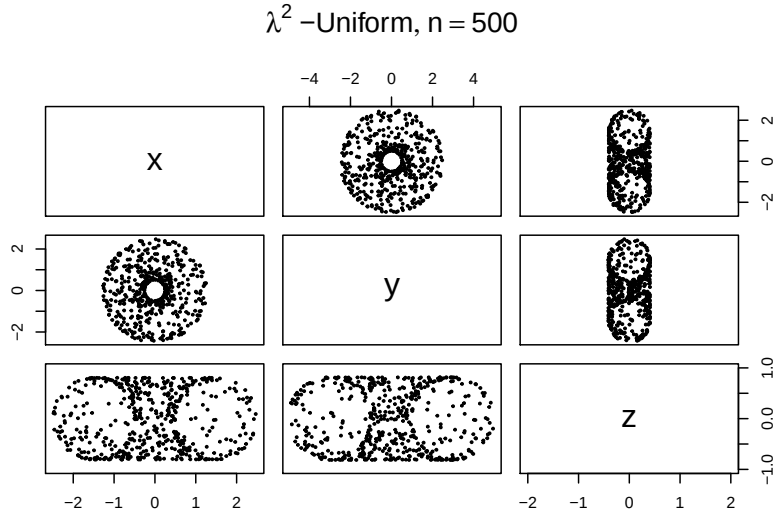


Figure 3.1: Top: Sample of size 500 from the $\mathcal{H}^2$-uniform measure on the torus. Bottom: sample using the λ^2-uniform distribution.

3.1.1 Uniform Distribution on the Sphere

For the particular case of the uniform distribution on the n-sphere, there is a simple algorithm that was introduced by Marsaglia [21] and Muller [23].

If X is a 3-dimensional random vector with the trivariate standard distribution (mean $(0,0,0)^T$ and the identity matrix as covariance matrix), then $X/\|X\|$ will have uniform distribution on the unitary sphere embeded in $\mathbb{R}^3$. For sampling from the uniform distribution on the sphere centered on $\mathbf{a} \in \mathbb{R}^3$ with radius $r > 0$, it suffices to take $\mathbf{a} + rX/\|X\|$.

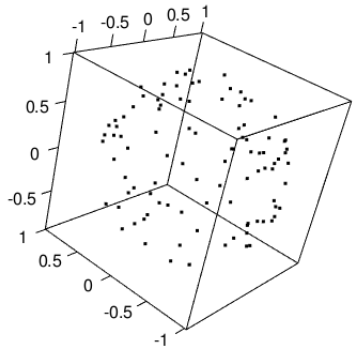


Figure 3.2: sample from the uniform distribution on the 2-sphere

This result extends immediately to dimensions greater than two: if X has multivariate normal distribution with mean $0 \in \mathbb{R}^k$ and the identity (of size $k \times k$) as covariance matrix, then $X/\|X\|$ is $\mathcal{H}^{k-1}$-uniformly distributed on the $(k-1)$-sphere.

These ideas are extended in Pérez-Angulo [26] to simulate PCD on the sphere and the torus, with a large class of distributions other than the uniform distribution.

3.2 Distributions Induced on a Manifold Using the Parametrization

In the previous section, we studied the notion of uniform distribution on a manifold and a method to sample from this distribution was presented. Now, we show some distributions on manifolds induced by distributions on the domain of a parametrization.

In the following sections, we will deal with a manifold $\mathcal{M}$ parametrized by g with domain A. Given a distribution F on A, under certain assumptions, $F \circ g^{-1}$ induces a a distribution on $\mathcal{M}$.

The distributions on A will be distinguished according to the election of the parameters. If each parameter is chosen independently of others, then we will say that we are in the independent case, otherwise we will say that we are in the dependent case.

3.2.1 Independent Case

In this case, a distribution function is specified for each parameter and the sample is generated according to the product measure. In the first example, each parameter is chosen from the uniform distribution. This method is very simple and natural.

Uniform distribution on the parameters

A simple and natural method for simulating points on a parametrized manifold is to simulate a random vector $X = (x_1, \ldots, x_k)$ with uniform distribution on A, the domain of g, and map the points to $\mathcal{M}$. In many cases, $A = \prod_{i=1}^{k} [a_i, b_i]$ and so choosing a point in A according to the uniform distribution on A, is equivalent to choosing a point, for each i, uniformly and

independently on $[a_i, b_i]$. For more general domains, the algorithm of acceptance-rejection can be used.

One consequence of the area formula is that the density of the uniform distribution gives a density proportional to $1/J_k f$ on the manifold. Hence, the measure of a region on $\mathcal{M}$ will be inversely proportional to the Jacobian.

The method described in this section is used in many applications to simulate points on manifolds; expecting that for a sufficiently large sample, the manifold will be covered uniformly. But, as will be seen later, this method does not always give these results. In fact, this method gives these results if and only if the Jacobian of the parametrization is constant.

Distribution with Other Margin Distributions

A more general simulation scenario can be established as follows. Given $F_1, \ldots, F_k$ real distribution functions, a point X_i is simulated according to F_i for each i and the vector $(X_1, \ldots, X_k)$ is mapped to $\mathcal{M}$. This method can be implemented in a simple way and only depends on the ability to simulate points according to a real distribution. This method is not as flexible as may be desired, but some regions of high concentration of points can be of interest.

This simulation can lead to examples of manifolds that are covered "slowly". In other words, there are examples where it is necessary to simulate a large number of points to obtain a point in some specific regions. Taking as margins the semicircle distribution, gives one such example.

3.2.2 Dependent Case

The first method in this section is based on Girko's circular law. For the second method, the theory of copulas is used. This theory enables to simulate random vectors with some kind of correlation and specified margins.

Distribution Induced by the Circular Law

We obtain a sample in the parameters that is not induced by a product measure; now the parameters are correlated to each other. In this section, Girko's circular law theorem is used. This theorem, roughly speaking, states that the eigenvalues of a $n \times n$ random matrix with i.i.d. entries, mean 0 and variance $1/n$ have as limiting distribution the uniform distribution on the complex disc. The points obtained are not independent and some 'repulsion' can be observed between them. Moreover, due to this repulsion, the points mapped to the manifolds exemplified are 'nearer' to the $\mathcal{H}^k$-uniform distributed points. In the following paragraphs, we give a brief exposition of the circular law, which is based on [7]. After presenting the main result on the circular law, a method for inducing a sample in the parameters is shown.

For a matrix $A \in \mathbb{M}_n(\mathbb{C})$, the empirical measure of its eigenvalues is defined as

$$\mu_A = \sum_{k=1}^{n} \frac{1}{n} \delta_{\lambda_k(A)},$$

where $\lambda_k(A)$ denotes the k-th eigenvalue ordered decreasingly by their norm.

For a random complex variable Z, the variance of Z is $Var(Z) = \mathbb{E}(|Z|^2) - |\mathbb{E}(Z)|^2$. Given a random complex variable Z with variance 1 and mean 0, for each n let $M \in \mathcal{M}_n(\mathbb{C})$ the random matrix such that $1 \leq i, j \leq n$, M_{ij} are i.i.d. random variables distributed as Z.

We are now ready to state Girko's circular law. In the theorem, the convergence is the weak convergence of probability measures with respect to bounded continuous functions.

Theorem 3.1. *$\mu_{n^{-1/2}M} \to \mathcal{C}_1$ as $n \to \infty$, where $\mathcal{C}_1$ is the uniform law on the unitary complex disc $\mathbb{C}$, with density*

$$z \mapsto \frac{1}{\pi} \mathbf{1}_{\{z \in \mathbb{C} || z| \leq 1\}}. \tag{3.3}$$

So, we obtain the following algorithm

- Simulate an $n \times n$ matrix M with i.i.d. entries $(2n)^{-1/2}(N_1 + iN_2)$, being N_1, N_2 i.i.d. random normal variables.
- Compute the eigenvalues of M.

- Accept the points in $\left[-2^{-1/2}, 2^{-1/2}\right] \times \left[-2^{-1/2}, 2^{-1/2}\right]$.
- Map the points linearly to $[0, 2\pi] \times [0, 2\pi]$.

Because the area of the square taken in consideration is 2, we obtain that the number of sampled points is approximately the $200/\pi \approx 63\%$ of the total number of points generated by taking the eigenvalues. Hence, if the goal is to obtain n points, then the generation of $n\pi/2$ points is necessary.

The following example shows the difference between two point clouds consisting on 200 points: one of them was simulated according to the uniform distribution on $[0, 2\pi]$ and the other one according to the method presented previously. Due to repulsion between the eigenvalues of the simulated matrix, repulsion between the points is observed.

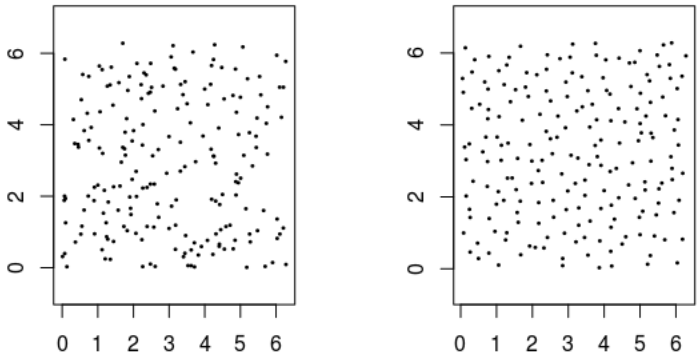


Figure 3.3: Comparison between sample from uniform distribution (left) and sample with repulsion between points (right)

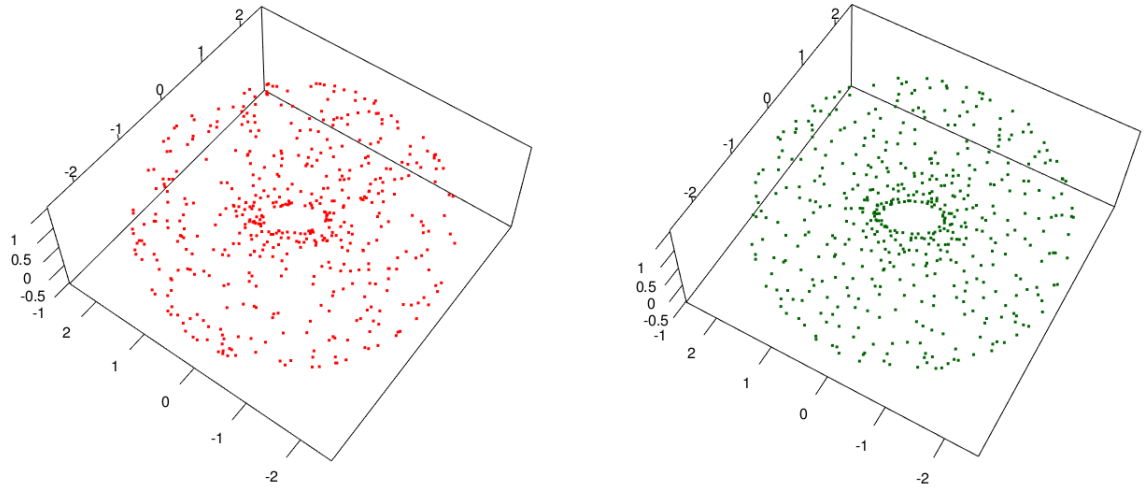


Figure 3.4: 500 points sampled from the uniform distribution on $[0, 2\pi]^2$ and mapped to the torus (left) vs. 500 points sampled with repulsion and mapped to the torus.

