

Functional Data Analysis using a Topological Summary Statistic: the Smooth Euler Characteristic Transform

Lorin Crawford^{1,2,3,†}, Anthea Monod^{4,†}, Andrew X. Chen⁴, Sayan Mukherjee^{5,6,7,8}, and Raúl Rabadán⁴

1 Department of Biostatistics, Brown University, Providence, RI, USA

2 Center for Statistical Sciences, Brown University, Providence, RI, USA

3 Center for Computational Molecular Biology, Brown University, Providence, RI, USA

4 Department of Systems Biology, Columbia University, New York, NY, USA

5 Department of Statistical Science, Duke University, Durham, NC, USA

6 Department of Computer Science, Duke University, Durham, NC, USA

7 Department of Mathematics, Duke University, Durham, NC, USA

8 Department of Bioinformatics & Biostatistics, Duke University, Durham, NC, USA

† E-mail: lorin_crawford@brown.edu; am4691@cumc.columbia.edu

Abstract

We introduce a novel statistic, the smooth Euler characteristic transform (SECT), which is designed to integrate shape information into regression models by representing shapes and surfaces as a collection of curves. Due to its well-defined inner product structure, the SECT can be used in a wider range of functional and nonparametric modeling approaches than other previously proposed topological summary statistics. We illustrate the utility of the SECT in a radiomics context by showing that the topological quantification of tumors, assayed by magnetic resonance imaging (MRI), are better predictors of clinical outcomes in patients with glioblastoma multiforme (GBM). We show that SECT features alone explain more of the variance in patient survival than gene expression, volumetric features, and morphometric features.

1 Introduction

The field of radiomics is focused on the extraction of quantitative features from medical images, typically constructed by tomography and digitally stored as shapes or surfaces. The problem of quantifying geometric features from a shape in a way that is amenable to statistical modeling has been a long-standing and fundamental challenge in both statistics as well as radiomics. In this paper, we address this problem by proposing a novel statistic, the smooth Euler characteristic transform (SECT), which is designed to integrate shape information into standard functional regression models. The idea behind the SECT is that physical properties from images and three-dimensional objects can be equivalently represented as a collection of smooth curves. This transformation to a collection of curves is key, because it allows the use of well-developed statistical tools from functional data analysis (FDA) to model shapes. Hence, the statistical contribution of this paper is twofold: (1) a novel summary statistic for shapes, and (2) a functional regression framework which adapts methods from FDA to allow the integration of shape as a covariate. Much of radiomics has been driven by applications in oncology, and in this study we will apply the SECT to a predictive analysis of survival outcomes of patients with glioblastoma multiforme (GBM) — a common glioma that materializes into cancerous tumor growths within the human brain.

The remainder of this paper is organized as follows. In Section 2.1, we explain the theoretical concepts used to develop the SECT and highlight its statistical utility. In Section 2.2, we outline how regression methodologies for functional covariates are naturally suited to model the curves resulting from the SECT. This connection with functional data allows us to specify a general functional kernel regression model that uses shape information and takes on a form that is known to be particularly powerful when conducting predictive inference. In Section 3, we apply our functional kernel regression framework to predict the clinical outcomes of GBM patients using gene expression data, existing morphometric and volumetric tumor image quantifications, and our proposed topological summaries. Here, we perform a comparative study between each covariate type across different models generated by various kernel functions. Finally, in Section 4, we close with a discussion on possible future research.

2 Functional Regression Models with Shape as Covariates

In this section, we will develop a regression method that allows us to include shape information as covariates. There are two key conceptual components in this framework. The first component is specifying a summary statistic, which we think of as a transform that maps shapes into a Hilbert space. This transform has two important properties: the map is injective and the summary statistic admits a well-defined inner product structure. The inner product structure allows us to adapt ideas from functional data analysis to specify a Bayesian regression model that uses the shape summary statistic as predictor variables.

2.1 Summary Statistics for Shape Data

The classic statistical model represents shapes as a collection of landmark points (see Figure 1(a)) (Kendall, 1984; Bookstein, 1997; Dryden and Mardia, 1998). This data representation was implemented partly due to the limited image processing technology of the time. Current imaging technologies have since greatly improved and now allow three-dimensional shapes to be represented as meshes — a collection of vertices, faces, and edges (see Figure 1(b)). Given the advancement in both imaging technology and computational tools to process imaging data, classical landmark-based approaches that previously required user-specification have become less relevant. Methods have since been developed that generate automated geometric morphometrics for mesh representations (e.g. Boyer et al., 2011; Al-Aifari et al., 2013; Lipman and Daubechies, 2011; Boyer et al., 2015). Yet, despite these advancements, both user-specified and automated landmark-based methods for geometric morphometrics are known to suffer from structural errors when comparing shapes that are highly dissimilar (Gao, 2015; Gao et al., 2016).

Most recently, an approach known as the persistent homology transform (PHT) (Turner et al., 2014) was developed to comprehensively address landmark-specification induced issues and maintain robust quantification performance for highly dissimilar and non-isomorphic shapes. The PHT allows for the comparison of shapes without requiring landmarks. However, because the PHT is a collection of persistence diagrams — multiscale topological summaries used extensively in topological data analysis (TDA) — it has a construction that is not directly amenable to generalized functional data models. In this subsection, we will describe the key theoretical concepts of the PHT with a focus on ideas that will be used to introduce our alternative shape summary statistic, the smooth Euler characteristic transform (SECT). Here, we will formally motivate why we propose the SECT for improved statistical utility. Briefly, the SECT is a variant of the PHT that captures the same complete topological and integral geometric information of a shape, but has the advantage of being a continuous, piecewise linear function that is an element in the Hilbert space \mathbf{L}^2 . The Hilbert space structure is well-suited for integration into a larger catalog of standard functional regression methodologies.

Persistent Homology Transform. We will now introduce the minimum theory of topology needed to obtain an intuition for persistence diagrams and persistent homology (Edelsbrunner et al., 2000). For a more detailed review and discussion of these concepts, see the Appendix. The key concept developed in persistent homology is that a great deal of geometry can be captured by a multiscale topological analysis.

The mesh representation of a shape is an example of a mathematical set known as a *simplicial complex* which has been studied extensively in computational geometry and topology. For the purposes of this paper, a simplicial complex can be thought of as a combinatorial representation of the geometry encoded in a mesh. In the context of our paper, the concept of *homology* is an abstract measure of a surface or shape. Homology typically is indexed by integers with the 0th-degree homology capturing the number of connected components in the shape, the 1st-degree homology capturing the number of loops, and the 2nd-degree homology capturing the number of voids. The notation we use in this paper specifies the k^{th} homology group for a simplicial complex K as $H_k(K)$, which corresponds to the collection of the k -dimensional holes of the simplicial complex. Here, 0-holes are a collection of connected components, for example. The key utility of *persistent homology* is to study the homology of a shape at different scales or resolutions as a tool to extract geometric information. The tool used to examine a shape at different scales is called a *filtration*. A filtration is a collection of simplicial complexes $\{K_t\}$ where the index t induces totally ordered sets, or $K_i \subseteq K_j$ for $i < j$. In persistent homology, the filtration $\{K_t\}$ encodes the shape at different resolution scales t , and tracks how the homology groups $H_k(K_t)$ of the filtration changes with t . More specifically, H_0 tracks the connected components, H_1 tracks the loops, and H_2 tracks the voids. The final output of computing

persistent homology is a collection of intervals — where each interval represents a topological feature that is “born” at the parameter value given by the left endpoint of the interval, and “dies” at the value at the right endpoint. The length of the interval corresponds to how long the topological feature persists. The representation of the intervals we consider is called a *persistence diagram*, which treats the ordered pair of the start and end of each interval as a point in the plane — where the x -axis corresponds to birth time and the y -axis is the death time. Thus, one can consider a persistence diagram as a collection of points above the diagonal, together with the set of points on the diagonal having infinite multiplicity (for regularity conditions). Again, for a more detailed review and discussion of these concepts, see the [Appendix](#).

