

Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

2016 - 28th Annual Conference Proceedings

TOPOLOGICAL METHODS FOR THE QUANTIFICATION AND ANALYSIS OF COMPLEX PHENOTYPES


Patrick S. Medina

Purdue University, medinap@purdue.edu

Rebecca W. Doerge

Purdue University, doerge@purdue.edu

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

 Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Medina, Patrick S. and Doerge, Rebecca W. (2016). "TOPOLOGICAL METHODS FOR THE QUANTIFICATION AND ANALYSIS OF COMPLEX PHENOTYPES," *Conference on Applied Statistics in Agriculture*.
<https://doi.org/10.4148/2475-7772.1484>

This Event is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

TOPOLOGICAL METHODS FOR THE QUANTIFICATION AND ANALYSIS OF COMPLEX PHENOTYPES

PATRICK S. MEDINA & R.W. DOERGE

ABSTRACT. Quantitative Trait Locus (QTL) mapping of complex traits, such as leaf venation or root structures, require the phenotyping and genotyping of large populations. Sufficient genotyping is accomplished with cost effective high-throughput assays, however labor costs often makes sufficient phenotyping prohibitively limited. In order to develop efficient high-throughput phenotyping platforms for complex traits algorithms and methods for quantifying these traits are needed. It is often desirable to study the spatial organization of these phenotypes from the images generated by high-throughput platforms. With the goal of quantifying the traits, many approaches try to identify several core traits useful in describing the phenotypic morphology. This simplification may lose important information about the phenotype. Rather than reducing the structural information, we introduce a novel method, the Persistence Intensity Array, for studying complex traits using tools from the emergent field of Topological Data Analysis. This approach uses the complete geometry of the phenotype and represents it as a simpler summary of the key topological shape features contained in the data. We demonstrate this method's efficacy by through a simulated QTL analysis.

1. INTRODUCTION

Advances in technology such as imaging, computing, robotics, and unmanned aerial vehicles has given rise to the widespread development of high-throughput plant phenotyping platforms. While these advances have closed the gap between the amount of phenotypic and genotypic information collected for statistical analysis, accurately quantifying complex phenotypes as data to be used in statistical methods remains a challenge. Toward this end, a common approach to quantifying complex phenotypes is scoring. Specifically, phenotypes with common attributes are assigned an integer value. For example, a diseased root structure may be assigned a value of one, while a healthy root structure is assigned a value of zero. Unfortunately, scoring phenotypes this way results in a loss of potentially important structural information. Further, methods of scoring may not be consistent across experiments or between the people who provide the evaluation.

A more current approach to high-throughput phenotyping, developed for plant roots, uses imaging software to compute the statistical distribution for the data that represent certain trait characteristics across the entire structure [1]. While this quantification approach is reproducible and consistent across experiments, it still reduces the amount of structural information acquired about each phenotype. In addition, since the approach in [1] is focused on plant root structures, it does not easily extend to other complex phenotypes.

In this paper we propose a novel approach for quantifying complex phenotypes that we call the “persistence intensity array;” it is reproducible, consistent across phenotypes, and can be easily implemented on a wide range of phenotypes. The proposed method is built on a new area of applied mathematics known as Topological Data Analysis (TDA) [2, 3]; it quantifies key topological shape features in data, and sets the stage for statistical inference. Here we introduce the basic ideas of TDA and the persistence intensity array, and demonstrate its effectiveness in quantifying complex phenotypes for eventual statistical analysis. Because we are developing a new methodology, we rely on simulated data for evaluation of the proposed approach.

1.1. Introduction to Topological Data Analysis. TDA is a new area and culmination of applied mathematics, statistics, and machine learning that studies the key topological shape attributes of a given collection of data for the purpose of high dimensional visualization, classification, and statistical inference on shapes. A brief overview of the TDA workflow and its’ use in phenotyping is provided. A thorough introduction to TDA can be found in [4] and in the review papers [3, 5]. A conceptual introduction to TDA may be found in [6].