Distributions Obtained Using the Theory of Copulas

As before, the real distribution functions $F_1, \ldots, F_k$ are given. The goal is to simulate point clouds where the components are correlated and the i-th component has F_i as distribution. In the following paragraphs, some notions of this theory are discussed to use them for our purpose. An extensive treatment of this topic can be seen in [20].

Definition 3.1. Let $k \geq 2$. A k-dimensional copula is a k-multivariate distribution on $[0, 1]^k$ such that their univariate margins are uniformly distributed on $[0, 1]$.

The following theorems show how to simulate points with some multivariate distribution with margins specified, having certain dependence between them.

Theorem 3.2 (Sklar)**.** *Let F be a k-multivariate distribution function with univariate margins $F_1, \ldots, F_k$. For each j, let A_j the range of F_j, i.e. $A_j = F_j(\mathbb{R})$. Then there exists a copula C such that for all $(x_1, \ldots, x_k) \in \mathbb{R}^k$,*

$$F(x_1, \ldots, x_k) = C(F(x_1), \ldots, F_k(x_k)). \tag{3.4}$$

Such C is uniquely determined on $A_1 \times \cdots \times A_k$; thus, when $F_1, \ldots, F_k$ are continuous, C is unique.

Theorem 3.3. *If $F_1, \ldots, F_k$ are univariate distribution functions and C is any k-copula, then the function $F : \mathbb{R}^d \to [0, 1]$ defined by 3.4 is a k-multivariate distribution function with margins $F_1, \ldots, F_k$.*

From the previous result, if F is a given k-variate distribution function with continuous margins, then a copula C can be obtained by

$$C(u_1, u_2, \ldots, u_k) = F(F_1^{-1}(u_1), \ldots, F_k^{-1}(u_d)).$$

Then, we have the following algorithm to simulate $(U_1, \ldots, U_k)$ with distribution C:

- Sample $(X_1, \ldots, X_k)$ from F.
- Evaluate each component in its corresponding margin:

$$(U_1, \ldots, U_k) = (F_1(X_1), \ldots, F_k(X_k)).$$

We illustrate this with a particular example of a Gaussian copula, and we use it to sample on a manifold.

Example 3.2. Let Σ be a $k \times k$ symmetric and positive-definite matrix with real entries, such that $\Sigma_{ii} = 1$ for $i = 1, \ldots, k$. Let $\Sigma = LL^T$ be its Cholesky decomposition (where L is a lower triangular matrix).

Note that if $Z = (Z_1, \ldots, Z_k) \sim \mathcal{N}_k(0, I_k)$ (which we know how to simulate), then $X = LZ \sim \mathcal{N}_k(0, LL^T) \sim \mathcal{N}_k(0, \Sigma)$.

As $\Sigma_{ii} = 1$ for $i = 1, \ldots, k$, we have that $X_1, \ldots, X_k$ are all marginally distributed as Gaussian standard random variables, and $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$. If Φ is the univariate standard Gaussian distribution function, then from the previous results we have that

$$(U_1, \ldots, U_k) = (\Phi(X_1), \ldots, \Phi(X_k))$$

such that each U_i has the uniform distribution on $[0, 1]$ as its margin distribution.

If we take $k = 2$, then for this example Σ has the form

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1, \end{pmatrix}$$

with $\rho \in (-1, 1)$ (if $\rho \notin (-1, 1)$ Σ is not positive definite). By varying ρ we control the dependence between X_1 and X_2; if $\rho = 0$ they are independent.

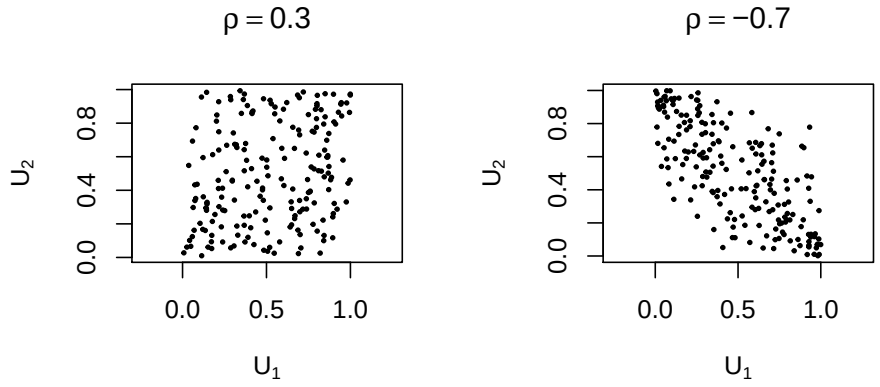
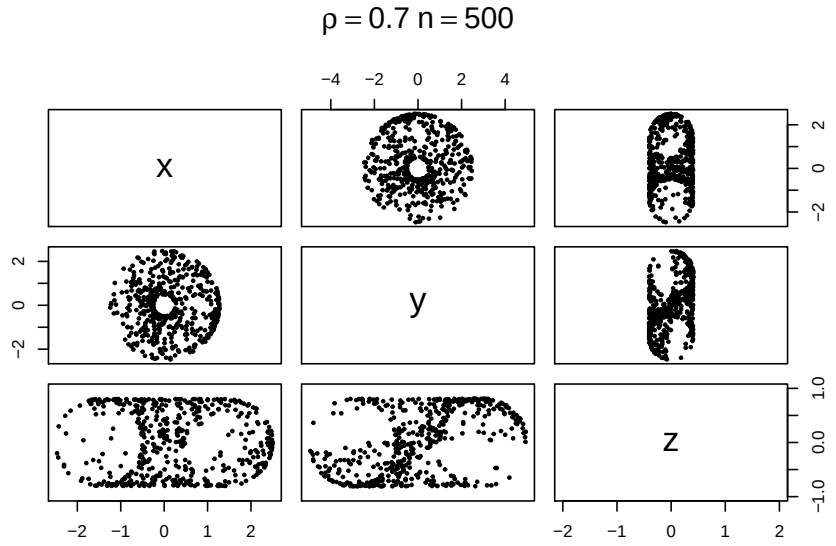


Figure 3.5: samples of size $n = 200$.

By rescaling the sample adequately, we can get a sample on a parametrized manifold, such as the torus:



3.3 Simulation on Stratified Spaces

A stratified space is essentially piece-wise a manifold, so the techniques used for simulation until now can be used for stratified spaces. In this section, we look into cases worth mentioning.

3.3.1 Uniform Distribution on a Stratified Space

Let $X \subset \mathbb{R}^d$ be a k-dimensional manifold stratified space ($k \leq d$), with filtration

$$X = X^k \supset \ldots \supset X^0 \supset X^{-1} = \emptyset.$$

We assume that $0 < \mathcal{H}^k(X) < \infty$, and that X^k has a finite number of components. Let $\Gamma_1, \ldots, \Gamma_l$ be the connected components of X_k. Because by definition X_k is a k-manifold, so are $\Gamma_1, \ldots, \Gamma_l$. We can then apply the previous techniques to each Γ_i. Suppose we are able to sample from a random variable Y_i with uniform distribution on Γ_i, for $i = 1, \ldots, l$.

Consider the following procedure. Let $v_i = \mathcal{H}^k(\Gamma_i)$ the k-dimensional Hausdorff measure of each Γ_i. Note that $\mathcal{H}^k(X) = \sum_{i=1}^{l} v_i$. Sample an index α from $\{1, \ldots, l\}$, with probabilities for each index j given by $v_j/\mathcal{H}^k(X)$. Next, make $W = Y_\alpha$.

Because X^{k-1} is a $(k-1)$-manifold, it has $\mathcal{H}^k$-measure equal to zero, so

$$\mathcal{H}^k(X) = \mathcal{H}^k(X \setminus X^{k-1}) + \mathcal{H}^k(X^{k-1}) = \mathcal{H}^k(X_k).$$

Then, by the total probability law we have

$$\begin{aligned}
\mathbb{P}[W \in A] = \mathbb{P}[W \in A \cap X_k] &= \sum_{i=1}^{l} \mathbb{P}[W \in A \cap X_k | \alpha = i]\mathbb{P}[\alpha = i] \\
&= \sum_{i=1}^{l} \mathbb{P}[Y_i \in A \cap \Gamma_i] \frac{v_i}{\mathcal{H}^k(X)} \\
&= \sum_{i=1}^{l} \frac{\mathcal{H}^k(\Gamma_i \cap A)}{H^k(\Gamma_i)} \frac{\mathcal{H}^k(\Gamma_i)}{\mathcal{H}^k(X)} \\
&= \sum_{i=1}^{l} \frac{\mathcal{H}^k(\Gamma_i \cap A)}{\mathcal{H}^k(X)} = \frac{\mathcal{H}^k(A \cap X_k)}{\mathcal{H}^k(X)} = \frac{\mathcal{H}^k(A)}{\mathcal{H}^k(X)}.
\end{aligned}$$

Thus, the uniform distribution over X is H^k-uniform.

It is worth noting that $\{\Gamma_i\}$ does not to be the partition on connected components of X_k.