We now formally define the persistent homology transform. Let M be a closed, compact subset (shape) of \mathbf{R}^d that can be written as a finite simplicial complex K . For any unit vector over the unit sphere $\nu \in S^{d-1}$, we define a filtration $K(\nu)$ of K parameterized by a height r as

$$K(\nu)_r = \{x \in K : x \cdot \nu \leq r\}, \quad (1)$$

where $\nu \in S^{d-1}$ is a unit vector. The parameter height function $r_\nu(x)$ is then defined as

$$\begin{aligned} r : \mathbf{R}^d &\rightarrow \mathbf{R} \\ \{x, \nu\} &\mapsto x \cdot \nu \end{aligned} \quad (2)$$

for the same fixed unit $\nu \in S^{d-1}$. Figure 2 depicts an example of a filtration by a given height function. The k^{th} dimensional persistence diagram $X_k(K, \nu)$ summarizes how the topology of the filtration $K(\nu)$ changes over the height parameter r . The filtration in this case is a sublevel set filtration — meaning, as r increases in the direction ν , more and more of the total shape is revealed and included in the persistence information.

Definition 1 (Turner et al. (2014)). The persistent homology transform (PHT) of $K \subset \mathbf{R}^d$ is the function

$$\begin{aligned} \text{PHT}(K) : S^{d-1} &\rightarrow \mathcal{D}^d \\ \nu &\mapsto (X_0(K, \nu), X_1(K, \nu), \dots, X_{d-1}(K, \nu)). \end{aligned}$$

The PHT measures the change in homology by the height filtration over all directions on the unit sphere. There are two key reasons why the PHT is useful. To illustrate the first, denote the space of persistence diagrams as \mathcal{D} (Mileyko et al., 2011; Turner et al., 2014; Bubenik, 2015), and consider two diagrams $X, Y \in \mathcal{D}$. The *Wasserstein distance* between the diagrams X and Y is given by

$$\text{dist}(X, Y) = \inf_{\varphi: X \rightarrow Y} \sum_{x \in X} \|x - \varphi(x)\|,$$

where φ is a bijection between points in X and Y . Let \mathcal{M}_d be the space of subsets of \mathbf{R}^d that can be represented as finite simplicial complexes. The PHT can then be used to define the following distance metric between shapes or surfaces

$$\text{dist}_{\mathcal{M}_d}^{\text{PHT}}(K_1, K_2) := \sum_{k=0}^d \int_{S^{d-1}} \text{dist}(X_k(K_1, \nu), X_k(K_2, \nu)) d\nu.$$

Note that this specification allows for comparisons and similarity studies between shapes. A vectorization of topological characteristics to solve this same problem has also been previously proposed (Carrière et al., 2015). The second key reason why the PHT is useful is that the transformation is injective and thus it preserves information about the shape. Specifically, it has been shown that the PHT is injective when the domain is \mathcal{M}_d for $d = 2, 3$ (Turner et al., 2014). In practice, the PHT is not computed using all directions on the sphere. The stability properties of persistent homology justifies that the use of finitely many directions provides a close approximation to the integral in the equation above.

While the PHT allows for a natural metric structure to compare shapes, it does not admit a simple inner product structure, since it is an infinite collection of persistence diagrams. More specifically, the geometry of the space of persistence diagrams is known to be an Alexandrov space with curvature bounded from below (Burago et al., 1992). This space does not have unique geodesics and therefore quantities such as the Fréchet

mean are not unique (Turner et al., 2014). The complicated structure of the PHT and the geometry of the space of persistence diagrams are an impediment to using the PHT in most statistical models. The key idea motivating the SECT is to provide a quantitative summary of shapes that is injective like the PHT, but also admits a well-defined inner product structure. Having a statistic that has such a structure will allow us to use standard statistical models and adapt methods used in functional data analysis (FDA). For select covariance functions (e.g. Reininghaus et al., 2015; Kwitt et al., 2015; Kusano et al., 2017) the PHT can be adapted to nonparametric statistical models, but this class is notably limited.

Smooth Euler Characteristic Transform. The result of the SECT is a collection of continuous, piecewise linear functions that can be considered an element in the Hilbert space \mathbf{L}^2 . The corresponding inner product structure allows us to apply the SECT to a much broader set of statistical methodologies. The SECT is based on the Euler characteristic (EC), a topological invariant that appears in many branches of mathematics. In terms of homology and persistent homology, the EC counts the ranks of the homology groups (i.e. Betti numbers, β_k) in an alternating sum.

Definition 2. The Euler characteristic (EC) χ for a finite simplicial complex K^d for $d = 3$ is defined by:

$$\chi(K^3) = V - E + F, \quad (3)$$

where V , E , and F are the numbers of vertices, edges, and faces, respectively.

An EC curve is constructed by tracking the progression of the EC as a function with respect to a filtration. In the context of closed compact subsets (shapes) $M \subset \mathbf{R}^3$ represented by finite simplicial complexes K , we consider the filtration in (1) via the height function in (2), and denote the extremal heights from this filtration in the direction $\nu \in S^2$ by

$$\begin{aligned} a_\nu &:= \min\{r_\nu(x), x \in K\}, \\ b_\nu &:= \max\{r_\nu(x), x \in K\}. \end{aligned}$$

Definition 3. Let K_ν^x denote the simplicial complex that represents the closed compact subset $M_\nu^x \subseteq M \subset \mathbf{R}^3$ generated by the sublevel set filtration in (1), defined by the height function $r_\nu(x)$ in (2) for varying x and fixed unit direction vector $\nu \in S^2$. The *EC curve* is defined by

$$\begin{aligned} \chi_\nu^K : [a_\nu, b_\nu] &\rightarrow \mathbf{Z} \subset \mathbf{R} \\ x &\mapsto \chi(K_\nu^x). \end{aligned} \quad (4)$$

The EC curve tracks the evolution of the EC up to (and including) the largest subcomplex of M contained in the sublevel set $r_\nu^{-1}((-\infty, x])$. See Figure 3 for an illustrative example. In considering a directional sweep over the surface of the sphere S^2 , and calculating the corresponding EC curves of the finite simplicial complex representations of M for every direction $\nu \in S^2$, the *Euler characteristic transform (ECT)* (Turner et al., 2014) is defined as follows:

$$\begin{aligned} \text{ECT}(M) : S^{d-1} &\rightarrow \mathbf{Z}^{\mathbf{R}} \\ \nu &\mapsto \chi(M(x, \nu)). \end{aligned} \quad (5)$$

Notice that the EC curve (4) and its corresponding ECT (5) are piecewise constant, integer-valued functions. These discontinuities can affect the stability of this representation (see Figure 3(b)). We therefore propose a reformulation of (5) that allows for a summary that can be used in a wider range of statistical analyses. This involves smoothing the function via the following procedure. First, take the mean value of the EC curve $\bar{\chi}_\nu^K$ over $[a_\nu, b_\nu]$. Next, subtract this mean from the value of the EC curve $\chi_\nu^K(x)$ at every $x \in [a_\nu, b_\nu]$. The result is a centered EC curve in the direction $\nu \in S^2$,

$$\begin{aligned} Z_\nu^K : [a_\nu, b_\nu] &\rightarrow \mathbf{R} \\ x &\mapsto \chi_\nu^K(x) - \bar{\chi}_\nu^K. \end{aligned} \quad (6)$$

Here, we denote the value of Z_ν^K to be zero outside the interval $[a_\nu, b_\nu]$. We then integrate the curve to specify the following construct:

Definition 4. The *centered, cumulative Euler characteristic curve* or *smooth Euler characteristic curve* (SEC), for a fixed direction $\nu \in S^{d-1}$, is defined for all $y \in \mathbf{R}$ as

$$\begin{aligned} \text{SEC}(M) : \mathbf{R} &\rightarrow \mathbf{L}^2 \\ F_\nu^K(y) &:= \int_{-\infty}^y Z_\nu^K(x) dx. \end{aligned} \quad (7)$$

The SEC is a continuous, piecewise linear function with compact support $[a_\nu, b_\nu]$ by construction. It is therefore an element of the Hilbert space \mathbf{L}^2 of square integrable functions on \mathbf{R} . The counterpart to Figure 3(b), smoothed by the procedure resulting in the SEC, is visually illustrated in Figure 3(c). We now formally define the smooth Euler characteristic transform.

