Bayesian Analysis (2024) **19**, Number 1, pp. 1–20

Bayesian Estimation of Topological Features of Persistence Diagrams*

Asael Fabian Martínez†

Abstract. Persistent homology is a common technique in topological data analysis providing geometrical and topological information about the sample space. All this information, known as topological features, is summarized in persistence diagrams, and the main interest is in identifying the most persisting ones since they correspond to the Betti number values. Given the randomness inherent in the sampling process, and the complex structure of the space where persistence diagrams take values, estimation of Betti numbers is not straightforward. The approach followed in this work makes use of features' lifetimes and provides a full Bayesian clustering model, based on random partitions, in order to estimate Betti numbers. A simulation study is also presented.

MSC2020 subject classifications: Primary 62F15, 62R40; secondary 62H30.

Keywords: Betti numbers, cluster analysis, lifetimes, outlier detection, random partitions, topological data analysis.

1 Introduction

Geometrical and topological methods are modern tools for analyzing highly complex data (Carlsson, 2009; Chazal et al., 2011; Boissonnat et al., 2018). While geometrical techniques capture quantitative information in data, topology reveals qualitative information. Both are useful to uncover patterns and relationships in data and, together with statistical and computational concepts and tools, sometimes complementary, form a powerful set of methods for analyzing modern data.

In particular, topological data analysis (TDA) is a modern field of applied mathematics with considerable interest and activity during the last two decades. It is a collection of tools in the field of data analysis that lies at the intersection of Algebraic Topology, Computational Geometry, Computer Science and Statistics. The main goal of TDA is to use ideas and results from Geometry and Topology to develop tools for revealing and describing relevant features of data objects with an intrinsic and complex structure. Given a cloud point data (dataset, in the terminology in this field), TDA methodologies are useful to understand their underlying space, so we can infer about its shape regardless of the choice of coordinates, deformations or presentations. This is done by estimating topological invariants related to the space, which capture the intrinsic clusters and connections among the clusters present in the cloud point data as

arXiv: 2204.01127

*The author thankfully acknowledges the support of PAPIIT project number IG100221.

†Departamento de Matemáticas, Universidad Autónoma Metropolitana, Unidad Iztapalapa. Av. Ferrocarril San Rafael Atlixco, 09310, Ciudad de México, Mexico, fabian@xanum.uam.mx

 https://doi.org/10.1214/22-BA1341

well as other connectivity information, including the classification of loops and higher dimensional surfaces within the space.

Since TDA is in its core an issue in inference, to discover an unknown structure feature based on sampled cloud point data, some natural questions arise for statisticians. One is how it differs from classical cluster analysis. As pointed out by Carlsson, "TDA uses cluster analysis in building its networks, and builds on cluster analysis to provide additional precision in the taxonomies that are created" (Carlsson, 2016).

There are several TDA methodologies. One of the most common focuses on what is called *persistent homology*, which makes use of algebraic tools in order to discover topological features of data, where the so-called *Betti numbers* codify the number of k-dimensional holes present in the underlying space for different values of k. The method was introduced by Edelsbrunner et al. (2002) and its theory has been further considered in, for example, Carlsson (2009), Edelsbrunner and Harer (2010), Ghrist (2008), Oudot (2015), Zomorodian and Carlsson (2005) and Zomorodian (2005). Furthermore, persistent homology has been applied in a variety of fields including interconnectedness in the banking system (de la Concha et al., 2018), manufacturing systems (Guoa and Banerjee, 2017) and computational biology with industrial and medical engineering applications (Gameiro et al., 2014; Xia and Wei, 2014). Other notable applications have been in data analysis (Bastian et al., 2012; Carlsson, 2009; Lesnick, 2013; Niyogi et al., 2011; Wang et al., 2011; Xu et al., 2012), image analysis (Carlsson et al., 2008; Frosini and Landi, 2013; Singh et al., 2008), detection of subtypes of cancer (Arsuaga et al., 2015; Nicolau et al., 2011), analysis of brain artery trees (Bendich et al., 2016), virus evolution (Chan et al., 2013; Ibekwe et al., 2014; Parida et al., 2015), complex networks (Horak et al., 2009), language processing (Zhu, 2013), sensor networks (de Silva and Ghrist, 2007), spectroscopy (Offroy and Duponchel, 2016), and soil science (Savica et al., 2017), among others.

Roughly speaking, persistent homology can be described as follows. Consider a finite cloud point data with pairwise distances between their points, select a scale $\epsilon > 0$ and join all points at a distance not more than 2ϵ. This gives an indication of a possible topological feature in the data, in particular an initial cluster classification; however, this depends on the scale, which could be hard to choose if the data are high-dimensional. Persistent homology examines data over all scales. The output of this computation is a summary called the *persistence diagram*, a set of pairs (*birth*, *death*). This and the *persistence barcode* encode life spans of topological features from their birth to death, from where one can choose, visually, pairs with a long persistence lifetime ($death - birth$). The procedure is performed over all the dimensions of the data. Topological features of short-scale duration are referred as *topological noise*, whereas the rest are called *topological signal*. Thus, discriminating between these features is of great interest since the topological signal is closely related with the Betti numbers of the underlying space.

A second natural question for statisticians is how randomness and uncertainty are handled. Any topological feature extracted from the cloud point data has random variation that needs to be taken into consideration for meaningful topological inference, and this is exactly the subject matter of Statistics. As referred in Otter et al. (2017), practitioners of TDA often have backgrounds in pure topology and are not well-versed

in statistical approaches to data analysis and there are several challenges for statistically interpreting results in applications of persistent homology, as few statistical tools are currently available. Conversely, TDA methodology has been slow to spread and develop within the statistical community and literature. In the setting of Statistics, persistence diagrams and barcodes are data summaries, or *statistics* based on cloud point data. Thus, notions of probability models for data and sampling distributions apply. A statistical approach to persistent topology was first considered in Bubenik and Kim (2007) where, among other aspects, theoretical persistence barcodes for parametric probability distributions on manifolds, as those considered in directional data, are derived, and then persistence barcodes are estimated using statistical inference principles, like maximum likelihood.

A general challenge for the statistical analysis of persistence diagrams is first to consider probability distributions on the set of persistence diagrams. This set is geometrically very complex, and it is difficult to consider parametric distributions for those diagrams as to allow practical statistical inference. Moreover, persistence diagrams are not in a vector space and therefore one cannot use basic statistical tools like means, variance and moments, but rather Fréchet means (see Mileyko et al., 2011; Munch et al., 2015; Turner et al., 2014). As an alternative, *persistence landscapes* are proposed in Bubenik (2015). They give topological summaries belonging to a space of functions and therefore law of large numbers and central limit theorems in function Banach spaces can be used to perform ad-hoc, non-parametric, classical statistical inference. Under a Bayesian framework, persistence diagrams are modeled in Maroulas et al. (2020) through Poisson point processes for hypothesis testing in classification. On the other hand, for potential use in Statistics, there is the probabilistic limit theorem for persistence diagrams, which is an area of increasing interest in the framework of stochastic topology (Bobrowski and Mukherjee, 2015; Hiraoka et al., 2018; Kahle, 2011; Yogeshwaran and Adler, 2015; Yogeshwaran et al., 2017; Bobrowski and Kahle, 2018).