For instance, in the example 2.3, one could take Γ_1 equal to the sphere and Γ_2 equal to the disc through the equator. In this case Γ_1 is the union of two connected components of X_2 and a subset of X_1, which has $\mathcal{H}^2$ measure equal to 0.

Remark 2. The support of the probability measure constructed here is not necessarily equal to X: the curve that pinches the sphere and the disc is not contained in the support.

Example 3.3. A simple and illustrative example is the polar rose, a parametric curve defined by

$$x(\theta) = \cos(k\theta)\cos(\theta), \tag{3.5}$$

$$y(\theta) = \cos(k\theta)\sin(\theta). \tag{3.6}$$

This is a parametric curve for every real $k > 0$, but we will consider only integer values for k: it has $2k$ 'petals' when k is even, and k 'petals' when k is odd.

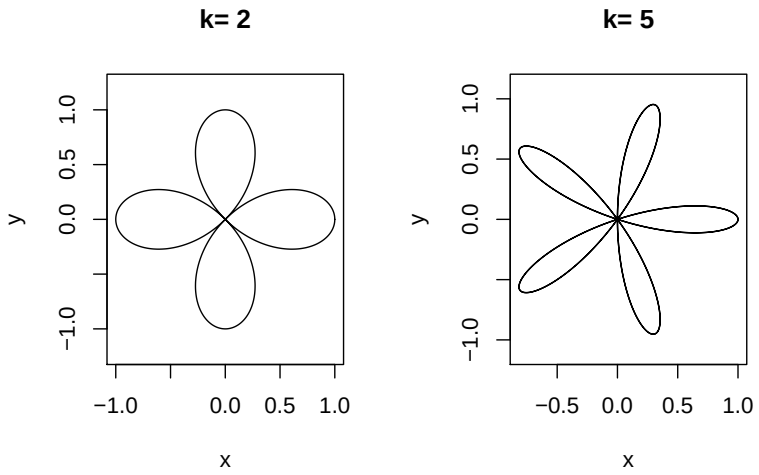


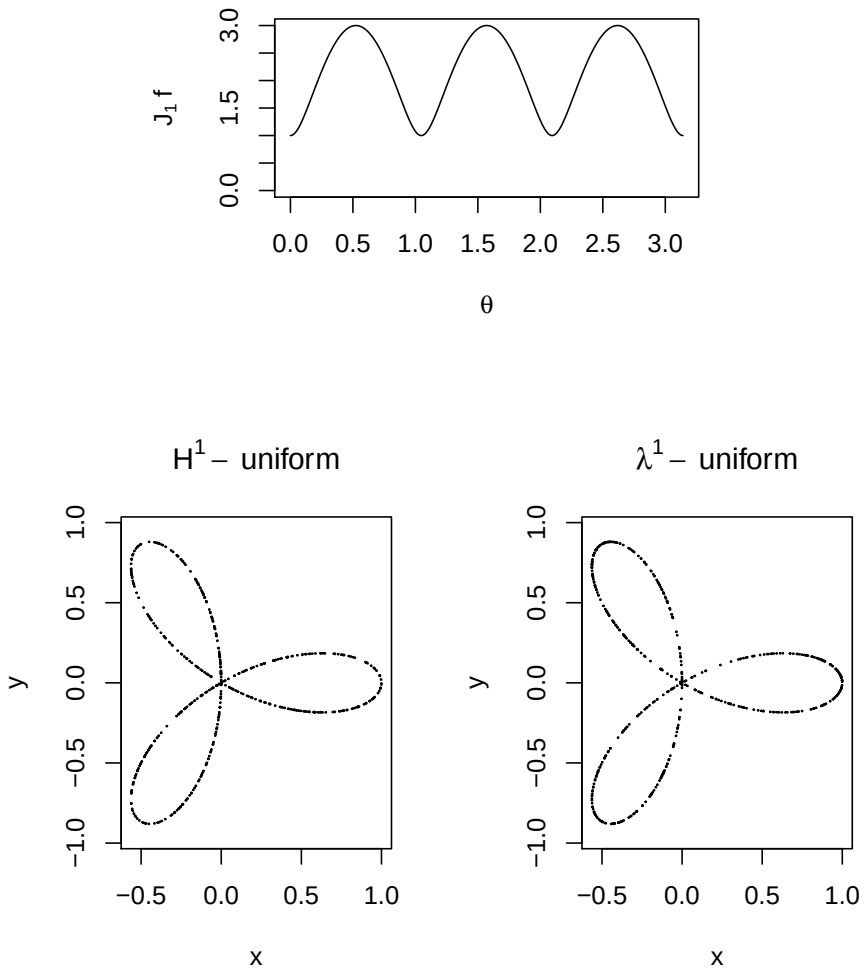
Figure 3.6: examples of the polar rose for $k = 2$ (4 petals) and $k = 5$ (5 petals).

Note that if k is even, the entire graph is obtained as an image under (3.5) and (3.6) of any interval of length 2π, whereas if k is odd, then it is obtained with any interval of length π. In particular, we can consider $[0, \pi]$ as the domain of $x(\theta)$ and $y(\theta)$ for k odd, and $[0, 2\pi]$ for k even.

If $X \subset \mathbb{R}^2$ is the entire graph, then it is easily verified that with $X^0 = \{(0, 0)\}$, $X^1 = X$, then the polar rose is a 1-dimensional manifold stratified space because the only point that does not satisfies the definition of a 1-manifold is the origin.

The components of X_1 are the single petals. The length (this is, the $\mathcal{H}^1$ measure) of every petal is the same, and the preimage of a single petal is an interval of length π/k (for k even and odd).

Because every petal is a parametrized 1-manifold and the preimage of $\{(0, 0)\}$ is a set of Lebesgue measure 0, it is equivalent (for this case) to use the method described in section 3.1. For $k = 3$, the Jacobian has the following shape:



In the figure above, the plot on the left shows an example of a sample of size 500 using this method. The plot on the right shows an example of 500 points sampled from the uniform distribution on $[0, \pi]$ mapped using the parametrization: note that the outside of the petals have more concentration of points than the inside.

3.3.2 Special Cases: $\bar{X}_k \subsetneq X$

As said in the previous section, example 2.3 presents some issues for simulating on all the space. Suppose that we have a probability measure on a k-dimensional stratified space X that is absolutely continuous with respect to $\mathcal{H}^k$, then X_i for $i \leq k$ are null sets for that

probability measure. Therefore, if the closure of X_k is not equal to $\mathbb{X}$, then other methods will be required to sample all of the space.

Both of the models presented in this section are briefly explained in [3].

Simulation by Thickening X

The first method proposed is to sample on a *thickened* version of X. For a fixed $\varepsilon > 0$, sample from the set

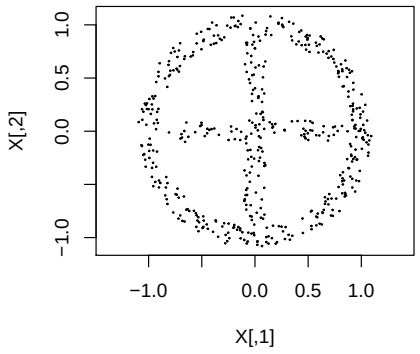
$$X_\varepsilon = \{x \in \mathbb{R}^d : \rho(x, X) < \varepsilon\},$$

where $\mathbb{R}^d$ is the ambient space and ρ is the Euclidean distance on $\mathbb{R}^d$.

Although one does not properly sample from the space X, this method is relevant in TDA-related contexts because it may be seen as sampling with noise.

In the general case, one can sample using the acceptance-rejection method. However, in that case we must be able to calculate the distance to X, which is not always easy to do. One strategy is to calculate the distance to every maximal stratum and then take the minimum of those distances. One difficulty is that even for sets in low dimensions, numerical methods could be needed for calculating $\rho(x, X)$. Another difficulty that can arise is that there could be a high number of rejections if ε is too small.

Example 3.4. Consider the subspace $X \subset \mathbb{R}^2$ consisting of the unitary circle along with two perpendicular diameters. Using the acceptance-rejection method, we simulate 500 samples from X thickened by $\varepsilon = 0.1$. To apply the acceptance-rejection method, we sample from the uniform distribution on $[-1-\varepsilon, 1+\varepsilon]^2$, and reject a point if it is at a distance greater than or equal to 0.1. For the example shown below, a sample of size 1238 was needed for obtaining 500 acceptances.



Mixture Model

We say that a stratum T of a manifold stratified space is *maximal* if it is not a subset of the closure of any other stratum. In this method, we focus on sampling from the maximal strata. In this case, $\bar{X}_k$ is a proper subset of X, so there are maximal strata of lower dimension.

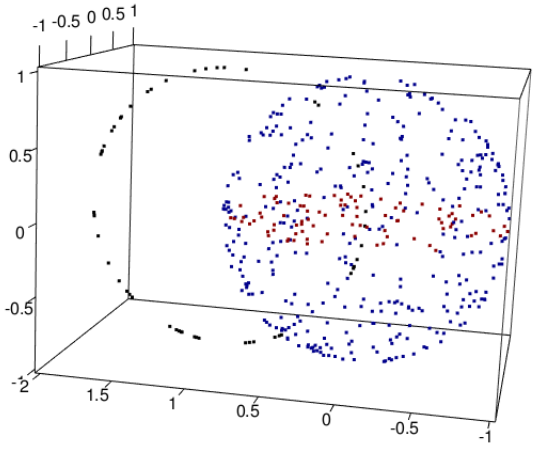
Let S_i be the collection of maximal strata of dimension i, for $i \in \{0, \ldots, k\}$. The strategy is as follows. Given non-negative weights $a_0, \ldots, a_k$ such that $\sum a_i = 1$ and probability measures $\mathbb{P}_i$ on S_i, sample an index t with probability $\mathbb{P}[t = i] = a_i$. Next, sample W from the probability $\mathbb{P}_i$.

Remarks.