Definition 5. The *smooth Euler characteristic transform* (SECT) for a simplicial complex K of a shape M is the map

$$\begin{aligned} \text{SECT} : \{K, S^{d-1}\} &\rightarrow L^2[a_\nu, b_\nu] \\ \nu &\mapsto F_\nu^K(b_\nu) \end{aligned} \quad (8)$$

for all $\nu \in S^{d-1}$. Each curve F_ν^K is also an element in the Hilbert space \mathbf{L}^2 . The following metric can therefore be used to define distances between two meshes K_1 and K_2 :

$$\text{dist}_{\mathcal{M}_d}^{\text{SECT}}(K_1, K_2) := \left(\int_{S^{d-1}} \|F_\nu^{K_1} - F_\nu^{K_2}\|^2 d\nu \right)^{1/2}. \quad (9)$$

The advantage of the SECT over the PHT is that SECT summaries are a collection of curves and have a Hilbert space structure — this means that their structure allows for quantitative comparisons using the full scope of functional and nonparametric statistical methodology. The SECT is also an injective map and the following corollary is an immediate consequence of Theorem 3.1 in Turner et al. (2014).

Corollary 1. *The smooth Euler characteristic transform is injective when the domain is \mathcal{M}_d for $d = 2, 3$.*

It is important to note here that enough directions $\nu \in S^{d-1}$ must be taken for this corollary to hold, since for any one fixed direction, it is not true that the EC curve (upon which the SECT construction depends) is injective. An illustration of this fact is depicted in Figure 4. To determine the number of directions to use in practice, we suggest performing a sensitivity analysis with many different combinations of numbers of directions and sublevel sets. In this particular study, we find prediction results (with SECT features as predictor variables) to be reasonably robust to our final choice of numerical parameters (see GBM example in Section 3).

2.2 Bayesian Functional Regression Models

In the previous subsection we specified the SECT which allows us to map shapes into a space that is represented by collection of curves and has a well-defined inner product structure. In this subsection, we will adapt ideas from functional data analysis (FDA) to specify a Bayesian regression model that uses shape summary statistics as covariates. The goal of FDA is to model data that are continuous functions such as curves, response surfaces, or images (e.g. Ferraty and Vieu, 2006; Ramsay, 2006). The key idea here is that these functions can be considered as elements in a Hilbert space for which one can specify statistical models using stochastic processes (Morris, 2015; Wang et al., 2016). In this paper, we will use a class of stochastic processes that is often referred to as kernel regression models or Gaussian processes (e.g. Wahba, 1990, 1997; Preda, 2007; Pillai et al., 2007; Yuan and Cai, 2010).

Functional Kernel Regression. Denote the SECT representation of a shape as $\{F_\nu\}_{\nu=1}^m$ measured over m directions. A generalized functional regression model considers a response variable \mathbf{y} and covariates that are square integrable functions $F_\nu(t)$ on the real interval domain \mathcal{T} where $t \in \mathcal{T}$. Namely, we consider the following

$$\mathbf{y} \sim p(\mathbf{y} | \boldsymbol{\mu}), \quad g^{-1}(\boldsymbol{\mu}) = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \int_{\mathcal{T}} \sum_{\nu=1}^m f(F_\nu(t)) dw(t) + \boldsymbol{\varepsilon} \quad (10)$$

where g is a smooth link function, $\boldsymbol{\eta}$ is vector of linear predictors, dw is a real-valued measure, and f is a smooth operator from \mathbf{L}^2 to \mathbf{R} to be estimated. A common practice is to assume f is linear in the covariates and can be defined via the following parametric form (Müller and Stadtmüller, 2005)

$$\boldsymbol{\eta} = \sum_{\nu=1}^m \langle F_{\nu}(t), \boldsymbol{\beta}_{\nu}(t) \rangle.$$

Notice that unlike traditional linear regression models, $\boldsymbol{\beta}_{\nu}(t)$ here is an unknown smooth parameter function that is also square integrable on the domain \mathcal{T} , with $\langle \cdot, \cdot \rangle$ denoting an inner product. In many applications, the assumption of a linear relationship between the latent predictors $\boldsymbol{\eta}$ and functional covariates $\{F_{\nu}\}_{\nu=1}^m$ may be too restrictive. For example, when modeling the topological landscape of brain tumors (as we will do in Section 3), it is reasonable to assume that interactions between modes of brain activity extend well beyond additivity (Friston et al., 2000). The nonlinear methodology we will use is functional kernel regression.

There are two key characteristics of a kernel regression model. The first key element is a positive definite kernel function, $k : \mathbf{L}^2 \times \mathbf{L}^2 \rightarrow \mathbf{R}$, where again \mathbf{L}^2 is the Hilbert space of the SECT functional covariates such that $F_{\nu}(t) \in \mathbf{L}^2$. The second key element is the reproducing kernel Hilbert space that is induced by the kernel function. Given the eigenfunctions $\{\psi_j\}_{j=1}^{\infty}$ and eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ of the integral operator defined by the kernel function

$$\lambda_j \psi_j(\mathbf{s}) = \int_{\mathcal{T}} k(\mathbf{s}, \mathbf{v}) \psi_j(\mathbf{v}) d\mathbf{v},$$

we can define a reproducing kernel Hilbert space (RKHS) as

$$\mathbf{H} = \overline{\left\{ f \mid f(F_{\nu}(t)) = \sum_{j=1}^{\infty} c_j \psi_j(F_{\nu}(t)) \text{ and } \|f\|_{\mathbf{H}}^2 = \sum_{j=1}^{\infty} c_j^2 / \lambda_j < \infty \right\}},$$

where $\|f\|_{\mathbf{H}}$ is called the RKHS norm. The appeal of the RKHS model is that the minimizer of the following optimization problem

$$\min_{f \in \mathbf{H}} \left[\frac{1}{n} \sum_{i=1}^n L(f(F_{\nu,i}(t)), y_i) + \lambda \|f\|_{\mathbf{H}} \right],$$

where L is a loss function and $\lambda > 0$ denotes a tuning parameter chosen to balance the trade-off between fitting errors and the smoothness of the function, takes on the following form

$$\hat{f}(F_{\nu}(t)) = \sum_{i=1}^n \alpha_i k(F_{\nu}(t), F_{\nu,i}(t)). \quad (11)$$

Thus, the power of the RKHS model is that an infinite-dimensional minimization problem can be turned into an optimization problem over just n parameters. We can specify a probabilistic model based on equations (10) and (11) with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{K})$, where \mathcal{N} denotes the multivariate normal distribution and the matrix \mathbf{K} is a covariance matrix with elements $\mathbf{K}_{ij} = k(F_{\nu,i}(t), F_{\nu,j}(t))$. This would be equivalent to assuming $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{K}^{-1})$, with $\boldsymbol{\eta} = \mathbf{K} \boldsymbol{\alpha}$ representing a functional random effect.