Another role for Statistics and Probability in persistent homology lies in the problem of disentangling topological noise from topological signals in persistence diagrams. In this direction, Fasy et al. (2014) and Chazal et al. (2018) obtained asymptotic results for the construction of confidence sets for persistence diagrams, using geometric, statistic and probabilistic tools like kernel estimates, concentration inequalities, bootstrap, empirical processes, distance to measure and kernel distance. Those sets show the topological noise corresponding to all topological dimension homologies simultaneously. The confidence sets in Chazal et al. (2018) are robust, and it is also indicated how to construct confidence sets for particular dimension homologies, using the so-called bottleneck bootstrap.

The methodology presented in this work focuses on this second problem of identifying the topological signal. The interest of this topological feature lies in the fact that the quantity of signal features corresponds to the Betti number for each fixed dimension homology. Hence, by providing an estimate of these numbers, it will be possible to understand the underlying space where the cloud point data live.

Our approach makes use of the topological features' lifetime spans. For every homology level, the lifetimes computed from the persistence diagram are used to identify

the topological noise and signal. Since, roughly speaking, the noise is of small magnitude when compared to the signal, and both of them inherit the randomness of the sampled cloud point data, the disentanglement of both features can be posed as a problem of outlier detection. The number of outlier lifetimes, thus, will correspond to the estimated Betti number. The detection of outliers is done by means of a full Bayesian model based on random partitions. The specific support for the random partition will allow to identify the topological signal from the groups containing the largest lifetime values.

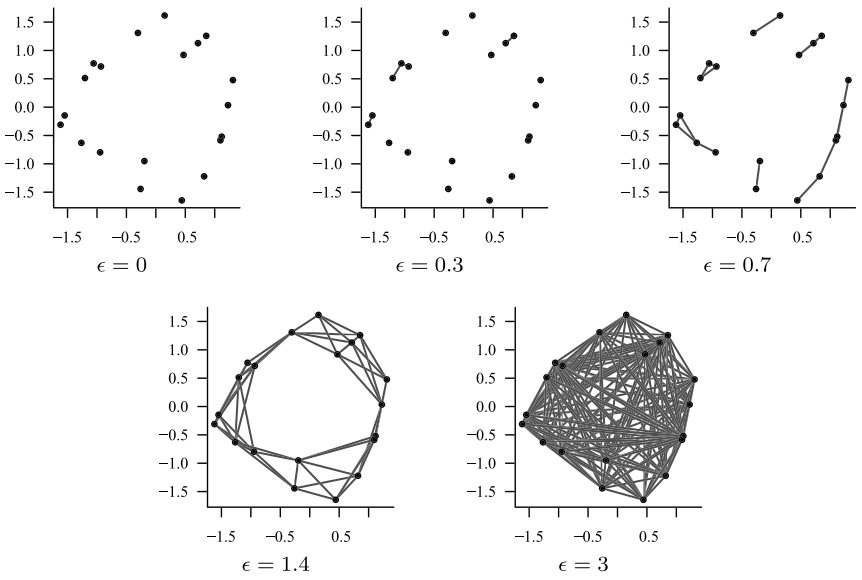
The paper is organized as follows. Section 2 briefly explains persistent homology and how it is related to clustering and provides more information about the underlying sample space. Section 3 presents the proposed Bayesian model for estimating the Betti numbers via outlier detection. A simulation study is also performed in Section 4, and Section 5 contains some final remarks and future work.

2 Topological data analysis in a nutshell

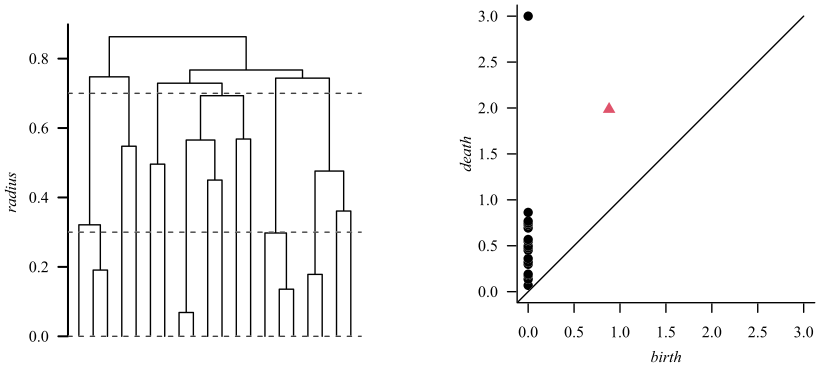
TDA and Statistics, at first glance, tackle the same problem of clustering, as already mentioned. This makes it a good starting point to better understand where TDA, in particular persistent homology, differs; we can also have a clear picture of its potential. For a more comprehensive treatment of TDA methods, we refer the reader to Nanda and Sazdanović (2014), Ferri (2017), and Wasserman (2018), among others. A theoretical treatment can be found, for example, in Zomorodian (2005). Also, the work Otter et al. (2017) overviews and compares the various methods available for computing persistent homology.

Cluster analysis aims to gather a set of items according to some similarity conditions. This set is partitioned into non-empty subsets, called groups or clusters, in such a way that all items in the same group are more similar among them than those in any other group. The degree of similarity is usually quantified according to some probability model or a specific distance function. Under the distance-based approach, a common method is the agglomerative hierarchical clustering with single linkage. Suppose we use the Euclidean distance, d. Given a cloud point data, we can build a graph using these points as vertices and its edges defined as follows: for a fixed $\epsilon \geq 0$, an edge xy is added if and only if $d(x, y) \leq 2\epsilon$ for any two different vertices x, and y. Clusters are obtained from the connected components of the resulting graph. Figure 1a illustrate this procedure. The challenging part is to choose an appropriate cut-off value for ϵ, since when $\epsilon = 0$, there will be as many clusters as observations, and when ϵ is large enough, there will be a single cluster containing all the observations.

Regarding the cut-off value, the persistent homology approach consists on keeping track of the evolution of the number of connected components as the value of parameter ϵ increases; it is stored in a topological summary, formally called *persistence diagram*. Those connected components persisting for more time are the more meaningful, so they will determine the clustering structure for the given cloud point data (see Figure 1c). The 0th persistence diagram, T^0, is a multiset of (*birth*, *death*) points, (b_i, d_i), in $\mathbb{R}^2_+$, i.e.