The main goal of TDA is to quantify shape features of an object so that the shape can be analyzed quantitatively. To do this, TDA uses simplicial homology, a subject of algebraic topology [7]. The shape features that are quantified are the 0^{th} homology group (i.e., connected components or clusters), the 1^{st} homology group (i.e., holes), the 2^{nd} homology group (i.e., voids), and their higher dimensional equivalents. For each feature, simplicial homology calculates the number of unique features that appear in a mathematical space and quantifies them as a Betti number. Differences in the Betti numbers of two mathematical spaces allows one to conclude that the spaces are topologically distinct. This fact provides a means for researchers to distinguish between two shapes. For example, Figure 1(a) illustrates a two-dimensional representation of a disk (left) and an annulus (right). Both shapes have a single connected component and the Betti number for the number of connected components, β_0 , is one. The annulus has a single large hole in the center and the disk does not. The Betti number for the number of holes, β_1 , in the annulus is one, while it is zero for

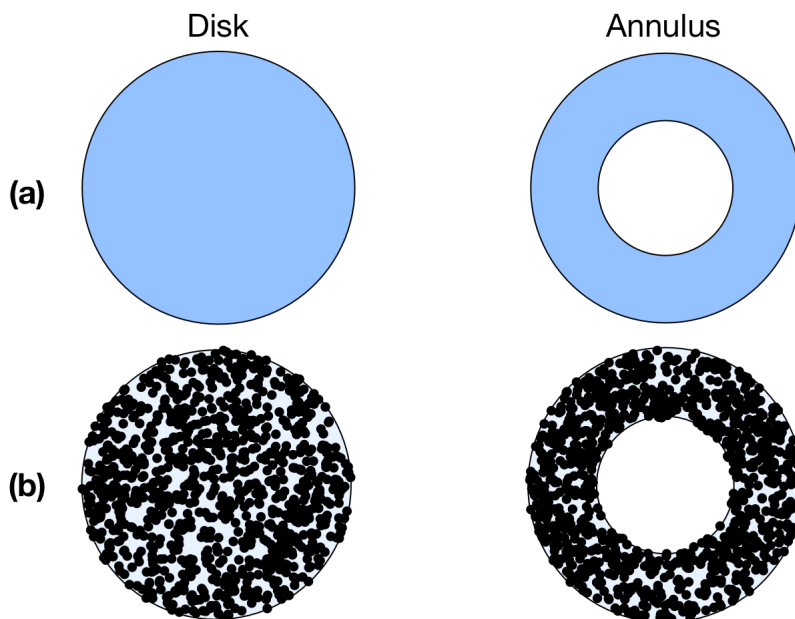


FIGURE 1. (a) A disk (left) and an annulus (right) embedded in \mathbb{R}^2 . These two shapes are topologically distinct, with Betti numbers $\beta_0 = 1$, $\beta_1 = 0$ for the disk, and $\beta_0 = 1$, $\beta_1 = 1$ for the annulus. (b) Samples drawn uniformly from the disk (left) and the annulus (right) using accept-reject sampling.

the disk. Hence, we can distinguish between these two spaces with only knowledge of their Betti numbers.

To compute homology of the underlying shape from a discrete collection of data, methods for constructing a representation of the shape are used. Common methods for achieving this are simplicial complexes [5], or approximating functions [8], such as the kernel density estimator. A commonly used simplicial complex used in TDA, the Vietoris-Rips complex, is illustrated in Figure 2(b) and (c), while an example of the kernel density estimator is illustrated in Figure 2(d).

One challenge of these methods for constructing a representation of the underlying shape is their dependency on the choice of parameters used for approximating the shape. Different choices of parameters may result in different homology calculations (Figure 2(b) and (c)). Since the underlying shape is not necessarily known beforehand, knowing which parameter results in an accurate representation of the