For concerns with computational complexity in high-dimensional data scenarios, one may factor the kernel matrix $\mathbf{K} = \mathbf{U} \mathbf{D} \mathbf{U}^{\top}$ where \mathbf{U} is an $n \times n$ unitary matrix of eigenvectors and $\mathbf{D} = \text{diag}(\delta_1, \dots, \delta_n)$ an $n \times n$ diagonal matrix with the corresponding eigenvalues in descending order. Given this decomposition, we will work with the following model

$$\tilde{\boldsymbol{\eta}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D}), \quad \tilde{\boldsymbol{\eta}} = \mathbf{U} \boldsymbol{\eta}. \quad (12)$$

This representation results in quicker estimation of regression parameters (e.g. Lippert et al., 2011; Zhou and Stephens, 2012, 2014), and is particularly helpful in Bayesian modeling and posterior sampling.

Posterior Inference and Sampling. We now state the complete specification of the functional kernel regression model

$$\tilde{\mathbf{y}} = \tilde{\boldsymbol{\eta}} + \tilde{\boldsymbol{\varepsilon}}, \quad \tilde{\boldsymbol{\varepsilon}} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}) \quad (13)$$

$$\tilde{\boldsymbol{\eta}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D}) \quad (14)$$

$$\sigma^{-2}, \tau^{-2} \sim \mathcal{G}(\kappa_1, \kappa_2), \quad (15)$$

where in addition to previous notation, $\tilde{\mathbf{y}} = \mathbf{U}\mathbf{y}$, $\tilde{\boldsymbol{\varepsilon}} = \mathbf{U}\boldsymbol{\varepsilon}$, and \mathcal{G} denotes a Gamma distribution with κ_1 and κ_2 representing the shape and rate parameters, respectively.

Note that in many applications, the kernel function is indexed by a bandwidth or smoothing parameter h , $k_h(\mathbf{s}, \mathbf{v})$. For example, the Gaussian kernel can be specified as $k_h(\mathbf{s}, \mathbf{v}) = \exp\{-h\|\mathbf{s} - \mathbf{v}\|^2\}$. This bandwidth parameter can be inferred; however, posterior inference over h is slow, complicated, and often mixes poorly (e.g. Liang et al., 2009). For simplicity, we will work with a fixed bandwidth that is chosen via cross validation. In posterior inference we consider the posterior distribution that arises in the limits $\kappa_1 \rightarrow 0$ and $\kappa_2 \rightarrow 0$. While these limits correspond to improper priors, the resulting posteriors are proper. Most importantly, they scale appropriately with shifting or scaling of the outcome vector \mathbf{y} (Servin and Stephens, 2007). In other words, conclusions and posterior inferences will be unaffected by changing the units of measurement of the response variable. This property is desirable in the settings we consider because outcomes are being characterized by topological features, which are quantified by the SECT and are therefore also invariant to scaling and shifting. This intersection strengthens the validity of our approach for incorporating topological quantification properties with functional data analysis.

Given the full model, we use Markov chain Monte Carlo (MCMC) via a Gibbs sampler to obtain approximate samples from the joint posterior of the all parameters given the observed data. Specifically, posterior samples of $\{\tilde{\boldsymbol{\eta}}, \sigma^2, \tau^2\}$ may be obtained by using the following closed form conditional densities:

- (1) $\tilde{\boldsymbol{\eta}} | \tilde{\mathbf{y}}, \omega, \sigma^2, \tau^2 \sim \mathcal{N}(\mathbf{m}^*, \mathbf{V}^*)$ where $\mathbf{m}^* = \tau^{-2} \mathbf{V}^* \tilde{\mathbf{y}}$ and $\mathbf{V}^* = \tau^2 \sigma^2 (\tau^2 \mathbf{D} + \sigma^2 \mathbf{I}_n)^{-1}$;
- (2) $\sigma^2 | \tilde{\mathbf{y}}, \tilde{\boldsymbol{\eta}}, \omega, \tau^2 \sim \mathcal{G}(a^*, b^*)$ where $a^* = n/2$ and $b^* = \tilde{\boldsymbol{\eta}}^\top \mathbf{D}^{-1} \tilde{\boldsymbol{\eta}}/2$;
- (3) $\tau^2 | \tilde{\mathbf{y}}, \tilde{\boldsymbol{\eta}}, \omega, \sigma^2 \sim \mathcal{G}(a^*, b^*)$ where $a^* = n/2$ and $b^* = \tilde{\mathbf{y}}^\top \tilde{\mathbf{y}}/2$.

The matrices in steps (1) and (2) are diagonal so they can be efficiently inverted at each MCMC iteration. This reduces the computational complexity of the algorithm to scale linearly with the number of samples, as opposed to cubically, when dealing with the inversion of a non-diagonal matrix. Iterating the above procedure B times results in a set of sampled draws $\{\tilde{\boldsymbol{\eta}}^{(b)}, \sigma^{2(b)}, \tau^{2(b)}\}_{b=1}^B$ from the posterior. While in this work, we will exclusively be concerned with performing prediction of unobserved outcomes, variable selection in kernel models has recently been established (Crawford et al., 2017), and could easily be adapted to a scenario in which the predictors are assumed to be collections of curves.

Prediction of Unobserved Outcomes. In the case of fully Bayesian kernel regression models, the posterior predictive distribution follows a multivariate normal that depends on a kernel matrix \mathbf{K}^* evaluated at all $n^* = n_S + n_T$ observations, respectively. Thus, whenever new samples are observed, the regression model is respecified and the posterior must to be recomputed (e.g. Liang et al., 2009). Depending on the size of the data, this can be a computationally expensive procedure. In our proposed functional kernel regression model, we will instead predict the outcome for new samples based on an approach developed in Speed and Balding (2014). We use the posterior estimates of the random effects for individuals in S to deterministically predict the random effects of those in set T (Speed and Balding, 2014). Namely,

$$\tilde{\boldsymbol{\eta}}_T^{(b)} = \mathbf{K}_{TS} \mathbf{K}_{SS}^{-1} \tilde{\boldsymbol{\eta}}_S^{(b)}, \quad b = 1, \dots, B.$$

Here, \mathbf{K}_{TS} and \mathbf{K}_{SS} are submatrices that are found by first computing $\mathbf{K}^* = [\mathbf{K}_{SS}; \mathbf{K}_{ST}; \mathbf{K}_{TS}; \mathbf{K}_{TT}]$ and then partitioning \mathbf{K}^* according to the TS and SS subscripts. The implied posterior prediction is then computed similarly to that of parametric models, $\{\mathbf{y}_T^{(b)} = \tilde{\boldsymbol{\eta}}_T^{(b)}\}_{b=1}^B$. With sampled parameters at each iterate, we can generate posterior predictive quantities and Monte Carlo approximations of marginal predictive means across a range of new sample values.

3 Application: Predicting Clinical Outcomes in Glioblastoma

To fully illustrate the utility of the SECT topological summary statistic, we apply the Bayesian functional kernel regression model to a glioblastoma multiforme (GBM) radiogenomic study with two measured clinical outcomes: disease free survival (DFS) and overall survival (OS). Radiogenomics aims to determine the relationship between clinical imaging and functional genomic variation. The aggressive nature of GBM, coupled with the necessity for invasive surgical procedures, makes it difficult to obtain molecular data on the disease. Imaging technologies provide a useful alternative data source and can additionally help with both the earlier detection and monitoring of tumors. Some recent work in radiogenomics has confirmed the utility of imaging data in GBM research, showing that the inclusion of geometric information improves predictions of patient survival outcomes, and can also be used to classify subtypes of the disease (e.g. Clark et al., 2013; Gutman et al., 2013; Mazurowski et al., 2013; Gevaert et al., 2014; Macyszyn et al., 2016). It is important to note, however, that in most of these previous studies, the shape information extracted from cancer images has been limited to gross spatial features (e.g. the presence of multifocal tumors, the location of recurrent lesions, or crude volumetric calculations). The SECT alternatively offers a novel contribution to radiomic research as a topological and integral geometric representation of cancer images. In this section, we will specifically assess whether topological features are better predictors of DFS and OS prognoses than three other key tumor characteristics: (1) gene expression, (2) tumor morphometry, and (3) tumor volume.