(a) Agglomerative hierarchical clustering with single linkage using different radius values ϵ



(b) Dendrogram

(c) Persistence diagram

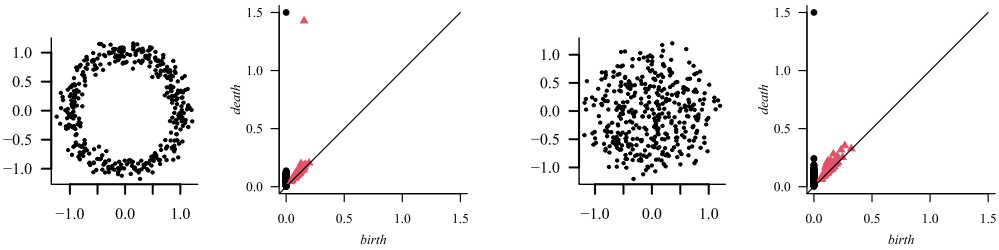
Figure 1: Illustration of hierarchical clustering, Panels (a) and (b), and its corresponding persistence diagram, Panel (c). Horizontal dashed lines in the dendrogram correspond to some values of ϵ.

the first entry indicates the time (value of ϵ) where a connected component is created, and the second one is its death time; so we can define this summary as

$$T^0 = \{(b_i, d_i) \in \mathbb{R}^2_+ : b_i \leq d_i, i = 1, \ldots, n\},$$

for n the sample size. Black dots in Figure 1c are a graphical representation of this set for the cloud point data on Figure 1a.

This described clustering procedure, actually, corresponds to the computation of the



(a) Data over a ring and its persistence diagram

(b) Data over a disc and its persistence diagram

Figure 2: Illustration of one-dimensional holes for cloud point data over different manifolds: (a) a ring, and (b) a disc.

0-homology under the TDA terminology. However, persistent homology is able to model higher dimensional relationships among data points. For example, Figure 2 shows two cloud point data sets, each one formed by a single cluster, but they do not have the same shape. This additional information cannot be obtained from traditional clustering procedures, but it is indeed useful to, for example, define a more adequate probabilistic model in each case. Higher levels of homology, h-homology for $h \geq 1$, are also captured during the computation of the persistent homology, and stored in the corresponding hth persistence diagram T^h. In the examples of Figure 2, both persistence diagrams, T^0 and T^1, are plotted together, represented by the black circles and red triangles respectively. This is one of the advantages of TDA methodologies. Using all this information together, one can have a more complete picture of the shape where data live.

Another application of TDA is related to dimensionality reduction. There are situations where each observation in a cloud point data takes values in an s-dimensional space, for example $\mathbb{R}^s$, but the meaningful features, or their intrinsic shape, say $\mathcal{M}$, is embedded in such a space, so $s' = \dim(\mathcal{M}) < s$. Persistence homology will only detect topological features until dimension homology s', even though it is computed from the bigger space of dimension s.

In any case, the main interest of persistence diagrams is the quantification of the most relevant features in the cloud point data, that is, determining the number of clusters, cycles, voids, and in general, the number of hth dimensional holes the data have. Indeed, this quantities correspond to the Betti numbers β_h, $h \geq 0$, inherent to the sample space. In the examples of Figure 2, there is only one cluster (black circle at the top of the plots), so $\beta_0 = 1$ in both cases. However, the two cloud point data sets are different by the presence of a one-dimensional hole, cycle, in the ring cloud point data (indicated by the red triangle far from the rest), so we have $\beta_1 = 1$, whereas for the other one, $\beta_1 = 0$.

Nevertheless, determining the value of the Betti numbers from persistence diagrams is not a straightforward task. Each of these numbers corresponds to the quantity of points *far enough* from the main diagonal in the graphical representation of its diagram; the rest of the points can be considered as *noise*. Therefore, the need of some formal method to classify these points accordingly is evident. In the following section, we

elaborate on this issue and present a methodology able to identify the points with the main objective of estimating the Betti numbers.

3 A methodology for modeling topological features

Randomness is an important issue that has to bear in mind when performing statistical inference in persistent homology. Indeed, typically the cloud point data $y = (y_1, \ldots, y_m)$ to be subject to topological analysis is a sample of random points in a metric space. Consequently, any topological feature extracted from it has random variations that need to be taken into consideration for meaningful inferences.

More specifically, two of the common cloud point data generating mechanisms are the following. Under the first scenario, each point y_i, $i = 1, \ldots, m$, is independently drawn with the same probability distribution F, usually the uniform distribution, supported on a manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$. Under the second one, each point y_i, $i = 1, \ldots, m$, has the form $y_i = u_i + z_i$, where u_i is independently drawn with the same probability distribution F supported on $\mathcal{M}$, as in the first setting, and z_i is an independent and identically distributed (iid) random perturbation drawn from some distribution G on $\mathbb{R}^d$. For example, G can be a d-variate Gaussian distribution with zero mean and covariance matrix $\sigma^2 I$, with I the identity matrix, and the dispersion $\sigma^2 > 0$ is typically small. Another possibility is to assume that z_i is randomly drawn in such a way that it is perpendicular to the tangent space of the manifold $\mathcal{M}$ at the point u_i. Notice that in these last two cases the cloud point data does not lie exactly over the manifold $\mathcal{M}$ but close to it.

In this framework, the main purpose of TDA is to infer topological features of the underlying manifold $\mathcal{M}$ from the given random cloud point data. In particular, as already mentioned, persistent homology keeps track of the evolution of topological features when varying a filtration parameter, in order to characterize the shape of the manifold through the evolution of its homological groups. However, this endeavor faces the difficulty that, due to the discrete nature and sampling variability of the cloud point data, topological features of short duration over the filtration emerge which are irrelevant regarding the true topology of the underlying manifold of interest $\mathcal{M}$. In Fasy et al. (2014) and Chazal et al. (2018), these features of random short lifetimes, which can be explained just by sampling variability, are termed *topological noise*, in contrast to the *topological signal* corresponding to persistent features.

Thus, discriminating between topological noise and topological signal is a crucial problem in TDA. For this, in Fasy et al. (2014) and Chazal et al. (2018), bootstrap confidence bands around the diagonal of persistence diagrams are introduced, and points (*birth*, *death*) inside the band are disregarded as due to topological noise. The so called Vietoris-Rips filtration is used to compute the persistent homology, see Edelsbrunner and Harer (2010). As described in Section 2, for each homology level h, the persistence diagram is represented by the multiset:

$$T^h = \{(b_i^h, d_i^h) \in \mathbb{R}_+^2 : b_i^h \leq d_i^h, i = 1, \ldots, n^h\},$$

for some $n^h < \infty$. The lifetimes of the homology classes in the persistence diagram are given by

$$l_i^h = d_i^h - b_i^h, \qquad i = 1, \ldots, n^h.$$

Randomness of the cloud point data leads to randomness of these lifetimes. Assume that $\mathcal{M}$ is a smooth manifold composed by a finite number of closed connected components, as it is common in practice. Also assume that the size m of the cloud point data is large in comparison with such number of connected components, as well as with the number of holes and other Betti numbers in $\mathcal{M}$. Then, most of the lifetimes l_i^h, $i = 1, \ldots, n^h$, except for a few, will be the result of birth and death of homological classes due to topological noise in sampling the manifold through the cloud point data.

With all these elements explained, a general statistical modelling approach for lifetimes associated with topological noise is introduced next. This provides, as a byproduct, a statistical tool for disentangling topological noise from topological signal, without requiring intensive computations involved in bootstrapping persistence diagrams.