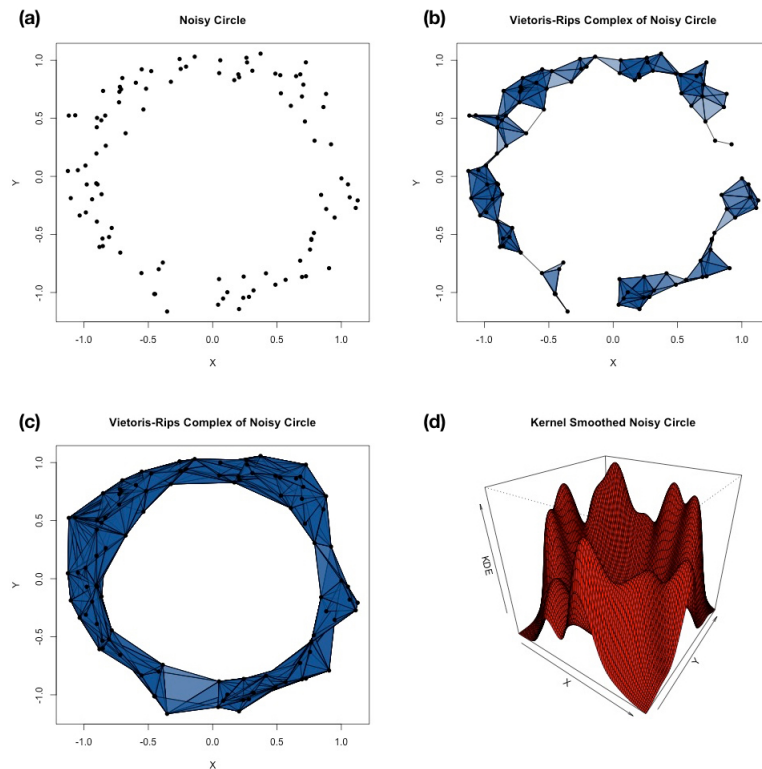


FIGURE 2. (a) One hundred points uniformly sampled from a circle with radius one and Gaussian noise added to each point. (b) The Vietoris-Rips complex is constructed with a parameter of 0.25. From this complex, we observe two connected components and one small hole, which is not consistent with our understanding of the topology from which the underlying points were sampled. (c) The Vietoris-Rips complex constructed with a parameter of 0.50. From this complex, we observe one connected component and one hole, which is consistent with our understanding of the topology of the underlying shape. (d) A kernel smoothed representation of the noisy circle.

underlying space is not known. To resolve this issue, the idea of persistent homology was introduced [2] for the purpose of tracking the change in homology as the parameter changes.

Persistent homology tracks the number and scale topological features present in data by recording the appearance (birth) and disappearance (death) of topological features as a scaling parameter changes (time) [2]. Persistence enables researchers

to study the topological properties of a shape or mathematical space from noisy data. This lays the groundwork for developing statistical methods for determining the topological features in data and for distinguishing between two mathematical spaces. Birth-death information computed for each homology group is stored in the summary statistic called the persistence diagram. A persistence diagram is a multiset of the birth-death values for each unique feature that appears in the data.

Figure 3 presents the persistence diagrams for the disk (left) and the annulus (right). The x -axis denotes the birth time of a feature and the y -axis is the death time. The black line across the diagonal is the line $y = x$. The black dots are the connected components and the red triangles are the holes. For the disk in Figure 3 there is a single connected component born at time zero that persists until a time before 0.40. The other connected components appear later in time and die quickly. In addition, a small hole is born later in time and dies quickly. These features are considered topological noise since their lifetimes are small and not consistent with the true homology of the disk. For the annulus in Figure 3 there is a connected component and a hole born at time zero that persists for a long time. These features are consistent with our knowledge of the true topology of the annulus. The remaining features are topological noise and are due to their small lifetimes and inconsistency with the true topology of the annulus.

Persistence diagrams have been useful for classification and machine learning [9, 10]. However, there are challenges in using the persistence diagram with statistical methods due to the fact that the mean is not guaranteed to be unique nor stable [11]. To overcome these issues, a new summary statistic, the persistence landscape, transforms the persistence diagram into a mathematically more useful space [12]. This said, the Persistence Landscape may be difficult to interpret, so instead, we focus on a growing number of kernel methods for persistence diagrams [10, 13, 14, 15], specifically the persistence intensity function [16].

1.2. The Persistence Intensity Function. The persistence intensity function transforms the persistence diagram to a function space that is more amenable to statistical methods. This is accomplished by applying a kernel smoothing function to the elements of the persistence diagram. In [16], the authors demonstrate that the intensity function contains sufficient topological information to classify different shapes and establish a hypothesis testing procedure for differentiating between samples of intensity functions. The formal definition of the persistence intensity function and theoretical results is found in [16].

2. INTRODUCTION OF THE PERSISTENCE INTENSITY ARRAY

One feature of the intensity function is that it combines all of the available topological information contained in the different homology groups. However, in the context of high-throughput phenotyping, a significant amount of topological information can be documented and knowledge of which topological features are most important for analysis is known from simple empirical observations. For example, if we consider quantifying leaf venation, it is clear that the veins are all connected, however the smaller veins may form intricate loops, which are useful in differentiating the leaves of different plant species. In this situation, more emphasis on the persistent homology of loops may be more relevant than that of the connected components.

Although a weighting mechanism on the different homology groups has been proposed [16], it remains unclear how the weights should be placed. Rather than combining this information in the intensity function and weighting, we propose computing the intensity function for each homology group separately. We call this new statistic the persistence intensity array. The principal difference between the persistence intensity function and the persistence intensity array is that the intensity array defines a function for each homology group. The intensity function can be recovered from the intensity array by summing across the components of the vector. A formal definition of the persistence intensity array and theoretical properties are the subject of future work.