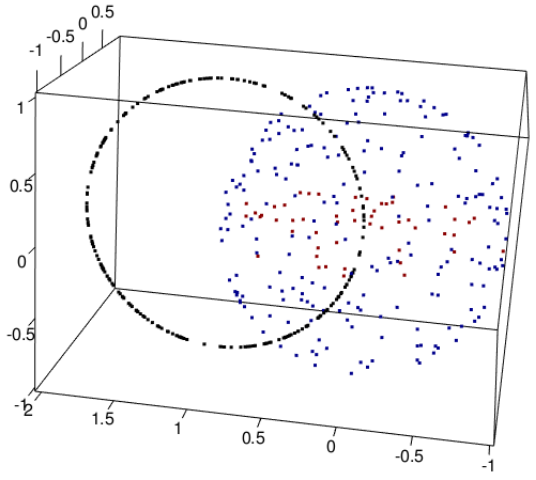
- If $S_i = \emptyset$ for some i, set $a_i = 0$ and leave $\mathbb{P}_i$ undefined because it makes no sense to define a probability measure on the empty set.
- S_i is a collection of i-manifolds, so $\mathbb{P}_i$ can be a probability measure corresponding to one of the methods presented in Sections 4.1 and 4.2.
- Depending on the purpose of the simulation, we can also include non-maximal strata in the mixture model.

Example 3.5. Consider the space from example 2.3. Using the same notation, the maximal strata are the sphere minus the equator (Γ_1), the disc (Γ_2) and the curve α. For this case, we consider $\mathbb{P}_1$ the uniform measure on α and $\mathbb{P}_2$ the uniform measure on $X_2 = \Gamma_1 \cup \Gamma_2$.

Sampling from $\mathbb{P}_1$ can be achieved using the method described in Section 3.3.1. By setting a_1, the weigh for $\mathbb{P}_1$, equal to 0.1, an example of a point cloud of size 500 is



If the weigh for $\mathbb{P}_1$ is too high, then there will be "too many" points in α compared to the number of points in X_2 (where "too many" is in a sense not to be formally defined here). See for example, the next point cloud, which also has size 500.



Chapter 4

Examples with Calculation of Persistent Homology

In this chapter, we empirically analyze the persistent homology obtained from the Vietoris-Rips filtration on PCD using three spaces: the polar rose, the Klein bottle and example 2.3. The last example is relevant because, in contrast to the first two, it has maximal strata of a lower dimension and, therefore, the mixture model is used for simulation. For the polar rose and the Klein bottle, we analyze aspects on convergence when the size of the point cloud increases, aided by the stability theorem and a concentration inequality proposed by Fasy et al. (2014). In the third example we study how the persistent homology changes when varying the weighs in the mixture model. For this chapter, we assume the reader is familiar with basic notions of Topological Data Analysis, such as Vietoris-Rips filtration, persistence diagrams, Hausdorff distance and Bottleneck distance. Otherwise, in [12], for example, there is a quick review of concepts and some relevant results used here. In [11], there is a more systematic introduction to Topological Data Analysis.

Everything was coded in `R` and the persistence intervals were calculated using code provided by Francisco Valente [27].

4.1 Analysis of Convergence

4.1.1 Polar Rose

In this section, we analyze how the persistence barcode shows the three 1-cycles in the polar rose with three petals for different values of n, the sample size.

We use both the $\mathcal{H}^1$-uniform distribution on the rose and the λ^1-uniform distribution on $[0, \pi]$. The support of both distributions is the entire rose and persistence barcodes describe what is expected, although, as we will illustrate, there are some variations among the barcodes.

When $n = 100$, the simulations frequently show three long persistence intervals for 1-dimensional components, as the bottom-right-hand plot in the figure below shows. However, as the size is small, the case of the top-right plot occurs often; that is, it is not clear that there are three 1-dimensional cycles.

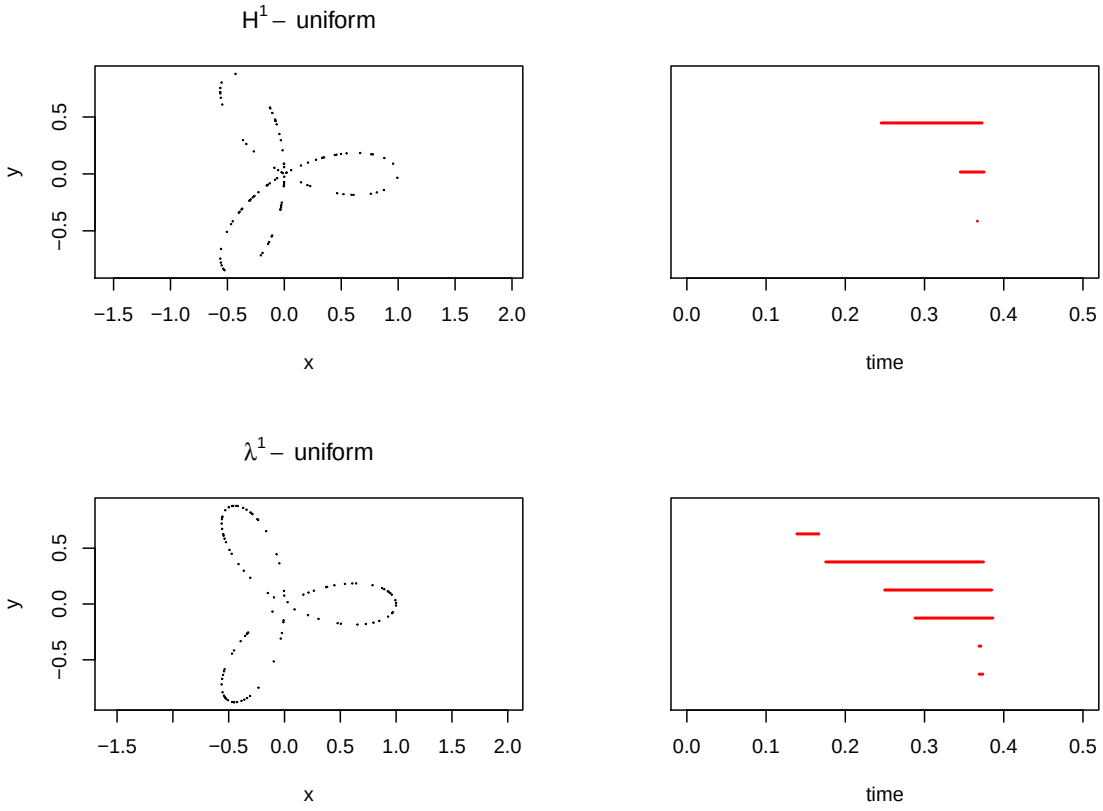
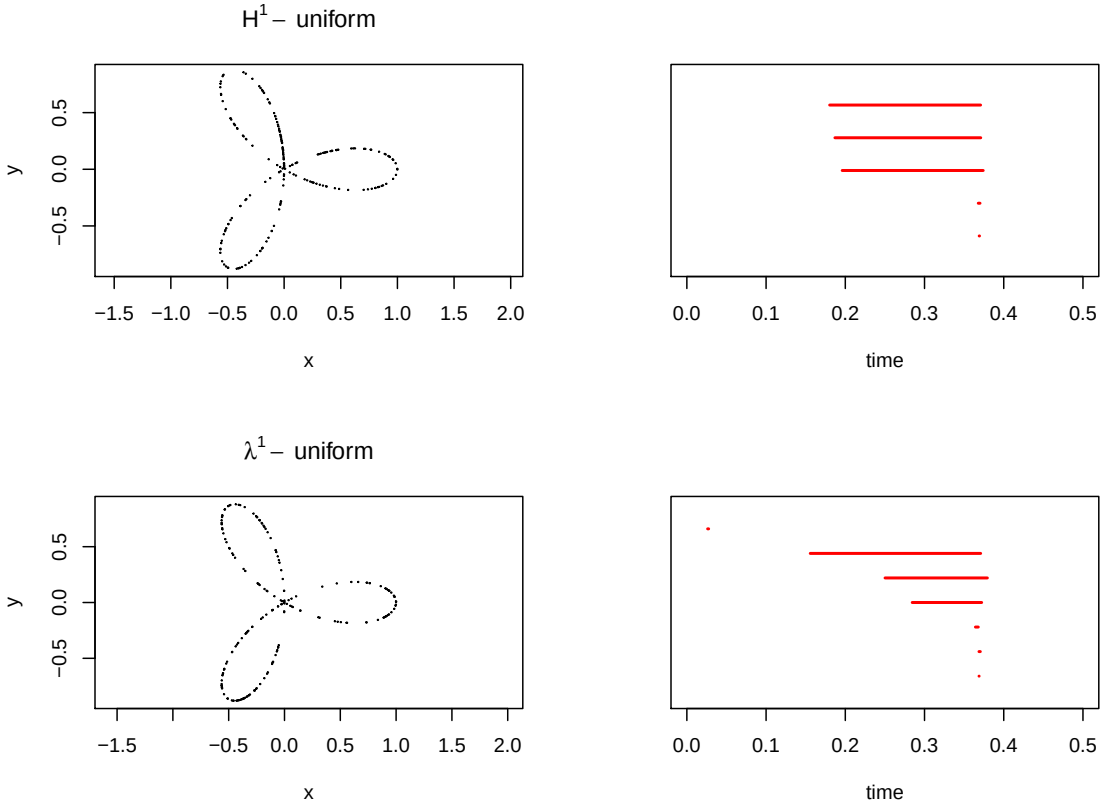