3.1 Topological signal detection through random partition modeling

Each lifetime can be identified as only one type of feature: topological noise or topological signal, and it is expected that the quantity of lifetimes being topological noise is much larger than those being topological signal. Focusing on the noise, they appear due to the sampling process, so it is possible to describe them according to some probabilistic model. In contrast, the topological signal will appear for any other reason not explained by the noise model. However, both type of features are collected together; thus, the topological signal becomes an *outlier* with respect to the topological noise. Hence, we will be able to disentangle both of them by applying some methodology designed for this purpose.

For the sake of completeness, given some arbitrary dataset, any observation which is inconsistent with the remainder is called an *outlier*. Under a probabilistic approach, it is assumed that $n - p$ observations, from a total of n, arise from some model, whereas the remaining p observations come from a different one. In general, $n \gg p$. The literature for methods tackling the problem of outlier detection is wide and comprises computational, probabilistic and statistical approaches; see for example, Wang et al. (2019) for a recent review, and Quintana and Iglesias (2003); Quintana (2006); Shotwell and Slate (2011) for some Bayesian nonparametric methodologies. In this work, a clustering approach is followed built upon Bayesian nonparametric commonly used tools, in particular, we make use of restricted random partitions as the methodological component, similarly to Fuentes-García et al. (2010); Wade et al. (2014).

Random partitions are probabilistic tools suitable to perform clustering (see, e.g. Lijoi et al., 2008; Müller et al., 2015), since their sampling space, known as *set partitions*, and denoted throughout this work by $\mathcal{P}$, encodes every possible arrangement of any set of items into a number of nonempty groups, or clusters. As a simple example, consider the set $\{y_1, y_2, y_3\}$, then all their possible arrangements are

$$\{\{y_1, y_2, y_3\}\},\ \{\{y_1\}, \{y_2, y_3\}\},\ \{\{y_1, y_2\}, \{y_3\}\},\ \{\{y_1, y_3\}, \{y_2\}\},\ \{\{y_1\}, \{y_2\}, \{y_3\}\},$$

where, for example, $\{\{y_1, y_2\}, \{y_3\}\}$ means that there are two clusters: one formed by observations y_1 and y_2, and the second one formed only by y_3. Establishing some notation, a set partition $\pi \in \mathcal{P}$ is a partition of some set of cardinality n having k nonempty subsets, groups or blocks, for some $1 \leq k \leq n$, where each group is denoted by π_j, $j = 1, \ldots, k$. Furthermore, simplifying the writing, partitions will be denoted by $\pi_1/\cdots/\pi_k$ instead of $\{\pi_1, \cdots, \pi_k\}$.

Outlier detection, for the context at issue, can be performed by means of a specific class of random partitions, whose support is a subset of $\mathcal{P}$. Let us assume the homology level h is fixed for the rest of the explanation. Since lifetimes are all positive numbers, it will be assumed they are ordered, i.e. $l_i \leq l_{i+1}$ for $i = 1, \ldots, n-1$, with n the total number of lifetimes. This allows us to locate the potential topological signal as the largest lifetime values; the rest of them will be the topological noise. Then, it is expected that the topological signal will be grouped in a few clusters π_j, all of them with a few elements or even being singletons. By ordering lifetime values, such clusters would be the most-right of them. However, any set partition is invariant to permutations of their blocks (one of the reasons of the very well known problem of *label-switching*) and we cannot stick to this rule.

Therefore, it is necessary to restrict the sampling space of the random partition. This new space will only contain set partitions $\pi \in \mathcal{P}$ such that every block π_j consists of consecutive items and $\max \pi_j < \min \pi_{j+1}$ for $j = 1, \ldots, k-1$ with k the number of groups in π. These conditions are known as the *no-gaps* assumption (cf. Fuentes-García et al., 2010; Martínez and Mena, 2014; Wade et al., 2014). In the small example, the partition $\{y_1, y_3\}/\{y_2\}$ does not satisfy this assumption. Let us denote by $\mathcal{R}$ the set of all no-gaps set partitions.

Random partitions can be used in conjunction with a model-based approach, meaning that a probabilistic model g_j is assigned to each group π_j. As a consequence, all observations $y_i \in \pi_j$ are distributed according to g_j [iid]. Therefore, given these elements, the proposed model for outlier detection can be written hierarchically as

$$\begin{aligned} l_i|\pi, \phi &\sim g(l_i|\phi_j)\mathbf{1}(l_i \in \pi_j) \ \text{[ind]}, \quad i = 1, \ldots, n, \\ \phi_j|\pi &\sim \nu_0 \ \text{[iid]}, \\ \pi &\sim \rho_0, \end{aligned} \tag{1}$$

where g is some probability distribution supported over $\mathbb{R}^+$ with driving finite-dimensional parameter ϕ_j, the distribution ν_0 is the prior for each parameter ϕ_j, and π is an $\mathcal{R}$-valued random partition with prior distribution ρ_0. Among the candidates for distribution g, the log-normal distribution of parameters $\phi_j = (\mu_j, \sigma_j^2) \in \mathbb{R} \times \mathbb{R}^+$ will be used, and its conjugate for ν_0 is chosen, i.e. a normal-gamma distribution of parameters (m, c, a, b). With respect to the prior distribution for the random partition, ρ_0, a restriction of the so-called exchangeable partition probability functions (EPPFs) is used, see Martínez and Mena (2014); Wade et al. (2014) for further details. EFFPs are a widely used class of distributions in Bayesian nonparametric methodologies. In particular, for the EPPF derived from the Dirichlet process Ferguson (1973), its corresponding

$\mathcal{R}$-valued distribution is written as

$$\Pr(\pi = \pi_1/\cdots/\pi_k) = \binom{n}{n_1 \cdots n_k} \frac{\theta^k}{k!(\theta)_{n\uparrow}} \prod_{j=1}^{k} \Gamma(n_j), \tag{2}$$

for any set partition $\pi = \pi_1/\ldots/\pi_k \in \mathcal{R}$, and where $n_j = \#\pi_j$ is the cardinality of block π_j for $j = 1\ldots, k$, $(x)_{r\uparrow} = x(x+1)\cdots(x+r-1)$ is the Pochhammer symbol or rising factorial, and $\theta > 0$ is the total mass parameter for the Dirichlet process.

In order to compute point estimates for this model, we resort to numerical procedures, in particular, Markov chain Monte Carlo (MCMC) techniques. Given the data, a sample of the posterior distribution

$$p(\pi|l_1, \ldots, l_n) \propto \rho_0(\pi) L(l_1, \ldots, l_n|\pi), \tag{3}$$

is obtained. In Appendix A (Martínez, 2022), the derivation of the complete MCMC sampling scheme is presented. For an alternative estimation procedure based on neural networks, see Fuentes-García et al. (2019).

3.2 Estimation of Betti numbers

Our approach classifies the topological signal as outliers with respect to the rest of the lifetime values, the topological noise; this signal is encoded in the largest values. Topological noise might contain some information regarding the underlying geometry of the sampling space, whereas the topological signal is of interest since it determines the value for the Betti numbers β_h, for each homology level $h \geq 0$. Therefore, if there is some topological signal in the data, it means there will be some lifetime value l_{n^*} being the last element conforming the topological noise, and letting the remaining l_j, for $j = n^* + 1, \ldots, n$, the signal, with n the lifetimes' sample size. Hence, the Betti number estimator, $\hat{\beta}_h$, corresponds to the size of the signal, that is $\hat{\beta}_h = n - n^*$.