3. SIMULATION STUDY

The motivation and purpose of this paper is to introduce the intensity array as a mechanism for quantifying phenotypes for use in statistical genetics analysis (e.g. quantitative trait locus mapping). Our focus is on demonstrating that the intensity array contains topological information about the underlying spaces from which the data were sampled, and motivating their usefulness for statistical methods in genetics.

3.1. Methods. One feature of using imaging technologies to study complex phenotypes is the vast sources of variation that contribute to the final representation of the phenotype captured by an image. These sources of variation include not only the genetic and environmental factors, but additional sources of variation such as the handling of the phenotype prior to imaging, the imaging technology used, and the image processing steps to isolate a relevant point cloud for use in topological analysis.

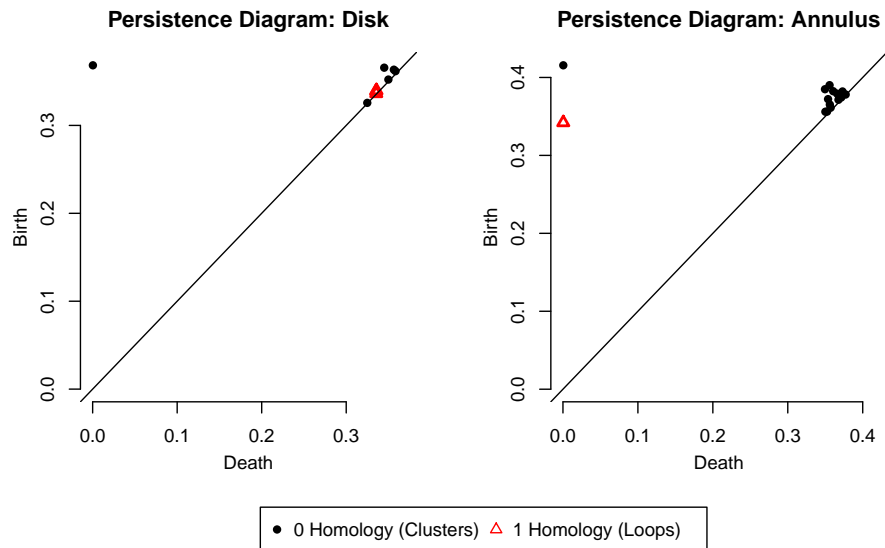


FIGURE 3. Persistence diagrams for the disk (left) and annulus (right). The 0 homology and 1 homology features are represented by a black dot and red triangle respectively. The disk detects a single connected component born at time zero and dies before time 0.40. The annulus detects both a single connected component and hole born at time zero. Both diagrams detect some minor topological features later in time, however these are noise. Both persistence diagrams accurately reflect the topology of the underlying shape from which they were sampled.

For the purpose of simulation, we seek to generate simple shapes that, while they may not reflect the complexity of natural phenotypes, they at least mimic the sources of variability that may be encountered when studying real phenotypes. To simulate the technical variability involved in dealing with the phenotype, imaging and image processing, a uniform sample of 6333 points is drawn from a disk with average radius of one. A uniform sample of 5000 points is drawn from an annulus with an average outer radius of one and inner radius of one half. The sampling was performed using accept-reject sampling [17]. The decision for the number of samples taken from each shape was made so that the shape was accurately reflected in the point cloud, as might be the case with imaging phenotypes, and so that the proportion of points sampled relative to the area of the shape is the same. An illustration of the accept-reject sampling scheme is presented in Figure 1(b). To simulate variability in the physical structures, the radius of the disk is a uniform random variable with support $[0.9, 1.1]$. Similarly, the outer and inner radius of the annulus are uniform random variables with support $[0.9, 1.1]$ and $[0.4, 0.6]$, respectively.

A binary covariate is used to generate the shape, with a value of one generating the annulus and a value of zero generating the disk. For each of the simulations in Sections 3.3 and 3.4 a sample of $2N$ shapes is taken with half of the shapes being an annulus and the remaining disks. The remaining covariates not associated with the shape are independent Bernoulli random variables with a probability of success equal to one half.

Persistence diagrams were computed for each shape using the `gridDiag()` function in the *R* package `TDA` [18]. The point cloud for each shape was smoothed using a kernel density estimator with smoothing parameter 0.15 and persistence homology was computed on the superlevel sets of the kernel smoothed shape using the `Dionysus` library. The advantage of this approach in working with images over other filtrations, such as the Rips complex, is that it helps reduce artificial topological noise and greatly simplifies the number of computations. The ability of this approach to recover information about the true structure of each sample is discussed in Section 3.2. A Gaussian kernel is used to compute the intensity array with a scaling parameter of 0.15. At this time, no method was used to choose the scaling parameter, rather they were chosen based on the appropriate number of features generated. Appropriate methods for choosing this parameter will be explored in future work. Software for the computation of intensity arrays was written in C++ and R. Numerical integration of the intensity array is done with a two dimensional version of the trapezoid rule to obtain a simple estimate. More appropriate numerical integration methods, such as Gaussian quadrature, will be implemented in future versions of the software.