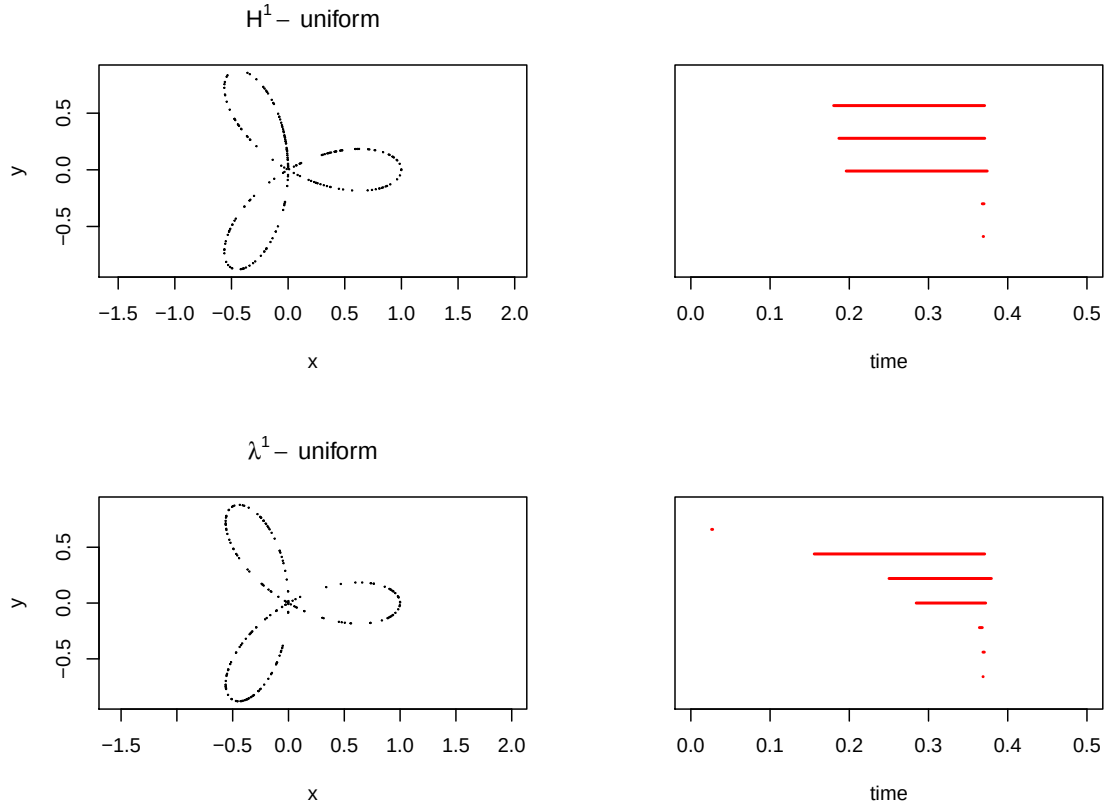
Figure 4.1: Samples of size 100 and their corresponding persistence barcodes.

In this particular example, the sample from the $\mathcal{H}^1$-uniform distribution but for low sample sizes this may happen with both distributions.

Given that what we are doing is of a random nature, for $n = 250$ we might also get not well-behaved case

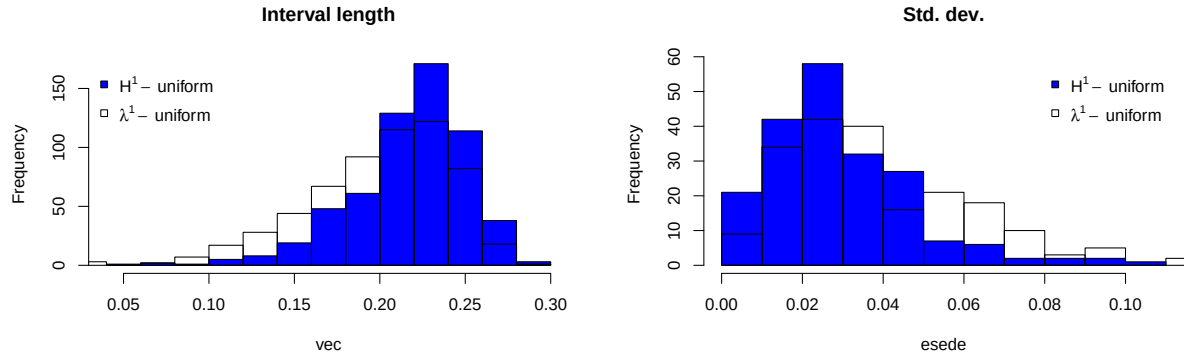


s as before, but less often. For this size, the following behavior is the most representative:



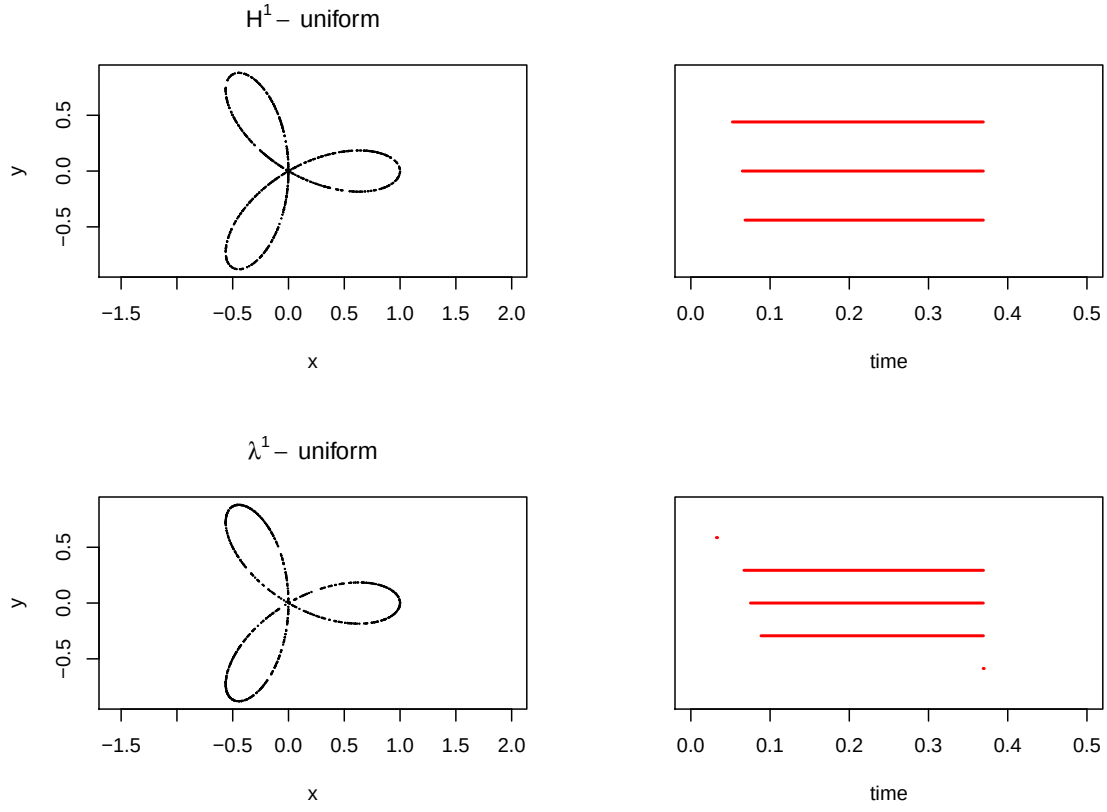
Because we get more "uniformly"-distributed samples in the geometrical sense with the $\mathcal{H}^1$-distribution, the three large 1-cycles have persistence intervals of similar length (top-right-hand plot), compared to the second sample, which presents more variation.

As empirical evidence of this behavior, for 200 samples of size 250 from both distributions, we stored the length of the three largest persistence intervals of 1-cycles; that is, we get 600 persistence intervals for each distribution. For each sample, we also store the standard deviation of the set of the three largest persistence interval lengths; that is, we get two sets of 200 values. The histograms of the corresponding values follow. They suggest that intervals obtained with the λ^1-uniform distribution are shorter and present larger variation.



It is important to mention that we do not intend to state that simulating with the $\mathcal{H}^1$-uniform distribution is "better" because the best method will depend solely on the purpose of the simulation. The goal of this section is to illustrate differences between two distributions with the same support from the point of view of persistent homology.

For large sample sizes, such as $n = 750$, there is little difference between them because the connected portions of the polar rose without a point are each time smaller with higher probability:



4.1.2 Klein Bottle

In this section, we consider the inmersion of the Klein bottle in $\mathbb{R}^3$, with the following parametrization found in [14]:

$$x = \begin{cases} 6\cos(u)(1+\sin(u)) + 4(1-\frac{1}{2}\cos(u))\cos(u)\cos(v) & \text{if } 0 \le u \le \pi, \\ 6\cos(u)(1+\sin(u)) + 4(1-\frac{1}{2}\cos(u))\cos(v+\pi) & \text{if } \pi < u \le 2\pi, \end{cases}$$

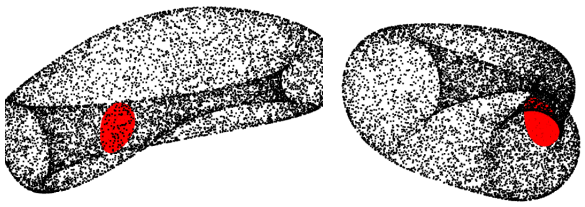
$$y = \begin{cases} 16\sin(u) + 4(1-\frac{1}{2}\cos(u))\sin(u)\cos(v) & \text{if } 0 \le u \le \pi, \\ 16\sin(u) & \text{if } \pi < u \le 2\pi, \end{cases}$$

$$z = 4(1-\frac{1}{2}\cos(u))\sin(v),$$

for $(u, v) \in [0, 2\pi]^2 \setminus R$, with

$$R := \left\{ (u,v) : \left(\frac{u-3.66}{0.19}\right)^2 + \left(\frac{v-\pi}{0.38}\right)^2 < 1 \right\}.$$

R is an approximation to the preimage under the parametrization of the region, as shown in red.



This embedding of the Klein bottle is a 2-dimensional manifold stratified space, with empty 0-dimensional stratum and a 1-dimensional stratum equal to the curve where it intersects itself.

We now make an empirical analysis of convergence based on the results given by Fasy et al. [12].

For a closed subset A of $\mathbb{R}^n$, we define the distance function to A as $d_A(x) = \inf_{y \in A} \|x - y\|$. Recall the definition of the Hausdorff distance:

Definition 4.1. The Hausdorff distance $H(K, K')$ between to closed subsets K, K' of $\mathbb{R}^n$ is defined as

$$H(K, K') = \max\left(\sup_{y\in K'}(\inf_{x\in K}||x-y||), \sup_{x\in K}(\inf_{y\in K'}||x-y||)\right).$$

For a function f, we denote (when it is defined) by $\mathrm{dgm}(f)$ the persistence diagram obtained from the filtration by sublevel sets of f, $\{f^{-1}((-\infty, \alpha])\}$, and by W_∞ the bottleneck distance between diagrams.