Working under the $\mathcal{R}$-valued partition approach allows us to locate the potential topological signal as the rightmost groups of a partition π, whereas the rest of them will contain the topological noise. Given our probabilistic framework, we need to provide some point estimate $\hat{\beta}_h$, $h \geq 0$. However, the sample space $\mathcal{R}$ is not an ordered set, hindering the computation of point estimates. Being $\mathcal{R}$ a discrete space, a simple and easily interpretable estimate is the mode, although some other options might be used (c.f. Dahl, 2006; Wade and Ghahramani, 2018).

Let $\tilde{\pi}$ be the posterior modal partition. Ideally, $\tilde{\pi}$ should consist on only two groups, say $\tilde{\pi}^{(n)}$ and $\tilde{\pi}^{(s)}$, one for each topological feature: noise and signal, respectively. This simplification causes some lost of information, though. Model (1) resembles mixture models (see, e.g. Müller et al., 2015; Martínez, 2019, for more details), and one of their advantages is their capability for fitting complex distributions, which is achieved by combining several mixing components, even though the resulting distribution exhibits a single mode. This is noteworthy to say since it might be expected that the topological noise behaves according to some probability distribution, even though it is not bounded

to be of the form of a single kernel function $g(\cdot|\phi_j)$. A similar rationale applies for the topological signal. As a result, the posterior modal partition $\tilde{\pi}$ can be conformed by more than two clusters. In particular, the topological signal, may be spread over a few blocks, so $\tilde{\pi}^{(s)} = \tilde{\pi}_\kappa \cup \cdots \cup \tilde{\pi}_k$, for some κ very close to k, the number of blocks in $\tilde{\pi}$. According to the definition of outlier, we can only expect that the size of $\tilde{\pi}^{(s)}$ is very small when compared with the sample size n. Therefore, by setting

$$\kappa = \min\left\{ s : \sum_{j=s}^{k} \frac{\#\tilde{\pi}_j}{n} \leq q \right\},$$

for some small $q > 0$, the relative proportion of expected outliers, e.g. 2% or 3%, the Betti number β_h can be estimated as

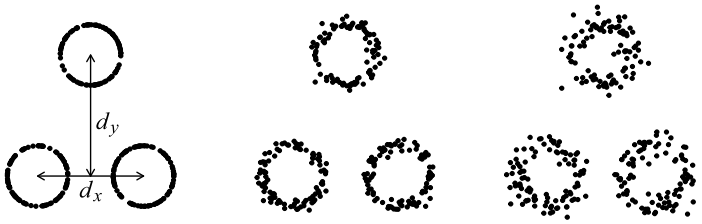
$$\hat{\beta}_h = \#\tilde{\pi}_\kappa + \cdots + \#\tilde{\pi}_k. \tag{4}$$

For the 0-homology level, it is important to highlight that the value obtained for $\hat{\beta}_0$ should be corrected by adding one. The computation of persistence diagrams requires a maximum radius ϵ, which is somehow arbitrarily fixed, and there will always be a lifetime having such a value. Then, it is necessary to remove it before any further processing, but it does count one connected component.

4 Simulation study

The proposed methodology is tested under some simulated scenarios, using manifolds having different shape and dimension. We start considering the 0th homology level using synthetic cloud point data uniformly distributed over the circle, varying the quantity of circles, their location, and the sample size. Additionally, a small noise is introduced in each dataset (see Figure 3) with the purpose of better understand the robustness of the model. For the sake of comparison, we also present the results provided by the R package TDA (Fasy et al., 2022), which includes the bootstrap methods of Fasy et al. (2014) and Chazal et al. (2018) providing 95% confidence bands for persistence diagrams. This estimator will be denoted by $\bar{\beta}_h$.

First examples are taken from a manifold conformed by r circles, $r = 1, 2, 3$, each of radius one; a cloud point data of size $n = 600$ is drawn for each case. In addition, two levels of noise are included, as depicted in Figure 3, consisting on a Gaussian perturbation of each point over the circle with standard deviation $\sigma = 0.1, 0.2$. For the cases $r = 2, 3$, the circles are separated from their centers by 5 units, according to Figure 3a. An MCMC run was executed for each cloud data point, taking a sample of size 5 000 after discarding a first batch of 10 000. Hyperparameter settings are as follows: $(0, 0.5, 1.1, 0.1)$ for the prior distribution ν_0, and $(1.1, 0.1)$ for the total mass parameter, θ, prior. Table 1 presents the posterior estimates for these datasets. It is expected that the 0th Betti number corresponds to the number of circles, i.e. $\beta_0 = r$. In all cases, the value of β_0 is correctly estimated by our approach, i.e. $\hat{\beta}_0$, by allowing a $q = 0.03$ relative proportion of outliers at most; a similar performance is seen for



(a) No noise (b) Noise $\sigma = 0.1$ (c) Noise $\sigma = 0.2$

Figure 3: Examples for the different noise levels using the same manifold. Distances, displayed in Panel (a), d_x and d_y will range from 1 to 5 units.

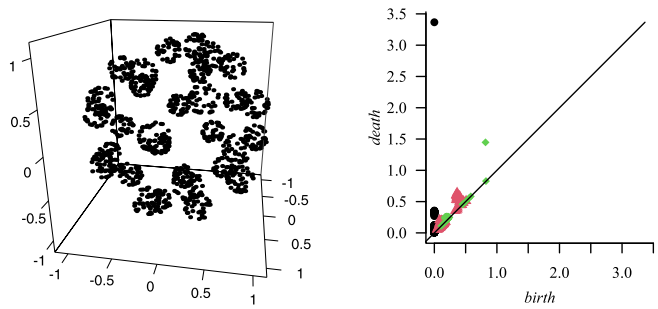
r	σ	$(n_1, \ldots, n_k)$	*prob.*	$\hat{\beta}_0$	$\bar{\beta}_0$
1	—	$(72, 207, 203, 117)$	0.019	1	1
	0.1	$(14, 167, 311, 107)$	0.017	1	1
	0.2	$(87, 208, 256, 48)$	0.018	1	1
2	—	$(69, 168, 211, 150, 1)$	0.019	2	2
	0.1	$(73, 226, 241, 58, 1)$	0.017	2	2
	0.2	$(89, 220, 247, 42, 1)$	0.017	2	2
3	—	$(39, 127, 204, 154, 73, 2)$	0.022	3	3
	0.1	$(15, 137, 254, 174, 17, 2)$	0.001	3	3
	0.2	$(40, 130, 179, 216, 32, 2)$	0.010	3	3

Table 1: Posterior estimates for the firsts toy examples, consisting on r circles, with Gaussian noise σ. Modal partition $\tilde{\pi}$ is presented in terms of block sizes $(n_1, \ldots, n_k)$ together with its probability. Last columns contain the estimated Betti numbers for our approach, $\hat{\beta}_0$, and the confidence-band based, $\bar{\beta}_0$.