3.2. Data Exploration. We begin our analysis by examining the persistence diagram and the intensity arrays. Figure 3 demonstrates the persistence diagram for a single disk and a single annulus generated as discussed in Section 3.1. The persistence diagram for the disk (left) reveals a single large connected component (black dots) born at time zero and that dies before 0.4. This should be expected since the disk is connected. In addition to the single connected component, we observe several components and holes born later and die relatively quickly. Due to their short lifespan we suspect these points are topological noise since they are not consistent with our knowledge of the true space.¹ The persistence diagram for the annulus (right) reveals a single dominant connected component and a single dominant hole born at time zero. These features are consistent with the geometry of the annulus. Again, we observe some topological noise that is born later in time. While each sample will vary from the persistence diagrams demonstrated in Figure 3, the key topological features are captured in the persistence diagram.

¹Statistical methods for determining if features are topological noise using confidence intervals are discussed in (Chazal et al., 2015) and implemented in the *R* package `TDA`.

TOPOLOGICAL METHODS IN STATISTICAL GENETICS

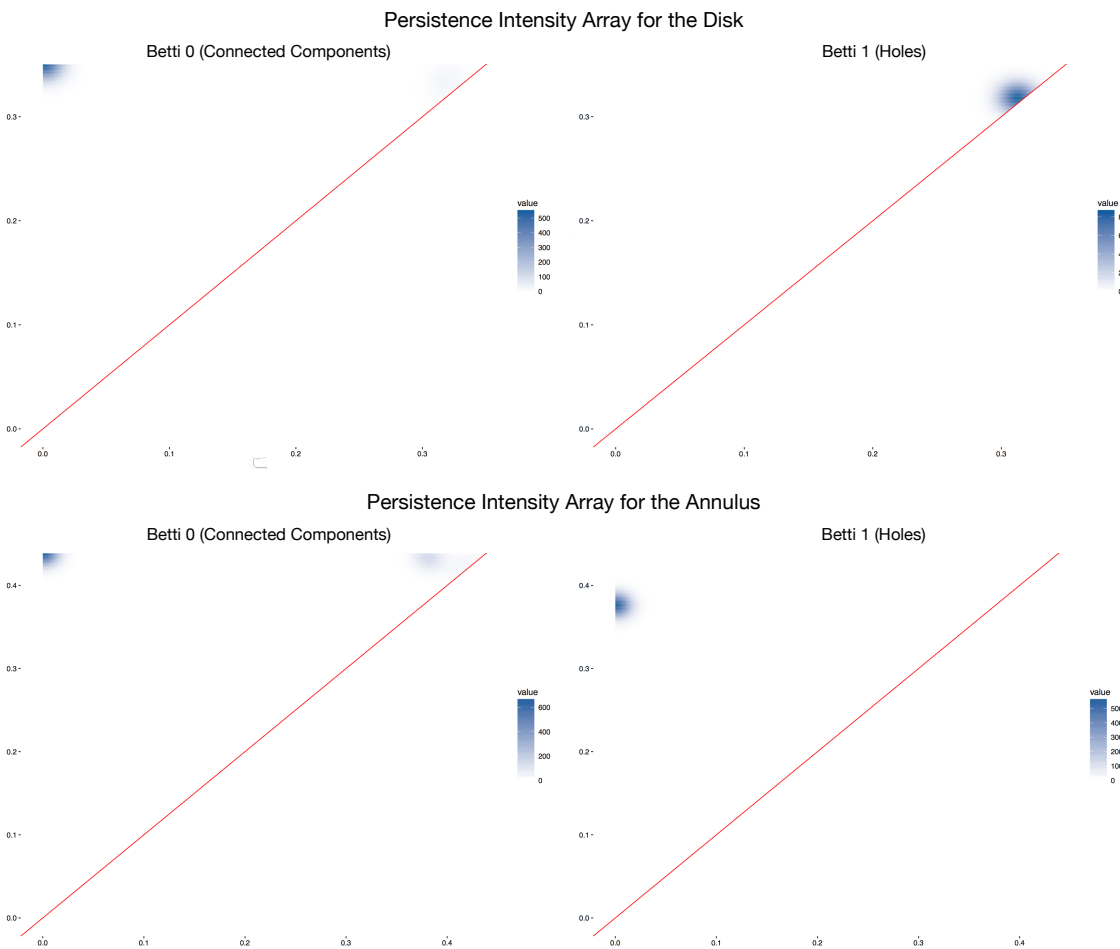


FIGURE 4. **Top:** Persistence Intensity Array for the connected components (left) and holes (right) of the disk. For the intensity array of the connected components, a dense region around the area where a large single connected component appeared in the persistence diagram of the disk. The 0-homology features that appeared later in life and were considered topological noise are not clearly visible. For the intensity array of the holes of the disk, a dark region appears around the area where a hole feature appeared as topological noise. However, the scale of this feature is small relative to the single dominant feature of the connected components. This is consistent with the feature being topological noise. **Bottom:** Persistence Intensity Array for the connected components (left) and holes (right) of the annulus. For the intensity array of the connected components and the holes, a dense region around the area where a large single connected component appeared in the persistence diagram of the disk. Both the 0-homology and 1-homology features that appeared later in life are faint, which is consistent with these features being topological noise.