Theorem 4.1 (Stability theorem)**.** *Let X be a topological space homeomorphic to a finite simplicial complex and $f, g : X \to \mathbb{R}$ be continuous functions. Then*

$$W_\infty(\mathrm{dgm}(f), \mathrm{dgm}(g)) \leq \|f - g\|_\infty,$$

The Hausdorff distance between two closed sets is related to the induced distance function as follows:

$$H(A, B) = ||d_A - d_B||_\infty.$$

Thus, if $S_n = \{X_1, \ldots, X_n\}$ is a sample from a probability distribution $\mathbb{P}$ with support on a manifold $\mathbb{M}$, then from the stability theorem we have

$$W_\infty(\mathrm{dgm}(d_{S_n}), \mathrm{dgm}(d_{\mathbb{M}})) \leq H(S_n, \mathbb{M}).$$

We are interested on finding out how $\mathbb{P}(H(S_n, \mathbb{M}) > t)$ behaves with respect to n and $t > 0$. The following functions are defined in [12]:

$$\rho(x, t) = \frac{P\left(B_x\left(t/2\right)\right)}{t^d}, \qquad \rho(t) = \inf_{x\in\mathbb{M}} \rho(x, t). \tag{4.1}$$

The motivation for defining those functions is to quantify how small the probability of a radius t on $\mathbb{M}$ can be.

A useful result provided in [12] follows:

Theorem 4.2. *For every $t > 0$*

$$\mathbb{P}\left(W_\infty(\mathrm{dgm}(d_{S_n}), \mathrm{dgm}(d_{\mathbb{M}})) > t\right) \leq \mathbb{P}(H(S_n, \mathbb{M}) > t) \leq \frac{2^d}{\rho\left(t/2\right) t^d} \exp\left(-n\rho(t)t^d\right).$$

From theorem 4.2, we are able to give a lower bound on the number of points needed for $\mathbb{P}(H(S_n, m) > t) \leq \alpha$, with $\alpha > 0$:

$$n = \frac{\log\left(\frac{2^d}{\rho(t/2)t^d\alpha}\right)}{\rho(t)t^d}.$$

If ρ has a small value, then the number of points must be large enough to accomplish the desired inequality.

The previous result is proven when $\mathbb{M}$ is a manifold but we aim to provide empirical insight of this result applied when $\mathbb{M}$ is a stratified space, such as the Klein bottle. ρ can be hard to calculate explicitly, so we will compare distributions by comparing the Hausdorff distance between them. This is justified by the fact that the Hausdorff distance between closed subsets satisfies the triangle inequality, so the Hausdorff distance between samples is bounded above by the sum of their distances to $\mathbb{M}$.

For the $\mathcal{H}^2$-uniform distribution on a manifold $\mathbb{M}$, we expect ρ to be bounded away from 0, so as n increases the Hausdorff distance between the sample and $\mathbb{M}$ effectively converges in probability to 0 (from theorem 4.2). As mentioned, in this case $\mathbb{M}$ is not a manifold but the whole space is the closure of the 2-dimensional stratum, so there is also convergence to 0. Note that this does not necessarily happen if $\mathbb{M}$ is a stratified manifold space with singular maximal strata, such as in example 2.3.

For the first example, we calculate $H(S_n, S'_n)$, where S_n is a sample of size n of the $\mathcal{H}^2$-uniform distribution on $\mathbb{M}$, and S'_n is a sample of size n obtained from the λ^2-uniform distribution on the domain of the Klein bottle parametrization.

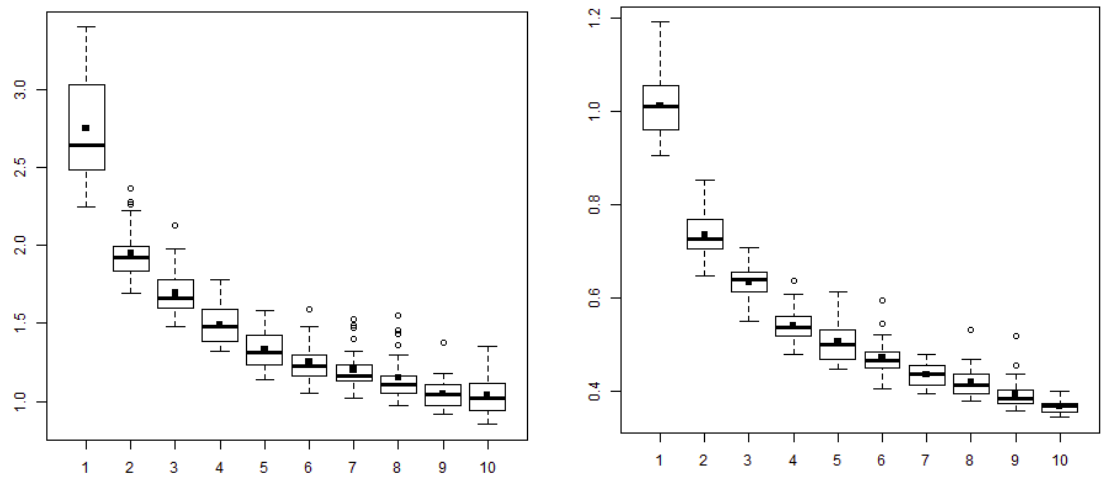


Figure 4.2: Boxplots of Hausdorff distance between samples from $\mathcal{H}^1$-uniform distribution and distribution induced from the λ^2-uniform on the domain. The sample sizes on the left plot are $n \times 1,000$; on the right, $n \times 10,000$. There is empirical convergence.

If we add Gaussian noise to the sample, then this convergence is no longer observed because the support of the distribution changes.

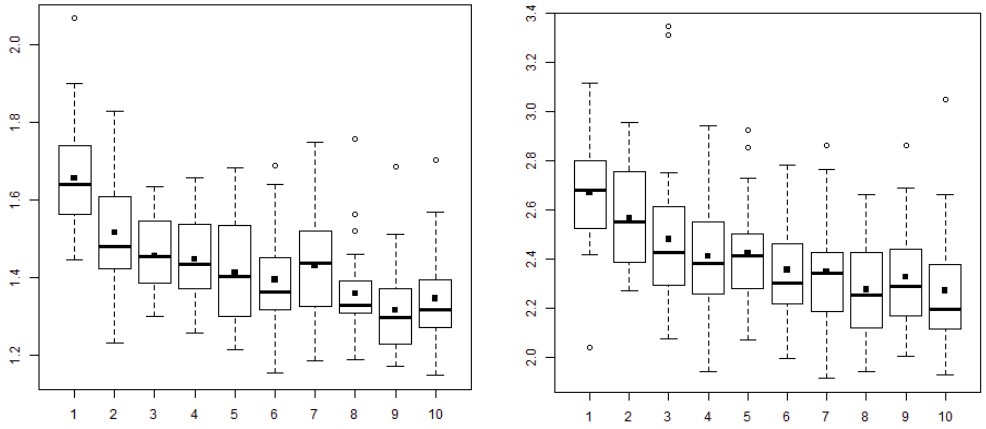


Figure 4.3: Hausdorff distance for samples with Gaussian noise with $\sigma^2 = 0.5$ for the left plot, and $\sigma^2 = 1$ for the right plot. Convergence is not clear. Sample size is $n \times 10,000$.

We now present cases where ρ is expected to have values close to 0.

For the next example we take one sample again with the $\mathcal{H}^2$-uniform measure on $\mathbb{M}$. The second sample is to be taken from the product distribution on $[0, 2\pi]^2$, with uniform

distribution on one parameter and the beta distribution $B(\alpha = 5, \beta = 3)$ on the other. It is not clear if it converges; if it does, it is slow.

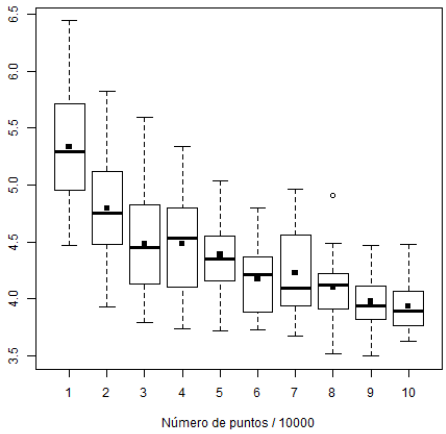


Figure 4.4: Sample size $= n \times 10,000$.

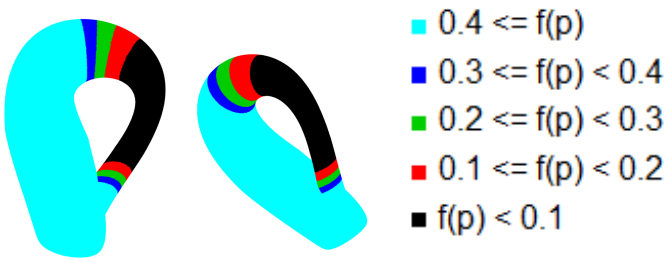


Figure 4.5: The colors in the Klein bottle correspond to the value of the density in the preimage of those regions: blue regions have higher values and black ones have lower values. The black region has a low probability measure, which might explain why the convergence is slow (if it does at all).

We now sample both parameters independently from $B(\alpha = 1.5, \beta = 1.5)$ (semicircle distribution) for the second point cloud. The plot suggests there is a convergence.

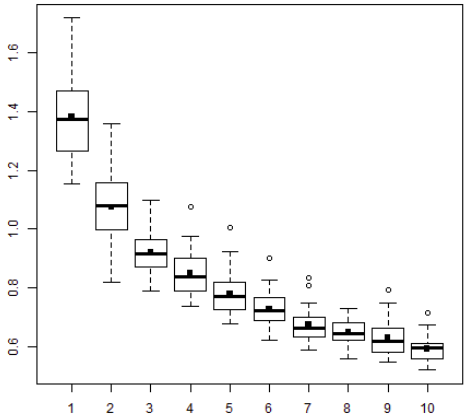


Figure 4.6: Sample size $= n \times 10,000$.

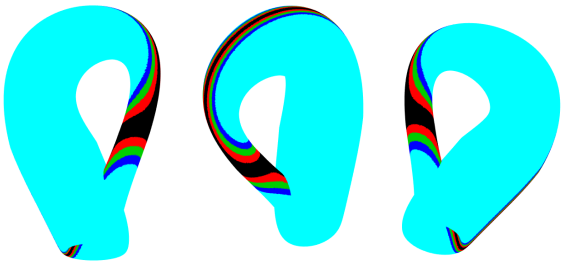
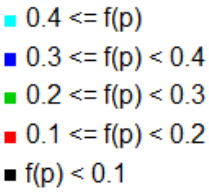


Figure 4.7: Although there is a region with lower probability (the black portion), the Hausdorff distance measurements suggests there is a convergence to 0.