r	σ	$(n_1, \ldots, n_k)$	*prob.*	$\hat{\beta}_0$	$\bar{\beta}_0$
2	—	$(69, 168, 211, 150, 1)$	0.019	2	2
	0.1	$(73, 226, 241, 58, 1)$	0.017	2	2
	0.2	$(89, 220, 247, 42, 1)$	0.017	2	2
3	—	$(39, 127, 204, 154, 73, 2)$	0.022	3	3
	0.1	$(15, 137, 254, 174, 17, 2)$	0.001	3	3
	0.2	$(40, 130, 179, 216, 32, 2)$	0.010	3	3

Table 2: Posterior estimates for the second toy examples, consisting on r circles of different radii, with Gaussian noise σ. Modal partition $\tilde{\pi}$ is presented in terms of block sizes $(n_1, \ldots, n_k)$ together with its probability. Last columns contain the estimated Betti numbers for our approach, $\hat{\beta}_0$, and the confidence-band based, $\bar{\beta}_0$.

the bootstrap-based estimator $\bar{\beta}_0$. The supplemental material, Appendix B, contains a



(a) Cloud point data (b) Persistence diagram

Figure 4: Cloud point data and persistence diagram over a spherical-Fibonacci manifold.

more exhaustive simulation study (Martínez, 2022).

A second example consists on a manifold made by spheres whose centers are points of a spherical Fibonacci lattice. Over each sphere, a random sample of 40 points is drawn; the manifold consists on 30 spheres, see Figure 4. The purpose of this example is testing the estimation of higher dimensional holes, namely cycles and voids, corresponding to β_1 and β_2 respectively. The MCMC specification is the same as before. Posterior estimates are presented in Table 3. Regarding β_0, one connected component is lost for our approach, $\hat{\beta}_0$; this might be due to the data sampling process. In fact, a closer inspection of the persistence diagram shows that there are only 28 persistent features. For the one-dimensional holes, β_1, no persistent features are detected. Finally, the study of two-dimensional holes shows an interesting fact. Its estimate is $\hat{\beta}_2 = 1$, indicating one void due to the Fibonacci lattice, but the second block actually indicates that the void for each individual small sphere is also detected. These results for the homology levels H_1 and H_2 are consistent with the theoretical topological features of a sphere, i.e. $\beta_1 = 0$ and $\beta_2 = 1$. On the other hand, it is worth saying that for each homology level, a different number of topological features are detected: $n^0 = 1\,199$ plus the one removed as explained, $n^1 = 330$, and $n^2 = 47$; except for n^0 which is always the number of observations, the rest is random.

Regarding the bootstrap-based method, it is not able to detect the small spheres as separated connected components, only the biggest one induced by all of them, since $\bar{\beta}_0 = 1$. No cycles are detected as expected. Finally, $\bar{\beta}_2 = 1$ meaning that only the biggest void is detected.

5 Concluding remarks

Topological data analysis is an emerging field of applied mathematics providing useful topological and geometrical information about the sample space. In particular, we have described persistent homology, one of the most common methodologies in TDA, where the main purpose is to discover topological features in such a space. The most relevant

h	$(n_1, \ldots, n_k)$	*prob.*	$\hat{\beta}_h$	$\bar{\beta}_h$
0	$(76, 258, 468, 369, 28)$	0.020	29	1
1	$(72, 143, 96, 19)$	0.021	0	0
2	$(16, 30, 1)$	0.521	1	1

Table 3: Posterior estimates for the spherical-Fibonacci manifold example, for the homology levels H_h, $h = 0, 1, 2$. Modal partition $\tilde{\pi}$ is presented in terms of block sizes $(n_1, \ldots, n_k)$ together with its probability. Last columns contain the estimated Betti numbers for our approach, $\hat{\beta}_0$, and the confidence-band based, $\bar{\beta}_0$.

features are summarized in the Betti numbers β_h, $h \geq 0$, quantifying the number of h dimensional holes.

Under a statistical viewpoint, determining the values for the Betti numbers is an inference problem. However, it has not been straightforward providing point estimates. The persistence diagram, the topological summary of persistent homology, takes values in a very complex space. Additionally, the observed cloud point data contains an inherent randomness, which is translated to the persistence diagram, so not all recorded features are relevant. Therefore, in any persistence diagram, topological noise and topological signal are mixed up.

While this work aims to close the gap between TDA and Statistics practitioners, its main contribution is to provide a statistical study of persistence diagrams by means of lifetime's topological features. This approach eases the disentanglement of the topological features and allows to identify and quantify the topological signal. Following a full Bayesian framework, the topological signal identification is treated as an outlier detection problem. The presented clustering model, based on random partitions, agglomerates the most persistent lifetimes as outliers, and their number is associated with the corresponding Betti number. Its performance is as good as the one based on confidence bands, but ours seems to capture more information in the noise-feature groups.

Furthermore, the methodology is tested by an extensive simulation study and some important remarks can be derived from it. A correct identification of the topological signal depends on the geometry of the manifold $\mathcal{M}$, that is, its relevant shape features should be clear enough otherwise the sampling process would veil their presence. In the circles manifold example, each component should be far from each other to be counted correctly.

A second important remark is the sampling process. All this work was performed by assuming the cloud point data was drawn uniformly from $\mathcal{M}$, the distribution F in Section 3. Indeed, this is a common assumption in most TDA literature and is due to the fact we wish to identify the shape of $\mathcal{M}$ so a sample covering the whole manifold is required. Clearly, affecting the sample by adding some noise or by using another distribution F will harden this task. We explore the first case in Appendix B (Martínez, 2022) and can see there are escenarios where the estimation fails. A deep study along these results and remarks, and their applications, are part of the ongoing work.

Supplementary Material

Supplementary Material for "Bayesian Estimation of Topological Features of Persistence Diagrams"(DOI: 10.1214/22-BA1341SUPP; .pdf). The derivation of the MCMC sampling scheme is presented, as well as an extensive simulation study is performed by taking several configurations for the r-circle simulated examples.

References

Arsuaga, J., Borrman, T., Cavalcante, R., Gonzalez, G., and Park, C. (2015). "Identification of Copy Number Aberrations in Breast Cancer Subtypes Using Persistence Topology." *Microarrays*, 4: 339–369. 2

Bastian, R., Hubert, M., and Heike, L. (2012). "Multivariate data analysis using persistence-based filtering and topological signatures." *IEEE Transactions on Visualization and Computer Graphics*, 18: 2382–2391. 2

Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). "Persistent homology analysis of brain artery trees." *The Annals of Applied Statistics*, 10: 198–218. MR3480493. doi: https://doi.org/10.1214/15-AOAS886. 2

Bobrowski, O. and Kahle, M. (2018). "Topology of random geometric complexes: a survey." *Journal of Applied and Computational Topology*, 1(3): 331–364. MR3975557. doi: https://doi.org/10.1007/s41468-017-0010-0. 3

Bobrowski, O. and Mukherjee, S. (2015). "The topology of probability distributions on manifolds." *Probability Theory and Related Fields*, 161. MR3334278. doi: https://doi.org/10.1007/s00440-014-0556-x. 3