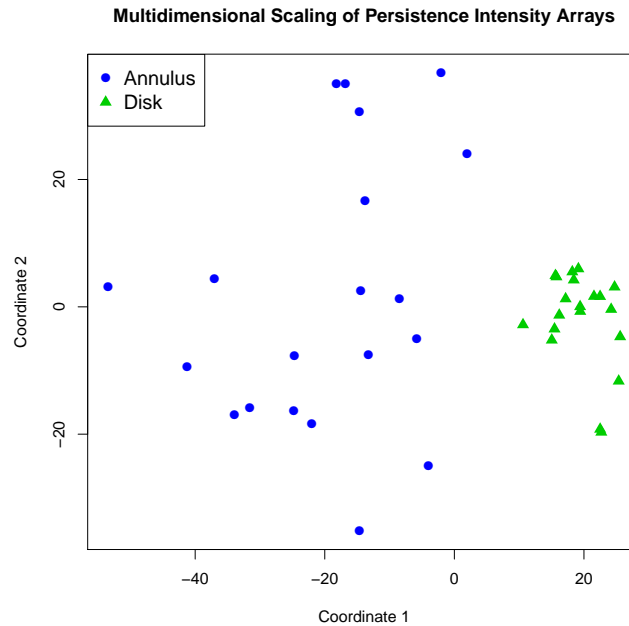


FIGURE 5. Multidimensional scaling of the Persistence Intensity Array for the annulus (blue circles) and the disk (green triangles). The intensity array contains sufficient topological information to be able to distinguish between these two topologically distinct shapes.

3.3. Clustering by Multidimensional Scaling. We continue by examining the intensity array’s ability to cluster topological objects using multidimensional scaling (MDS). Persistence intensity arrays were computed for twenty randomly generated disks and annuli using the methods outlined in Section 3.1. Since the intensity arrays are a vector of functions, the distance between two intensity arrays can be measured as the p -norm of their difference. Formally, if $\hat{\kappa}^1(b, d)$ and $\hat{\kappa}^2(b, d)$ are two intensity arrays, the p -norm distance Δ_p between them is

$$(1) \quad \Delta_p(\hat{\kappa}^1(b, d), \hat{\kappa}^2(b, d)) = \left(\sum_{k=0}^K \int_{\mathbb{R}^2} |\hat{\kappa}_k^1 - \hat{\kappa}_k^2|^p \right)^{1/p}.$$

MDS is performed on the distance matrix constructed between the 40 samples, with $p = 2$. The MDS is clearly able to separate the samples of the disk and annulus into their respective groups (Figure 5). This evidence suggests that the intensity array contains relevant topological information.

3.4. Multivariate Regression. The ability of multivariate regression to detect association between a covariate and the two dimensional shape it generates is tested. The shapes of interest are the disk and annulus. In each simulation, a single covariate is responsible for determining the physical shape of the response. The remaining covariates are binary noise. In these initial simulations we study the effectiveness of multivariate regression in detecting the covariate responsible for generating the shape in two scenarios: an increase in sample size; and as the number of random covariates increase.

Since the intensity array is a function of birth-death locations, we reduce the array to a vector of numbers by integrating the array over \mathbb{R}^2 for the purpose of using multivariate regression. The integrated array may be interpreted as a measure of the total intensity of a topological feature in the j^{th} homology group. The Hotelling-Lawley test is used to detect association between the covariates and response [19].