In the next example, the second point cloud is simulated by sampling each parameter from the arcsine distribution ($B(0.5, 0.5)$). The plot suggests convergence to 0, which is expected from the fact that in this case the density on the parameters is bounded away from 0.

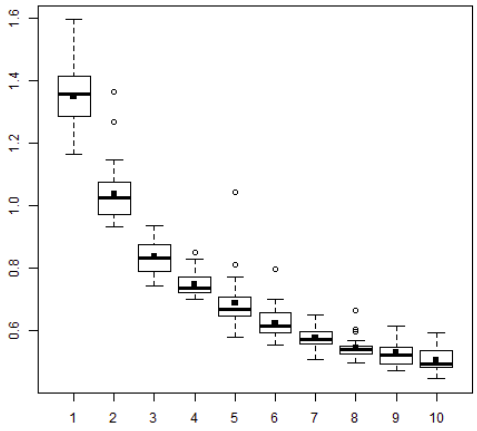


Figure 4.8: Sample size $= n \times 10,000$.

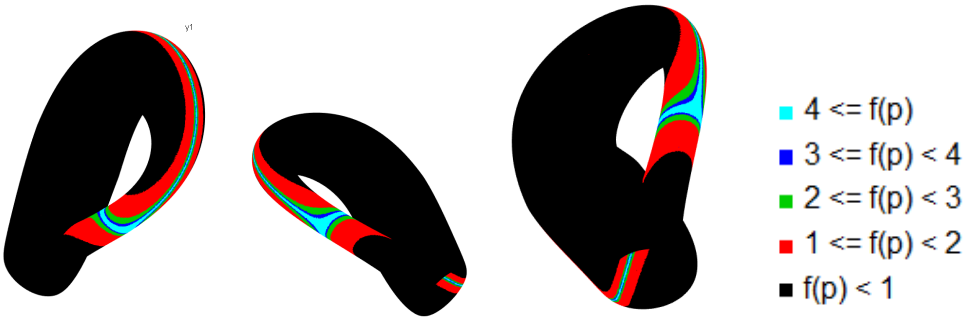


Figure 4.9: Even though the density on the parameters take low values on the preimage of the black region, the bottle is satisfactorily covered with the sample.

4.2 Change of Persistent Homology in Mixture Models

In this last example, we illustrate how the homology changes when varying the weighs in a mixture model described in Section 3.3.2. With the same notation as in example 3.5, we take a sample of size 700 and vary the weight a_1.

Taking $a_1 = 0.1$ we have a (visually) "balanced" sample; that is, there is a reasonable number of points in the curve α, and a reasonable amount of points in $X_2 = \Gamma_1 \cup \Gamma_2$.

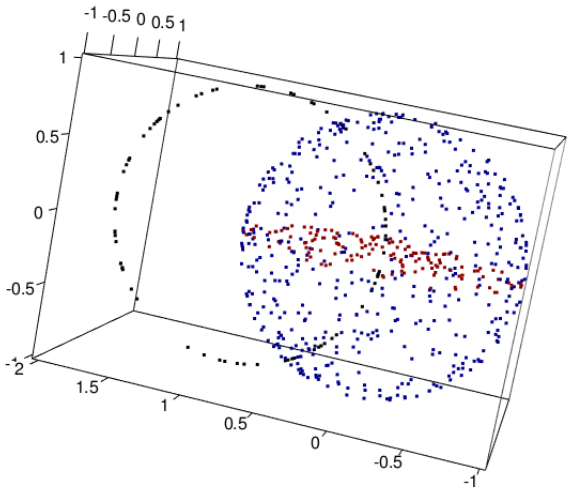
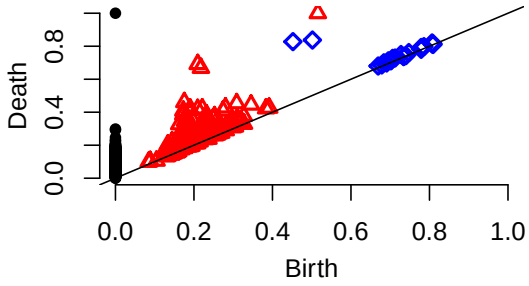


Figure 4.10: Sample of size 700 for $\mathbb{P}[X \in \alpha] = 0.1$.

This balance between the number of points in a singular stratum is also reflected in the persistence diagram, which detects two 2-cycles, three 1-cycles and one connected component that has a larger lifetime than the other components of its kind.



If a_1 is much smaller, say, 0.01, then we get too few points on α and this is reflected in the 1-cycles detected. Because we still have a large sample on X_2, the expected 2-cycles are still present.

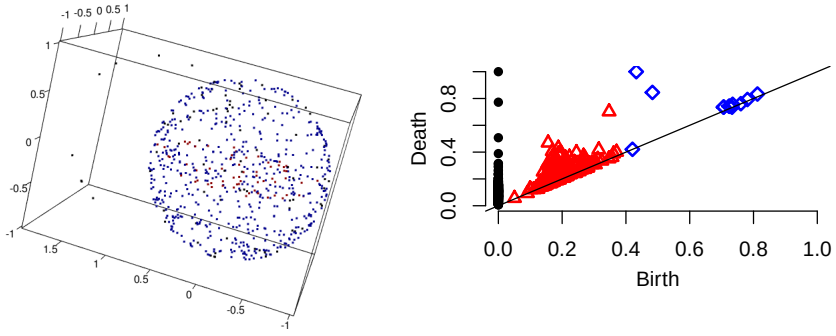


Figure 4.11: Sample and persistence diagram for $a_1 = 0.001$.

Meanwhile, if we increase the value of a_1, then the sample quickly tends to be concentrated on the curve α. For example, take $a_1 = 0.5$. We notice that the lifetime of one 1-cycle increases because it is born earlier than before, whereas the lifetime of the 2-cycles decreases as they are born later.

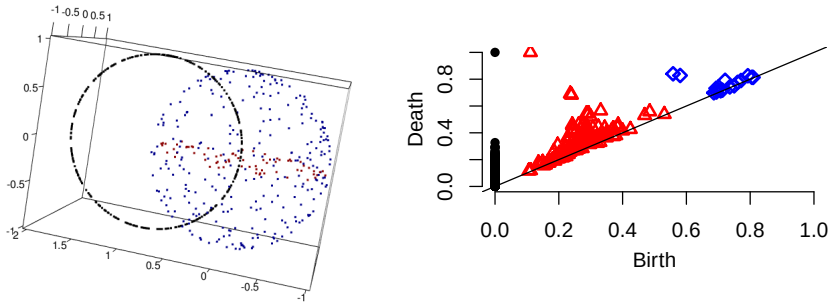


Figure 4.12: Sample and persistence diagram for $a_1 = 0.5$.

This behavior is more dramatic when a_1 is even larger; for example, $a_1 = 0.8$

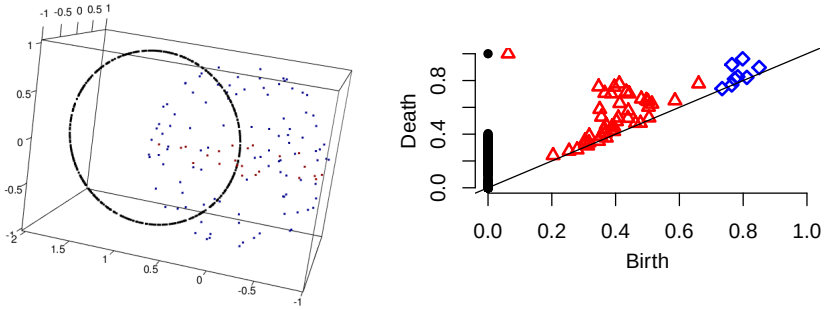


Figure 4.13: Sample and persistence diagram for $a_1 = 0.8$.

Appendix A

Acceptance-Rejection Method

In this appendix, we present relevant aspects of the acceptance-rejection method for sampling. First, Section A.1 describes the method in its general version, as presented by [9]. In the case of sampling from manifolds, there are some relevant considerations that we mention in Section A.2.

A.1 The General Case

The acceptance-rejection method provides a method to sample from a given probability distribution using samples from another distribution. This method is commonly used in contexts different to sampling from manifolds.

In this section, we will describe the mathematical aspects of this method when both probability distributions have density function (this method can also be used with discrete probability distributions).

We wish to sample from a probability distribution $\mathbb{P}_f$ on $\mathbb{R}^d$ that has density $f : \mathbb{R}^d \to \mathbb{R}$, and we have a generator of (pseudo)random numbers from the distribution $\mathbb{P}_g$ on $\mathbb{R}^d$ that has density $g : \mathbb{R}^d \to \mathbb{R}$. The densities f, g are such that there exists c with $cg(x) \geq f(x)$ for all $x \in \mathbb{R}^m$ (integrating both sides over $\mathbb{R}^m$ we can see that if there is such c, then $c \geq 1$).

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with distribution $\mathbb{P}_g$, and $U_1, U_2, \ldots$ a sequence of i.i.d. random variables with uniform distribution on $[0, 1]$, with both sequences independent from each other.

Let Y be the first X_i such that $U_i cg(X_i) \leq f(X_i)$, then Y has distribution $\mathbb{P}_f$. To prove this, the following theorems are used; they are stated and proved, for example, in Devroye [9].

Theorem A.1. *Let X be a random vector with density f and U be an independent random variable with uniform distribution on $[0,1]$.*

- *$(X, cUf(X))$ is λ^{m+1}-uniformly distributed on $A = \{(x,u) : x \in \mathbb{R}^d,\ 0 \leq u \leq cf(x)\}$, where $c > 0$ is an arbitrary constant.*
- *If (X,U) is a random vector uniformly distributed on A, then X has density f on $\mathbb{R}^d$.*

Theorem A.2. *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$ and $A \subset \mathbb{R}^d$ be a Borel set such that $\mathbb{P}(X_1 \in A) = p > 0$. Let Y be the first X_i in A. Then, Y has a distribution determined by*

$$\mathbb{P}(Y \in B) = \frac{\mathbb{P}(X_1 \in A \cap B)}{p}, \quad B \in \mathcal{B}(\mathbb{R}^d).$$

In particular, if X_1 is uniformly distributed in $A_0 \supset A$, then Y is uniformly distributed in A.