3.1 Radiomic and Gene Expression Data

Magnetic resonance images (MRIs) of primary GBM tumors were collected from ~ 40 patients archived by the The Cancer Imaging Archive (TCIA) (Clark et al., 2013), which is a publicly accessible data repository containing medical images of cancer patients with matched genomic and clinical data collected by The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network, 2008). These patients were selected based on two sets of criteria, namely: (1) individuals had post-contrast T1 axial MRIs taken at the time of their diagnosis, and (2) there were matching (mRNA) gene expression data and clinical correlates (e.g. recorded DFS and OS) available on cBioPortal (Gao et al., 2013). There are two key factors that influenced our decision to use this particular subset of samples. First, the T1-weighted MRI with Gadolinium contrast is one of the most commonly-used imaging modalities often implemented to assay lesions with vascular activity (Adin et al., 2015). Second, exclusively using MRIs taken at the time of diagnosis allows us to avoid any potential confounding factors related to treatment effects that may manifest on postoperative imaging (Macyszyn et al., 2016). Note that considering patients that have undergone different treatment regimens could introduce false positives into the model, particularly when analyzing their gene expression levels.

We segmented the TCIA MRIs using a computer-assisted segmentation program (Chen and Rabadan, 2017) to extract tumor lesions from the surrounding brain tissue. Briefly, this algorithm first converts MRI images to a grayscale, and then thresholds to generate binary images. Morphological segmentation is then applied to delineate connected components. More specifically, the program selects contours corresponding to enhanced tumor lesions, which are lighter than healthy brain tissue. For instance, necrosis presents as dark regions nested within the indicated lesion. An example of the raw image obtained from TCIA is given in Figure 5(a), while the final segmented result is given in Figure 5(b).

Data Preprocessing. We now briefly detail the four key tumor characteristics and two clinical outcomes of interest. We also describe the preprocessing of their specific data features:

- **Gene Expression:** Following previous genomic studies, the mRNA gene expression levels of the selected TCGA samples were preprocessed using the MAS5.0 normalization procedure (Irizarry et al., 2003) to correct for potential lab specific batch effects and other possible confounders. The final data set consisted of 9215 genes which passed a pre-specified hybridization accuracy threshold and showed reasonably varying expression across the assay.
- **Morphometric Features:** This study utilizes the same morphometric features outlined in previous imaging studies (Han et al., 2010; Chang et al., 2011). More specifically, these references also detail

the pipeline to extract and compute these features. This resulted in 212 morphometric predictors corresponding to shape and texture.

- **Volumetric Features:** We also consider 5 tumor volumetric measures. The first is the enhancing volume for each MRI slice. This metric is summed over lesions in the multifocal case. The next features we compute are the core volume of the enhancing and necrotic regions, the longest lesion diameter, and the shape factor of the tumor. Here, we define the shape factor to be the longest lesion diameter divided by the diameter of a sphere of the same volume.
- **SECT Topological Summary Statistics:** Each collection of patient MRIs consisted of approximately 23-25 segmented slices of two dimensional greyscale images (with the exact numbers varying slightly by patient). For each individual slice, we predetermine the fixed number of sublevel sets to be 100, and compute a different SECT topological summary statistic for 72 pre-specified directions evenly sampled over the interval $[0, 2\pi]$. Again, these final quantities were chosen after conducting a sensitivity analysis. Finally, we average over the smooth EC curves for a fixed direction across all slices. More specifically, we index a direction ν for a set of slices \mathbf{s} , respectively. Then we compute the SECT as \bar{F}_ν by averaging over all \mathbf{s} slices, where $F_{\nu, s}$ is the smooth EC curve for the s^{th} slice in direction ν . For an image from patient i , we obtain 72 vectors of length 100 which we concatenate into one 7200-vector.
- **Disease Free Survival (DFS):** The period after a successful treatment during which there are no signs or symptoms of the cancer that was treated.
- **Overall Survival (OS):** The entire period after the start of treatment during which the cancer patient is still alive.

It is important to note that DFS is more commonly used over OS in adjuvant cancer clinical trials, because it offers earlier presentation of data (Sargent et al., 2005). This stems from the particular idea that events due to disease recurrence occur earlier than death from disease, and thus are inherently have cleaner signal (Birgisson et al., 2011).

3.2 Prediction Results

We now compare the predictive accuracy for both clinical outcomes when fitting each tumor feature type with the Bayesian functional kernel regression model detailed in Section 2.2. Here, we use two types of metrics to assess performance: (1) root mean squared error of prediction (RMSEP), and (2) the tabulated frequency for which a given data type exhibits the lowest RMSEP, which we denote as Optimal%. To perform prediction analyses for each outcome, we randomly split the data 250 different times into 75% training and 25% out-of-sample test sets. In each case, the clinical outcomes are centered and scaled to have mean 0 and variance 1 to facilitate the interpretation of results.

In order to illustrate the robustness of the SECT, we apply the Bayesian kernel regression method while using a wide range of kernel functions with which we build the covariance matrix \mathbf{K} . The goal is to show that the power of the SECT summary statistic is robust to the choice of kernel function. Specifically, the kernel functions we consider include:

- the Gaussian kernel $k(\mathbf{s}, \mathbf{v}) = \exp\{-h\|\mathbf{s} - \mathbf{v}\|^2\}$;
- the linear kernel $k(\mathbf{s}, \mathbf{v}) = \mathbf{s}^\top \mathbf{v}/p + h$, where p is the number of features collected for a particular data type (e.g. Keerthi and Lin, 2003; Jiang and Reif, 2015);
- the neural network hyperbolic tangent (sigmoid) kernel $k(\mathbf{s}, \mathbf{v}) = \tanh(\mathbf{s}^\top \mathbf{v}/n + h)$, where n is the number of samples observed in the training set (Lin and Lin, 2003);
- the conditionally positive definite log kernel $k(\mathbf{s}, \mathbf{v}) = \log(\|\mathbf{s} - \mathbf{v}\|^h + 1)$ for images (Souza, 2010).

Once again note that each of these kernel functions are indexed by a bandwidth or smoothing parameter h which is selected via cross validation. More specifically, h is chosen from grid search with values between 0 and 5 and step sizes equal to that of 0.1. Here, a value of 0 denotes a rigid kernel function, while 5

represents a smoothed estimator. Note that, as expected, power and predictive accuracies for all covariate types starts to decay for “undersmoothed” kernel functions containing too many spurious data artifacts (e.g. $h \ll 1$). Predictive results were marginally affected for bandwidths $h \geq 1$. This same issue can be seen in the performance of other RKHS based models (e.g. Jones et al., 1996; Crawford et al., 2017). For each Bayesian model, we set the model hyper-parameters to very small numbers $\kappa_1 = 1 \times 10^{-10}$ and $\kappa_2 = 1 \times 10^{-10}$ in order to mirror the limits $\kappa_1 \rightarrow 0$ and $\kappa_2 \rightarrow 0$. We then run a Gibbs sampler for 20,000 MCMC iterations with a burn-in of 10,000 posterior draws. We note that longer MCMC chains had little effect with respect to inference for any of these models. Numerical results for DFS and OS while using the Gaussian kernel are presented in Table 1. Here we present the mean RMSEP and corresponding standard errors across testing split to show how each tumor characteristic performs while taking into account variability. Similar results for the other kernel functions can be found in the Appendix (see Tables A1-A3).