The example presented here serves as a simplification of the problem encountered in QTL analysis for complex phenotypes. Specifically, we explore the situation in a backcross, B_1 , design [20] to detect an association between the quantified phenotype and the observed marker data. In the scenario outlined at the start of this section, the observed genetic marker responsible for generating the shape of the phenotype may be assumed to be at the QTL. The annulus is generated by a QTL with a heterozygous genotype, while the disk is generated by a QTL with a homozygous genotype. While this example may be a simplification, it demonstrates the ability of the intensity array in detecting relationships between genetic markers and complex phenotypes.

3.4.1. Effect of Sample Size. The first set of simulations tests the effect of sample size on multivariate regression's ability to detect the true covariate responsible for generating the shape of the response variable. Sample sizes used are 14, 20, 50, and 100. In each simulation we observe the true generating covariate to be statistically significant. Tables 1(a-d) in Section 5.1 of the Appendix contain the results for each simulation.

3.4.2. Increasing the Number of Random Covariates. The second set of simulations tests the ability of multivariate regression to detect the true covariate responsible for generating the shape as the number of random covariates increases. This approach is further tested by keeping the sample sizes small. In each simulation we observe the true generating covariate to be statistically significant. Tables 2(a-c) in Section 5.2 of the Appendix contain the results for each simulation.

4. DISCUSSION AND FUTURE WORK

This work introduces the idea of the Persistence Intensity Array, a generalization of the Persistence Intensity Function, and demonstrates its' usefulness in detecting covariates associated with topological characteristics of a shape in a multivariate multiple regression model. In future work, we will formally define the Persistence Intensity Array and study its' statistical properties. Of particular interest will be the statistical properties associated with using the intensity array as a response variable in the multivariate multiple regression model. Although our focus will be on ensuring that statistical hypothesis testing procedures used in regression apply to the intensity array, we remain committed to real world applications that include high-throughput phenotyping and eventual QTL analysis.

5. APPENDIX

The following sections contain the tables of the output for the two scenarios tested in Sections 3.4.1 and 3.4.2 using the `manova()` command in *R*. In each of the tables, T1 is the true covariate that is used to generate the disk or the annulus, while the covariates starting with a "R" is random noise. Within each table, the first column contains the degrees of freedom associated with each covariate, the second column contains the value of the Hotelling-Lawley test statistic, the third column is the approximate *F* statistic, the fourth and fifth columns are the degrees of freedom for the numerator and the denominator of the *F*-distribution, and the sixth column is the *p*-value. In each situation, the null hypothesis is that the regression coefficient associated with each covariate is zero. The regression coefficient of the true covariate is significant in each situation.

5.1. Table of Results for the Effect of Sample Size.

Table 1a: 14 samples; 2 covariates (1 true, 1 random)

	Df	Hotelling-Lawley	approx F	num Df	den Df	Pr(>F)	
T1	1	7.0359	35.180	2	10	2.984e-05	***
R1	1	0.0784	0.392	2	10	0.6856	
Residuals	11						

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1b: 20 samples; 2 covariates (1 true, 1 random)

	Df	Hotelling-Lawley	approx F	num	Df	den	Df	Pr(>F)
T1	1	6.9390	55.512	2	16	6.337e-08	***	
R1	1	0.1774	1.419	2	16	0.2708		
Residuals	17							

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1c. 50 samples; 2 covariates (1 true, 1 random)

	Df	Hotelling-Lawley	approx F	num	Df	den	Df	Pr(>F)
T1	1	3.02602	69.599	2	46	1.224e-14	***	
R1	1	0.00017	0.004	2	46	0.9961		
Residuals	47							

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1d. 100 samples; 2 covariates (1 true, 1 random)

	Df	Hotelling-Lawley	approx F	num	Df	den	Df	Pr(>F)
T1	1	3.5253	169.215	2	96	<2e-16	***	
R1	1	0.0244	1.169	2	96	0.315		
Residuals	97							

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

5.2. Table of Results for Increasing the Number of Random Covariates.

Table 2a. 20 samples; 5 covariates (1 true, 4 random)

	Df	Hotelling-Lawley	approx F	num	Df	den	Df	Pr(>F)
T1	1	7.0610	45.897	2	13	1.284e-06	***	
R1	1	0.1785	1.160	2	13	0.3438		
R2	1	0.2220	1.443	2	13	0.2717		
R3	1	0.3554	2.310	2	13	0.1385		
R4	1	0.2351	1.528	2	13	0.2535		
Residuals	14							

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2b. 20 samples; 10 covariates (1 true, 9 random)

	Df	Hotelling-Lawley	approx F	num	Df	den	Df	Pr(>F)
T1	1	11.2452	44.981	2	8	4.448e-05	***	