Boissonnat, J.-D., Chazal, F., and Yvinec, M. (2018). *Geometric and Topological Inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press. MR3837127. doi: https://doi.org/10.1017/9781108297806. 1

Bubenik, P. (2015). "Statistical topological data analysis using persistence landscapes." *Journal of Machine Learning Research*, 16: 77–102. MR3317230. 3

Bubenik, P. and Kim, P. T. (2007). "A statistical approach to persistent homology." *Homology, Homotopy and Applications*, 9: 337–362. MR2366953. 3

Carlsson, G. (2009). "Topology and data." *Bulletin of the American Mathematical Society*, 46: 255–308. MR2476414. doi: https://doi.org/10.1090/S0273-0979-09-01249-X. 1, 2

Carlsson, G. (2016). "Why TDA and Clustering Are Not The Same Thing." www.ayasdi.com/why-tda-and-clustering-are-different. Accessed 19 December 2021. 2

Carlsson, G., Ishkhanov, T., De Silva, V., and Zomorodian, A. (2008). "On the local behavior of spaces of natural images." *International Journal of Computer Vision*, 76: 1–12. MR3715451. doi: https://doi.org/10.1007/s11263-007-0056-x. 2

Chan, J. M., Carlsson, G., and Rabadan, R. (2013). "Topology of viral evolution." *Proceedings of the National Academy of Sciences*, 110: 18566–18571. MR3153945. doi: https://doi.org/10.1073/pnas.1313480110. 2

Chazal, F., Cohen-Steiner, D., and Mérigot, Q. (2011). "Geometric inference for probability measures." *Foundations of Computational Mathematics*, 11: 733–751. MR2859954. doi: https://doi.org/10.1007/s10208-011-9098-0. 1

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2018). "Robust Topological Inference: Distance To a Measure and Kernel Distance." *Journal of Machine Learning Research*, 18(159): 1–40. MR3813808. 3, 7, 11

Dahl, D. B. (2006). "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model." In Do, K.-A., Müller, P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 201–218. Cambridge University Press. MR2706330. 10

de la Concha, A., Martinez-Jaramillo, S., and Carmona, C. (2018). "Multiplex Financial Networks: Revealing the Level of Interconnectedness in the Banking System." In Cherifi, C., Cherifi, H., Karsai, M., and Musolesi, M. (eds.), *Complex Networks & Their Applications VI*, 1135–1148. Springer International Publishing. 2

de Silva, V. and Ghrist, R. (2007). "Coverage in sensor networks via persistent homology." *Algebraic & Geometric Topology*, 7: 339–358. MR2308949. doi: https://doi.org/10.2140/agt.2007.7.339. 2

Edelsbrunner, H. and Harer, J. L. (2010). *Computational Topology: An Introduction*. American Mathematical Society. MR2572029. doi: https://doi.org/10.1090/mbk/069. 2, 7

Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). "Topological persistence and simplification." *Discrete Computation & Geometry*, 28: 511–533. MR1949898. doi: https://doi.org/10.1007/s00454-002-2885-2. 2

Fasy, B. T., Kim, J., Lecci, F., Maria, C., Millman, D. L., and Rouvreau, V. (2022). *TDA: Statistical Tools for Topological Data Analysis*. R package version 1.8.7. 11

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishan, S., and Singh, A. (2014). "Confidence sets for persistence diagrams." *Annals of Statistics*, 42: 2301–2339. MR3269981. doi: https://doi.org/10.1214/14-AOS1252. 3, 7, 11

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. MR0350949. 9

Ferri, M. (2017). "Persistent Topology for Natural Data Analysis - A Survey." In Holzinger, A., Goebel, R., Ferri, M., and Palade, V. (eds.), *Towards Integrative Machine Learning and Knowledge Extraction/*, 117–133. Springer International Publishing. 4

Frosini, P. and Landi, C. (2013). "Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval." *Pattern Recognition Letters*, 34: 863–872. MR2872256. doi: https://doi.org/10.1007/978-3-642-23672-3_36. 2

Fuentes-García, R., Mena, R. H., and Walker, S. G. (2010). "A Probability for Classification Based on the Dirichlet Process Mixture Model." *Journal of Classification*, 27: 389–403. MR2748990. doi: https://doi.org/10.1007/s00357-010-9061-9. 8, 9

Fuentes-García, R., Mena, R. H., and Walker, S. G. (2019). "Modal posterior clustering motivated by Hopfield's network." *Computational Statistics & Data Analysis*, 137: 92–100. MR3921062. doi: https://doi.org/10.1016/j.csda.2019.02.008. 10

Gameiro, M., Hiraoka, Y., Izumi, S., Kramar, V., K. Mischaikow, and Nanda, V. (2014). "Topological measurement of protein compressibility via persistence diagrams." *Japan Journal of Industrial and Applied Mathematics*, 32: 1–17. MR3318898. doi: https://doi.org/10.1007/s13160-014-0153-5. 2

Ghrist, R. (2008). "Barcodes: the persistent topology of data." *Bulletin of the American Mathematical Society*, 45: 61–75. MR2358377. doi: https://doi.org/10.1090/S0273-0979-07-01191-3. 2

Guoa, W. and Banerjee, A. (2017). "Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs." *Journal of Manufacturing Systems*, 43(2): 225–234. 2

Hiraoka, Y., Shirai, T., and Trinh, K. D. (2018). "Limit theorems for persistence diagrams." *The Annals of Applied Probability*, 28(5): 2740–2780. MR3847972. doi: https://doi.org/10.1214/17-AAP1371. 3

Horak, D., Maletic, S., and Rajkovic, M. (2009). "Persistent homology of complex networks." *Journal of Statistical Mechanics: Theory and Experiment*, 3: P03034. MR2495860. doi: https://doi.org/10.1088/1742-5468/2009/03/p03034. 2

Ibekwe, A. M., Ma, J., Crowley, D. E., Yang, C. H., Johnson, A. M., Petrossian, T. C., and Lum, P. Y. (2014). "Topological data analysis of escherichia codi and non-survival in soils." *Frontiers in Cellular and Infection Microbiology*, 4(122). 2

Kahle, M. (2011). "Random geometric complexes." *Discrete & Computational Geometry. An International Journal of Mathematics and Computer Science*, 45(3): 553–573. MR2770552. doi: https://doi.org/10.1007/s00454-010-9319-3. 3

Lesnick, M. (2013). *Studying the shape of data using topology*. The Institute Letter Summer 2013, Institute for Advanced Study. 2

Lijoi, A., Mena, R. H., and Prünster, I. (2008). "A Bayesian Nonparametric Approach for Comparing Clustering Structures in EST Libraries." *Journal of Computational Biology*, 15(10): 1315–1327. MR2461978. doi: https://doi.org/10.1089/cmb.2008.0043. 8

Maroulas, V., Nasrin, F., and Oballe, C. (2020). "A Bayesian Framework for Persistent Homology." *SIAM Journal on Mathematics of Data Science*, 2(1): 48–74. MR4060450. doi: https://doi.org/10.1137/19M1268719. 3

Martínez, A. F. (2019). "Clustering via Nonsymmetric Partition Distributions." In Antoniano-Villalobos, I., Mena, R. H., Mendoza, M., Naranjo, L., and Nieto-Barajas,