Overall, our study shows that the SECT topological summaries result in the most accurate predictions for survival. More specifically, under the Gaussian kernel function, using the SECT to predict DFS and OS resulted in the lowest average RMSEPs: 0.803 and 0.958, respectively. This led to the SECT being the optimal tumor characteristic 69% of the time for DFS and 42% of the time for OS. There are a few possible explanations for these results. Gene expression is known to be highly variable, particularly in GBM (Verhaak et al., 2010), while physical and shape traits of tumors are comparatively more stable. Thus, volumetric, morphometric, and topological features have an inherent advantage. In our application, the robustness of the SECT to choice of metric is particularly relevant because the geometric structure of the brain is known to be fibrous — meaning that the brain is made up of, and connected by, cerebral fiber pathways (Wedeen et al., 2012). This brings into the question the validity of assuming the usual Euclidean metric when attempting to quantify shape. Both volumetric and morphometric analyses require the specification of a metric and, in the case where the usual assumption of a Euclidean measure does not apply, an appropriate one must be constructed — which is not always a straightforward task. Moreover, in fibrous settings, there is also the possibility for the further requirement of defining a geodesic. Examining topological properties, as opposed to metric-based properties, bypasses these technical difficulties. It also avoids the introduction of statistical confounders associated with erroneous assumptions of metrics, geodesics, or measurement errors. Altogether, incorporating a topological measure that is not based on a metric results in the flexibility to compare tumors of different sizes more seamlessly. Subsequently, this also implicitly allows for comparisons between different stages of the disease without needing to account for time of progression. We hypothesize that these flexible characteristics favor the SECT as a better predictor of prognosis and survival.

One important implication from our results is that there may be correlations between the topological makeup of tumors and the heterogeneity potentially arising from the activation of different molecular recurrence mechanisms. An example of this correlation occurs in certain multicentric and multifocal tumors — tumors with lesions in opposing hemispheres of the brain that arise from the same oncogenic effects — that exhibit heterogeneity within only one hemisphere. This variation can be clinically relevant. Case studies have described scenarios where recurring multicentric tumors appear in only one hemisphere of the brain, taking the form of multifocal lesions (Lee et al., 2017). This suggests that clinical prognostics correlate directly with the multicentricity and multifocality of tumors, and draw direct connections between topological shape traits and the mutation status of oncogenic relapse drivers. Hence, there is evidence that the existence of underlying characteristics of therapeutic resistance and relapse mechanisms of GBM go beyond the simple consideration of proximity (in a geometric sense). For instance, a particular path to recurrence in GBM may be due to ambient effects inherent to a particular hemisphere of the brain. The prediction results we present in this work suggest that the topological features extracted by the SECT may be better than simple geometric summaries at providing insight into biological phenomena at the molecular level.

4 Discussion

In the present study, we developed a topological summary statistic transform which maps shapes into a space that admits an inner product structure that is amendable to standard functional and nonlinear regression models. To demonstrate the practical utility of our approach, we predicted survival of GBM patients using our specified functional kernel regression model. In this application, we compared the predictive accuracy using both molecular biomarkers and shape covariates. The SECT was shown to explain more of the variance

in DFS of patients than all other covariates in a wide variety of models defined by various kernel functions. For the Gaussian kernel function in particular, the SECT outperformed the other measures in accounting for the variance in both DFS and OS.

Despite these results, several interesting future directions and open questions still remain. For example, in the current study, we focus solely on measuring how well topological features predict survival. However, it would be useful to infer which particular spatial regions of the tumor are most relevant to predicting these outcomes. Standard variable selection methods can be used to infer the directions and segments of the Euler curves that are most relevant. An important open problem is to recover, or partially reconstruct, a shape from the SECT summary statistics. To ensure power, utilizing protected data from current consortium studies with a large number of participants would be of high interest (e.g. Gounder et al., 2015). Similarly, the distance measure for the SECT stated in (9) provides a framework for comparing the shapes of tumors, and correlating geometric properties with molecular and clinical features. Understanding the relationship between therapeutic strategies, signaling pathway dependence, and tumor shapes would provide useful information about different forms of GBM and their etiologies. We conjecture that greater general knowledge about tumor shape may help in distinguishing true progression from pseudoprogression. Progression meaning the growth of the tumor itself, while in pseudoprogression the tumor is infiltrated by immune cells and other factors.

Software and Data Availability

Software to compute the SECT from images is publicly available in both R and Matlab code and located on the repository <https://github.com/RabadanLab/SECT>. The MRI images were segmented using the Medical Imaging Interaction Toolkit with augmented tools for segmentation (MITKats), which was written C++ and is also located at <https://github.com/RabadanLab/MITKats>. The Bayesian functional regression model was fit using a combination of the Bayesian approximate kernel regression (BAKR) (Crawford et al., 2017) and Bayesian generalized linear regression (BGLR) (Pérez and de los Campos, 2014) softwares. These are each written out in R and Rcpp code. The segmented tumor images, in addition to the volumetric and morphometric data, are also publicly available on the Rabadán Lab GitHub repository.

Acknowledgements

LC, AM, and RR are supported by the National Cancer Institute Physical Sciences–Oncology Network (NCI PS–ON) under Grant No. 5U54CA193313-02; AM is the PI on Pilot Grant Subaward No. G11124 for research on radiomics and radiogenomics. LC would like to acknowledge the support of start up funds from Brown University. AXC is supported by the Columbia University Medical Scientist Training Program (MSTP). SM would like to acknowledge the support of grants NSF IIS-1546331, NSF DMS-1418261, NSF IIS-1320357, NSF DMS-1045153, and NSF DMS-1613261. This work used a high-performance computing facility partially supported by grant 2016-IDG-1013 (“HARDAC+: Reproducible HPC for Next-generation Genomics”) from the North Carolina Biotechnology Center. The authors wish to thank Mao Li (Donald Danforth Plant Science Center) and Christoph Hellmayr (Duke University) for help with the formulation of code, as well as Francesco Abate (McKinsey & Co.), Katharine Turner (Swiss Federal Institute of Technology), and Jiguang Wang (Hong Kong University of Science and Technology) for helpful conversations and input on a previous version of the manuscript. The authors would also like to acknowledge The Cancer Imaging Archive (TCIA) and The Cancer Genome Atlas (TCGA) initiatives for making the imaging and the clinical data used in this study publicly available. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Appendix

In this section, we provide more formal details on the mathematics underlying the persistent homology concepts introduced in Section 2.1. For a complete discussion on frequently used methodologies in TDA and

applied topology, the interested reader may refer to a detailed literature (e.g. Ghrist, 2008; Carlsson, 2009, 2014).

Simplicial Complexes, Homology and Persistence

In simplicial homology, the shape of simplicial complexes are studied. A k -*simplex* is the convex hull of $k + 1$ affine independent points v_0, v_1, \dots, v_k , and is denoted by $\sigma = [v_0, v_1, \dots, v_k]$. Examples of k -simplices are points, lines, and triangles. The 0-simplex $[v_0]$ is the vertex v_0 , the 1-simplex $[v_0, v_1]$ is the edge between the vertices v_0 and v_1 , and the 2-simplex $[v_0, v_1, v_2]$ is the triangle bordered by the edges $[v_0, v_1]$, $[v_1, v_2]$ and $[v_0, v_2]$.

Definition A1. A simplicial complex K is a countable set of simplices such that:

1. Every face of a simplex in K is also in K ;
2. If two k -simplices σ_1, σ_2 are in K , then their intersection is either empty or a face of both σ_1 and σ_2 .