Now we shall see that Y is distributed as $\mathbb{P}_f$. By the first part of theorem A.1, $(X, cUg(X)) \in \mathbb{R}^{d+1}$ has uniform distribution on the area under cg. Then, by theorem A.2, $(Y, cUg(Y))$ has uniform distribution in the area under f, and by the second part of theorem A.1, Y has density f.

Therefore, with the next algorithm (Devroye [9]) we can simulate a random variable with distribution $\mathbb{P}_f$:

In practice, it is desirable to have a low number of rejections. Note that

$$\begin{aligned}\mathbb{P}(f(X) \geq cUg(X)) &= \int_{\mathbb{R}^d} \mathbb{P}\left(U \leq \frac{f(x)}{cg(x)}\right) dx \\ &= \int_{\mathbb{R}^d} \frac{f(x)}{cg(x)} dx = \frac{1}{c}\int_{\mathbb{R}^d} f(x)dx = \frac{1}{c}.\end{aligned}$$

If N is the number of iterations needed to get Y, then we have

$$\mathbb{P}(N = i) = (1 - 1/c)^{i-1}\, 1/c. \tag{A.1}$$

Algorithm 1 Acceptance-rejection

Require: $n \geq 0$, c such that $cg \geq f$

Ensure: Y pseudorandom distributed as $\mathbb{P}_f$

  repeat

    $X \leftarrow$ pseudorandom from distribution $\mathbb{P}_g$

    $U \leftarrow$ pseudorandom from unif distribution on $[0, 1]$

  until $cUg(X) \leq f(X)$

  $Y \leftarrow X$

Thus, the expected number of iterations is $1/(1/c) = c$, and lower values of c will yield faster algorithms.

Remark. In (A.1), N is a geometric random variable with success probability $1/c$. If the goal is to get a sample $Y_1, \ldots, Y_n$ of size n, then the number of iterations required will be $N_1 + \ldots + N_n$, where N_i is the number of iterations required for getting Y_i. As $N_1, \ldots, N_n$ are independent, $N_1 + \ldots + N_n$ is a negative binomial random variable, and its expected value is nc.

A.2 The Manifold Case

The densities obtained from the parametrization of a manifold can differ greatly from other known distributions, so it may be difficult to get samples from them by other methods. Therefore, we describe how to use a version of the acceptance-rejection method in the context of sampling from manifolds.

Let f, g be density functions such that we know how to sample from g, and we wish to sample from f. Suppose that there are constants $p, q > 0$ (not necessarily known) such that we know $pg(x)$ and $qf(x)$ for every $x \in \mathbb{R}^d$ (or in some subset of interest), and such that $pg(x) \geq qf(x)$ for all $x \in \mathbb{R}^d$.

By integrating both sides, we get $p > q$, so $c := p/q \geq 1$. Then,

$$\frac{p}{q} g(x) = cg(x) \geq f(x).$$

so we can use pg, qf for sampling with the acceptance-rejection method.

Let us recall what we have in our context. We wish to sample from the uniform distribution over a manifold $\mathcal{M}$ parametrized by ϕ with domain $A \subset \mathbb{R}^k$. We already know that the density from which we wish to sample is $f := J_k\phi / \operatorname{vol}(\mathcal{M})$. We (must) know how to calculate $J_k\phi(x) = \operatorname{vol}(\mathcal{M}) f(x)$; that is, we know how to calculate f times a (maybe unknown) constant. Let m be a constant such that $m \geq J_k\phi(x)$ for all $x \in A$. There is a constant p such that $m = pg(x)$, $x \in A$, where g is the density for the uniform distribution on A.

The following version of acceptance-rejection can then be used to simulate a random variable with density f:

Algorithm 2 Modified acceptance-rejection

Require: $n \geq 0$, m such that $m \geq J_k\phi(x)$

Ensure: Y pseudorandom with density $J_k\phi / \operatorname{vol}(\mathcal{M})$

repeat

$X \leftarrow$ pseudorandom from unif distribution on A

$U \leftarrow$ pseudorandom from unif distribution on $[0, 1]$

until $mU \leq J_k\phi(X)$

$Y \leftarrow X$

Once again, in our context, $J_k\phi = \operatorname{vol}(\mathcal{M}) f$ and $m = pg$, where g is the uniform distribution over A (which has constant density g with support A). In many cases, A (the domain of a parametrization) is, for example, a k-dimensional rectangle, so sampling from the uniform distribution on A is easy.

Remark. Knowing $\operatorname{vol}(\mathcal{M})$ is only necessary when calculating the expected number of iterations needed to generate Y. The density for the uniform distribution over A is the constant function (over A) $g(x) = 1/\operatorname{vol}(A)$. So, if $m = pg(x) = p/\operatorname{vol}(A)$, then we have $p = m \operatorname{vol}(A)$. As shown earlier, the expected number of iterations needed for the first acceptance is

$$c = \frac{p}{q} = m\frac{\operatorname{vol}(A)}{\operatorname{vol}(\mathcal{M})}.$$

Therefore, it is desirable to have tighter bounds of the Jacobian to reduce the number of rejections in the previous algorithm.

Bibliography

[1] Bendich, P., D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and D. Morozov: *Inferring local homology from sampled stratified spaces.* In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 536–546. IEEE, 2007.

[2] Bendich, P., E. Gasparovic, J. Harer, and C.J. Tralie: *Scaffoldings and spines: organizing high-dimensional data using cover trees, local principal component analysis, and persistent homology.* In *Research in Computational Topology*, pages 93–114. Springer, 2018.

[3] Bendich, P., S. Mukherjee, and B. Wang: *Stratification learning through homology inference.* 2010. `https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2273/2714`.

[4] Bhattacharya, R.N., M. Buibas, I.L. Dryden, L.A. Ellingson, D. Groisser, H.W.M. Hendriks, S. Huckemann, H. Le, X. Liu, D.E. Osborne, *et al.*: *Extrinsic data analysis on sample spaces with a manifold stratification.* Bulletin of the Transilvania University of Bragov, Series III, 4:1–15, 2011.

[5] Billingsley, P.: *Probability and measure.* John Wiley & Sons, 1979.

[6] Biscay, R., M. Nakamura, V. Pérez Abreu, and F. Reveles: *Persistencia, probabilidad e inferencia estadística para análisis topológico de datos*, 2016. `http://atd.cimat.mx/es/node/436`.

[7] Bordenave, C. and D. Chafaï: *Lecture notes on the circular law.* Modern aspects of random matrix theory, 72:1, 2014.

[8] Brubaker, Marcus, Mathieu Salzmann, and Raquel Urtasun: *A family of MCMC methods on implicitly defined manifolds.* In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 161–172, 2012.

[9] Devroye, L.: *Non-uniform random variate generation.* Springer, 1986.

[10] Diaconis, P., S. Holmes, M. Shahshahani, *et al.*: *Sampling from a manifold.* In *Advances in modern statistical theory and applications: a Festschrift in honor of Morris L. Eaton*, pages 102–125. Institute of Mathematical Statistics, 2013.

[11] Edelsbrunner, H.: *A short course in computational geometry and topology.* Number Mathematical methods. Springer, 2014.

[12] Fasy, B.T., F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, A. Singh, *et al.*: *Confidence sets for persistence diagrams.* The Annals of Statistics, 42(6):2301–2339, 2014.

[13] Federer, H.: *Geometric measure theory.* Springer, 2014.

[14] Franzoni, G.: *The Klein bottle: variations on a theme.* Notices of the AMS, 59(8):1094–1099, 2012.

[15] Friedman, G.: *Singular intersection homology.* Book in progress, 2018.

[16] Gómez-Larranaga, J.C., F. González-Acuna, and W. Heil: *Classification of simply-connected trivalent 2-dimensional stratifolds.* In *Topology Proceedings*, volume 52, pages 329–340, 2018.

[17] Gómez-Larrañaga, J.C., F. González-Acuña, and W. Heil: *Models of simply-connected trivalent 2-dimensional stratifolds with an implementation code.* arXiv preprint arXiv:1805.06302, 2018.

[18] Hughes, B. and S. Weinberger: *Surgery and stratified spaces.* Surveys on surgery theory, 2:319–352, 2000.

[19] Ibanez, R., E. Abisset-Chavanne, J.V. Aguado, D. Gonzalez, E. Cueto, and F. Chinesta: *A manifold learning approach to data-driven computational elasticity and inelasticity.* Archives of Computational Methods in Engineering, 25(1):47–57, 2018.

[20] Jaworski, P., F. Durante, W. Härdle, and R. Rychlik: *Copula theory and its applications.* Springer, 2009.

[21] Marsaglia, G.: *Choosing a point from the surface of a sphere.* The Annals of Mathematical Statistics, 43(2):645–646, 1972.

[22] Martin, S., A. Thompson, E.A. Coutsias, and J.P. Watson: *Topology of cyclo-octane energy landscape.* The journal of chemical physics, 132(23):234115, 2010.

[23] Muller, M.E.: *A note on a method for generating points uniformly on n-dimensional spheres.* Communications of the ACM, 2(4):19–20, 1959.

[24] Otter, N., M.A. Porter, U. Tillmann, P. Grindrod, and H.A. Harrington: *A roadmap for the computation of persistent homology.* EPJ Data Science, 6(1):17, 2017.

[25] Patrangenaru, V., P. Bubenik, R.L. Paige, and D. Osborne: *Challenges in topological object data analysis.* Sankhya A, pages 1–28, 2018.

[26] Pérez-Angulo, J.M.: *Análisis topológico de datos: Robusticidad y análisis de sensibilidad de algoritmos.* Master's thesis, CIMAT, 2016.

[27] Valente-Castro, F.: *Ripseronr*, 2018. `https://github.com/holt0102/RipserOnR`.

[28] Zhu, B., J.Z. Liu, S.F. Cauley, B.R. Rosen, and M.S. Rosen: *Image reconstruction by domain-transform manifold learning.* Nature, 555(7697):487, 2018.