L. E. (eds.), *Selected Contributions on Statistics and Data Science in Latin America*, 69–80. Springer. 10

Martínez, A. F. (2022). "Supplementary Material for "Bayesian Estimation of Topological Features of Persistence Diagrams"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1341SUPP. 10, 13, 14

Martínez, A. F. and Mena, R. H. (2014). "On a Nonparametric Change Point Detection Model in Markovian Regimes." *Bayesian Analysis*, 9(4): 823–858. MR3293958. doi: https://doi.org/10.1214/14-BA878. 9

Mileyko, Y., Mukherjee, S., and Harer, J. (2011). "Probability measures on the space of persistence diagrams." *Inverse Problems*, 27. MR2854323. doi: https://doi.org/10.1088/0266-5611/27/12/124007. 3

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Clustering and Feature Allocation*, 145–174. Springer International Publishing. 8, 10

Munch, E., Turner, K., Bendich, P., Mukherjee, S., Mattingly, J., and Harer, J. (2015). "Probabilistic Fréchet means for time varying persistence diagrams." *Electronic Journal of Statistics*, 9: 1173–1204. MR3354335. doi: https://doi.org/10.1214/15-EJS1030. 3

Nanda, V. and Sazdanović, R. (2014). *Simplicial Models and Topological Inference in Biological Systems*, 109–141. Springer Berlin Heidelberg. MR3204630. doi: https://doi.org/10.1007/978-3-642-40193-0_6. 4

Nicolau, M., Levine, A. J., and Carlsson, G. (2011). "Topological based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival." *Proceeedings of the National Academy of Sciences*, 108: 7265–7270. 2

Niyogi, P., Smale, S., and Weinberger, S. (2011). "A topological view of unsupervised learning from noisy data." *SIAM Journal on Computing*, 40: 646–663. MR2810909. doi: https://doi.org/10.1137/090762932. 2

Offroy, M. and Duponchel, L. (2016). "Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry." *Analytica Chimica Acta*, 910: 1–11. 2

Otter, N., Porter, M., Tillmann, U., Grindod, P., and Harrington, H. (2017). "A roadmap for the computation of persistent homology." *EPJ Data Science*, 6(17). MR4000203. doi: https://doi.org/10.1137/18M1224350. 2, 4

Oudot, S. Y. (2015). *Persistence Theory: From Quiver Representations to Data Analysis*. AMS Mathematical Surveys and Monographs. MR3408277. doi: https://doi.org/10.1090/surv/209. 2

Parida, L., Utro, F., Yorukoglu, D., Carrieri, A. P., Kuhn, D., and Basu, S. (2015). "Topological signatures for population admixture." *Research in Computational Molecular Biology*, 261–275. MR3354646. doi: https://doi.org/10.1007/978-3-319-16706-0_27. 2

Quintana, F. A. (2006). "A predictive view of Bayesian clustering." *Journal of Statistical Planning and Inference*, 136(8): 2407–2429. MR2279815. doi: https://doi.org/10.1016/j.jspi.2004.09.015. 8

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian clustering and product partition models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574. MR1983764. doi: https://doi.org/10.1111/1467-9868.00402. 8

Savica, A., Tothb, G., and Duponchelc, L. (2017). "Topological data analysis (TDA) applied to reveal pedogenetic principles of European topsoil system." *Science of The Total Environment*, 586: 1091–1100. 2

Shotwell, M. S. and Slate, E. H. (2011). "Bayesian Outlier Detection with Dirichlet Process Mixtures." *Bayesian Analysis*, 6(4): 665 – 690. MR2869961. doi: https://doi.org/10.1214/11-BA625. 8

Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., and Ringach, D. L. (2008). "Topological analysis of population activity in visual cortex." *Journal of Vision*, 8(11): 1–18. 2

Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2014). "Féchet means for distributions of persistence diagrams." *Discrete & Computational Geometry*, 52: 44–70. MR3231030. doi: https://doi.org/10.1007/s00454-014-9604-7. 3

Wade, S. and Ghahramani, Z. (2018). "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)." *Bayesian Analysis*, 13(2): 559 – 626. MR3807860. doi: https://doi.org/10.1214/17-BA1073. 10

Wade, S., Walker, S. G., and Petrone, S. (2014). "A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting." *Scandinavian Journal of Statistics*, 41(3): 580–605. MR3249418. doi: https://doi.org/10.1111/sjos.12047. 8, 9

Wang, B., Summa, B., Pascucci, V., and Vejdemo-Johansson, M. (2011). "Branching and circular features in high dimensional data." *IEEE Transactions on Visualization and Computer Graphics*, 17: 1902–1911. 2

Wang, H., Bah, M. J., and Hammad, M. (2019). "Progress in Outlier Detection Techniques: A Survey." *IEEE Access*, 7: 107964–108000. 8

Wasserman, L. (2018). "Topological Data Analysis." *Annual Review of Statistics and Its Application*, 5(1): 501–532. MR3774757. doi: https://doi.org/10.1146/annurev-statistics-031017-100045. 4

Xia, K. L. and Wei, G. W. (2014). "Persistent homology analysis of protein structure, flexibility and folding." *International Journal of Numerical Methods in Biomedical Engineering*, 30: 814–844. MR3247713. doi: https://doi.org/10.1002/cnm.2655. 2

Xu, L., Zheng, Y., and Dongyun, Y. (2012). "A fast algorithm for constructing topological structure in large data." *Homology, Homotopy and Applications*, 14: 221–238. MR2954674. doi: https://doi.org/10.4310/HHA.2012.v14.n1.a11. 2

Yogeshwaran, D. and Adler, R. J. (2015). "On the topology of random complexes built over stationary point processes." *Annals of Applied Probability*, 25(6): 3338–3380. MR3404638. doi: https://doi.org/10.1214/14-AAP1075. 3

Yogeshwaran, D., Subag, E., and Adler, R. J. (2017). "Random geometric complexes in the thermodynamic regime." *Probability Theory and Related Fields*, 167: 107. MR3602843. doi: https://doi.org/10.1007/s00440-015-0678-9. 3

Zhu, X. (2013). "Persistent homology: An introduction and a new text representation for natural language processing." *Proceedings of the 23rd IJCAI, IJCAI13, AAAI Press*, 1953–1959. 2

Zomorodian, A. (2005). *Topology for Computing*. Cambridge University Press. MR2111929. doi: https://doi.org/10.1017/CBO9780511546945. 2, 4

Zomorodian, A. and Carlsson, G. (2005). "Computing Persistent Homology." *Discrete & Computational Geometry*, 33(2): 249–274. MR2121296. doi: https://doi.org/10.1007/s00454-004-1146-y. 2

Acknowledgments

The author is very grateful to Professors Victor Pérez-Abreu, Rolando Biscay and Miguel Nakamura for introducing him to this new world of topological data analysis and for all their support during his postdoctoral stay at CIMAT. This acknowledgment also goes to the anonymous referees for their constructive comments that improved the quality of this paper, and to Professor Ramsés H. Mena for his valuable comments and suggestions.