Given a shape M with a finite simplicial complex representation (mesh) K , a *simplicial k -chain* is a formal linear combination (assumed over \mathbf{Z}_2 in this paper) of k -simplices in K . A set of k -chains forms a vector space $C_k(K)$. The *boundary map* is $\partial_k : C_k(K) \rightarrow C_{k-1}(K)$ is then given by

$$\partial_k([v_0, v_1, \dots, v_k]) = \sum_{j=0}^k [v_0, \dots, v_{-j}, \dots, v_k]$$

with linear extension, where v_{-j} denotes that we drop element j . Elements of $B_k(K) = \text{im } \partial_{k+1}$ are called *boundaries*, and elements of $Z_k(K) = \ker \partial_k$ are called *cycles*.

Definition A2. The k^{th} homology group of M is defined by the quotient group

$$H_k(K) := Z_k(K) / B_k(K).$$

The intuition behind a homology group is that it contains information about the structure of K . The zeroth homology group $H_0(X)$ is generated by elements that represent connected components of X . For example, if X has three connected components, then $H_0(X) \cong \mathbf{Z}_2 \oplus \mathbf{Z}_2 \oplus \mathbf{Z}_2$, where \cong denotes group isomorphism. For $k \geq 1$, the k^{th} homology group $H_k(X)$ is generated by elements representing k -dimensional “holes” or “loops” in X . A k -dimensional hole can be thought of as the result of taking the boundary of a $(k + 1)$ -dimensional body. The ranks of the homology groups (i.e. the number of generators) are called the *Betti numbers*, and are denoted by $\beta_k(X) := \text{rank}(H_k(X))$. The notation $H_*(X)$ refers to all the homology groups simultaneously. In Figure A1, we display the homology of a torus constructed from a simplicial complex.

Persistent Homology

A *filtration* of simplicial complexes \mathcal{K} is a family of spaces $\mathcal{K} = \{K_t\}_{t=a}^b$ such that $K_{t_1} \subseteq K_{t_2}$ if $t_1 < t_2$. As the parameter t increases, the homology of the spaces K_t may change (e.g. components are added and merged, cycles are formed and filled up). The *persistent homology* of \mathcal{K} is denoted by $\text{PH}_*(\mathcal{K})$ and keeps track of the progression of homology groups generated by the filtration. More specifically, the persistent homology contains the information about the homology of the individual spaces $\{K_t\}$, as well as the mappings between the homology of K_{t_1} and K_{t_2} for every $t_1 < t_2$. Note that persistent homology is also equivalently referred to as *persistence*.

Intuitively, the main idea behind persistent homology is to study homology across multiple scales. Rather than restricting ourselves to only one instance of a space, in persistent homology we study the evolution of the topological structure over a filtration of the space. This amounts to beginning with a rigid proximity rule connecting observed data points, and then continuously relaxing this rule — all while studying the corresponding topological progression.

Barcodes and Persistence Diagrams

The persistence of the data is encoded in objects that are parameterizations of these homology groups known as *barcodes*: collections of intervals which correspond to the lifetimes of topological features. The left endpoint of a bar is the *birth time* of an element in $\text{PH}_*(\mathcal{K})$ and can be thought of as the value of t where this element appears for the first time. Conversely, the right endpoint of a bar is the *death time* and represents the value of t where an element vanishes, or merges with another existing element.

Figure A2 illustrates the idea behind persistent homology and provides an example of a barcode. Here, rather than studying the persistence of simplicial complexes directly as described above, the simplicial complexes are constructed from unions of balls around sampled points. We can study the persistence on either the unions of balls or the corresponding formed simplicial complexes, since the Nerve Lemma (Borsuk, 1948) provides homotopy equivalence between the two. In this case, the filtration parameter is the radius r of the balls, (see Figure A2(a)). The underlying space is an annulus from which $n = 50$ samples $P_1, \dots, P_n \in \mathbf{R}^2$ were drawn from a uniform distribution. Unions of closed balls $X_r = \bigcup_i B_r(P_i)$ are then formed around these data points with radii r growing from r_1 through to r_5 , which causes X_r to grow. As this happens, connected components merge, and cycles are formed and then are filled up. The barcode in Figure A2(b) captures a summary of all homology features in this process. Specifically, in this image there are two bars that are significantly longer than the others (one in H_0 and one in H_1) indicating that the underlying space has a single connected component and a single cycle, which is equivalent to homology of the annulus.

Barcodes can therefore be considered as summary statistics of the data generating process, as they sufficiently allow for a reduction in dimension of the ambient space. This information can alternatively be represented by a *persistence diagram*, which takes the birth and death times of each bar in a barcode as an ordered pair (x, y) and produces a scatterplot. This then provides a multi-scale topological summary of the data space. In a persistence diagram, the points lie in \mathbf{R}^2 and all the points on the diagonal $x = y$ have infinite multiplicity. The diagonal is included for technical reasons, and involves defining a metric on the space of diagrams or barcodes. Points on the diagonal may be intuitively thought of as topological noise, since they are points that are born and die immediately.

Since summary statistics are direct parallels to the invariants of a topological space, considering such topological approaches in data analytics is a way of reducing dimensionality in high-dimensional statistical problems (i.e. where the number of predictors is far greater than the number of observations).

Figures and Tables

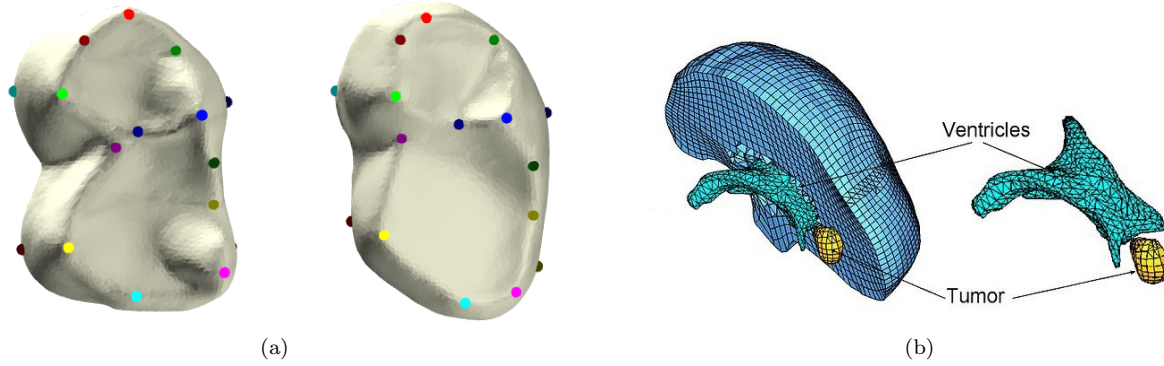


Figure 1: **Classic representations of three-dimensional shapes.** (a) A landmark representation of two molars first published in Boyer et al. (2011). (b) A mesh representation of a brain tumor and ventricles.

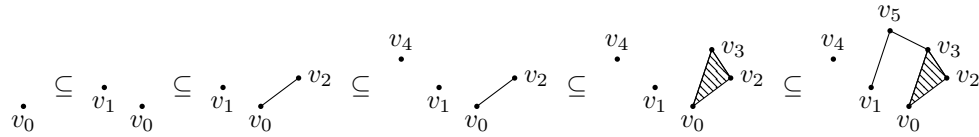


Figure 2: **The filtration of K by height in direction ν .** Each simplex is included at its maximal height. This figure was first published in Turner et al. (2014).