R1	1	0.0213	0.085	2	8	0.9190
R2	1	0.2801	1.121	2	8	0.3724
R3	1	0.1290	0.516	2	8	0.6154
R4	1	0.6946	2.778	2	8	0.1213
R5	1	0.6310	2.524	2	8	0.1413
R6	1	0.5536	2.214	2	8	0.1717
R7	1	0.1106	0.442	2	8	0.6574
R8	1	0.2104	0.842	2	8	0.4658
R9	1	0.2865	1.146	2	8	0.3651
Residuals	9					

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2c. 30 samples; 20 covariates (1 true, 19 random)

	Df	Hotelling-Lawley	approx F	num Df	den Df	Pr(>F)	
T1	1	8.8322	35.329	2	8	0.000107	***
R1	1	0.3650	1.460	2	8	0.288083	
R2	1	0.1484	0.594	2	8	0.574885	
R3	1	0.0001	0.000	2	8	0.999553	
R4	1	0.1613	0.645	2	8	0.549833	
R5	1	0.0406	0.162	2	8	0.852880	
R6	1	0.0740	0.296	2	8	0.751483	
R7	1	0.1220	0.488	2	8	0.630966	
R8	1	0.1721	0.688	2	8	0.529828	
R9	1	0.0668	0.267	2	8	0.772137	
R10	1	0.0079	0.031	2	8	0.969193	
R11	1	0.2000	0.800	2	8	0.482303	
R12	1	0.1351	0.540	2	8	0.602434	
R13	1	0.0568	0.227	2	8	0.801625	
R14	1	0.0784	0.314	2	8	0.739310	
R15	1	0.0549	0.220	2	8	0.807435	
R16	1	0.0375	0.150	2	8	0.863137	
R17	1	0.5127	2.051	2	8	0.191004	
R18	1	0.1920	0.768	2	8	0.495385	
R19	1	0.4732	1.893	2	8	0.212314	
Residuals	9						

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

REFERENCES

- [1] A. Bucksch, J. Burridge, L. M. York, A. Das, E. Nord, J. S. Weitz, and J. P. Lynch. Image-based high-throughput field phenotyping of crop roots. *Plant Physiology*, 166(2):470–486, 2014-10-01.
- [2] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. 28(4):511–533, 2002.
- [3] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009-04.
- [4] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, Providence, RI, 2010.
- [5] Robert Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008-01.
- [6] P. S. Medina and R. W. Doerge. Statistical Methods in Topological Data Analysis for Complex, High-Dimensional Data. *ArXiv e-prints*, jul 2016.
- [7] James R. Munkres. *Elements of Algebraic Topology*. Perseus, Reading, MA, 1984.
- [8] Frédéric Chazal, Brittany T. Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.
- [9] Jose A. Perea, Anastasia Deckard, Steve B. Haase, and John Harer. Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. *BMC Bioinformatics*, 16(1), 2015-12.
- [10] Jan Reininghaus, Stefan Huber, Ulrich Bauer, and Roland Kwitt. A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4748, 2015.
- [11] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014-07.
- [12] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015-01.
- [13] R. Anirudh, V. Venkataraman, K. Natesan Ramamurthy, and P. Turaga. A Riemannian Framework for Statistical Analysis of Topological Persistence Diagrams. *ArXiv e-prints*, May 2016.
- [14] G. Kusano, K. Fukumizu, and Y. Hiraoka. Persistence weighted Gaussian kernel for topological data analysis. *ArXiv e-prints*, January 2016.
- [15] Roland Kwitt, Stefan Huber, Marc Niethammer, Weili Lin, and Ulrich Bauer. Statistical topological data analysis—a kernel perspective. In *Advances in Neural Information Processing Systems*, pages 3070–3078, 2015.
- [16] Yen-Chi Chen, Daren Wang, Alessandro Rinaldo, and Larry Wasserman. Statistical analysis of persistence intensity functions. *arXiv preprint arXiv:1510.02502*, 2015.
- [17] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, NY, 2nd ed edition, 2004.
- [18] Brittany T. Fasy, Jisu Kim, Fabrizio Lecci, Clement Maria, Vincent Rouvreau. The included GUDHI is authored by Clement Maria, Dionysus by Dmitriy Morozov, PHAT by Ulrich Bauer, Michael Kerber, and Jan Reininghaus. *TDA: Statistical Tools for Topological Data Analysis*, 2016. R package version 1.5.
- [19] Alvin C. Rencher. *Methods of Multivariate Analysis*. Wiley series in Probability and Mathematical Statistics. J. Wiley, 2nd ed edition, 2002.

- [20] Rongling Wu, Chang-Xing Ma, and George Casella. *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*. Springer, 2007.