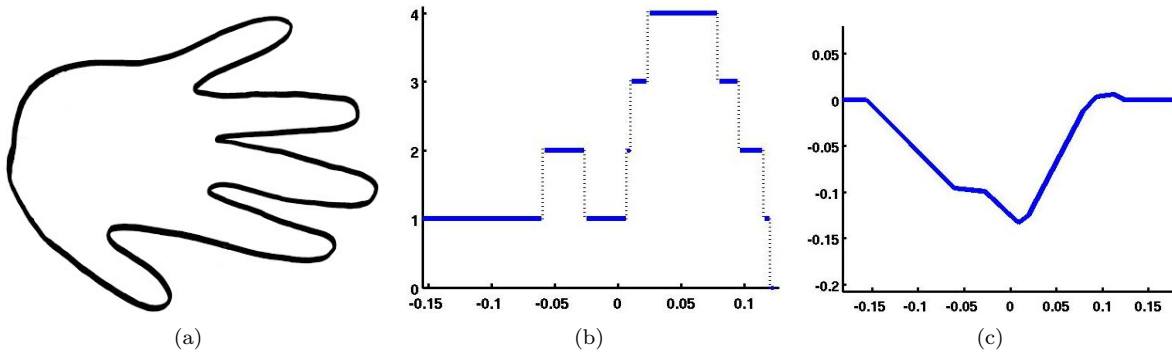


Figure 3: **Illustrative example of the evolutionary tracking of topological features for a given shape.** (a) A 2D contour of a hand. (b) Euler characteristic (EC) curve of the 2D contour of a hand. (c) The associated smooth Euler characteristic (SEC) curve.

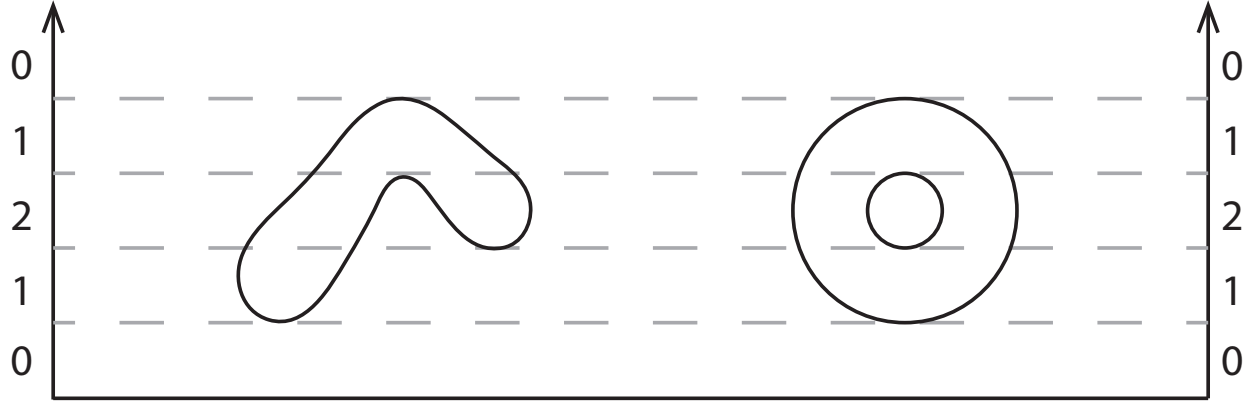


Figure 4: **Counterexample for injectivity of the EC curve for a fixed direction.** The vertical axis on both sides of the figure show the direction of the filtration of both shapes by height (sublevel set filtration). The numbers on the side of the axis show the evolution of the EC for both shapes. We see that although the EC changes in exactly the same manner for the filtration over both shapes, yielding identical ECTs for a fixed direction $\nu \in S^{d-1}$, the shapes that generated the ECTs differ. It is important to note that enough directions must be considered for injectivity of the ECT and SECT to hold.

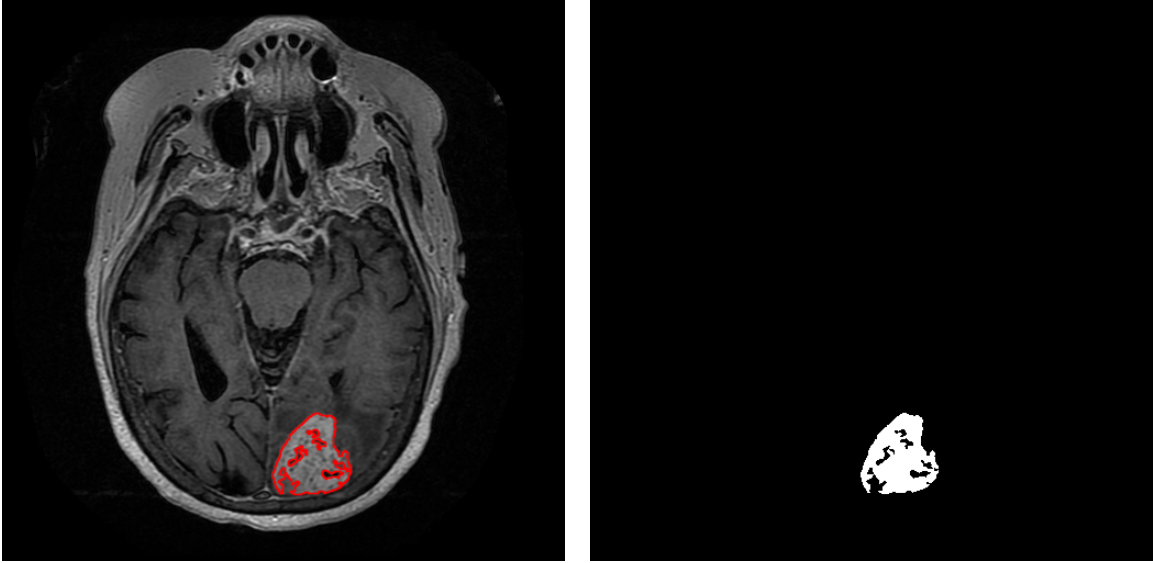


Figure 5: **Example of performance of the segmentation algorithm.** (a) Original MRI from the TCIA and TCGA; (b) Final segmented image via the proposed segmentation algorithm.

	Disease Free Survival		Overall Survival	
Data Type	RMSEP	Optimal%	RMSEP	Optimal%
Gene Expression	0.944 (0.035)	0.20	0.981 (0.030)	0.27
Morphometrics	0.942 (0.035)	0.07	0.965 (0.029)	0.15
Volume	0.939 (0.035)	0.04	0.964 (0.029)	0.16
SECT	0.803 (0.035)	0.69	0.958 (0.028)	0.42

Table 1: **Results for predicting disease free survival (DFS) and overall survival (OS) using the Gaussian kernel function.** The first and third panels show comparisons of root mean squared errors of prediction (RMSEP) for the four considered data types. The second and fourth panels detail the percentage of the time that a model exhibits the lowest RMSEP. This is denoted as Optimal%. All values in bold represent the method with the lowest RMSEP or the method that most frequently performs best, respectively. These values are based on 100 random different 80-20 splits for each clinical outcome. Standard errors for each model are given the parentheses.

Appendix: Figures and Tables

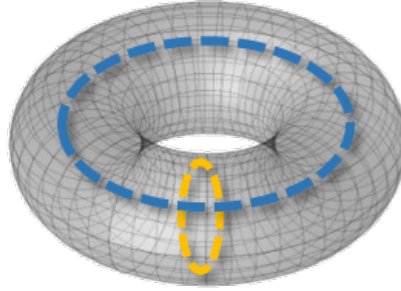


Figure A1: **An illustrative example of homology using the 2-dimensional torus and its cycles.** The torus has a single connected component and a single 2-cycle (the void locked inside the torus). In addition it has two distinct 1-dimensional cycles (or closed loops) represented by the two curves in the figure. Consequently the Betti numbers of the torus are $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$. This figure was first published in Bobrowski et al. (2017).

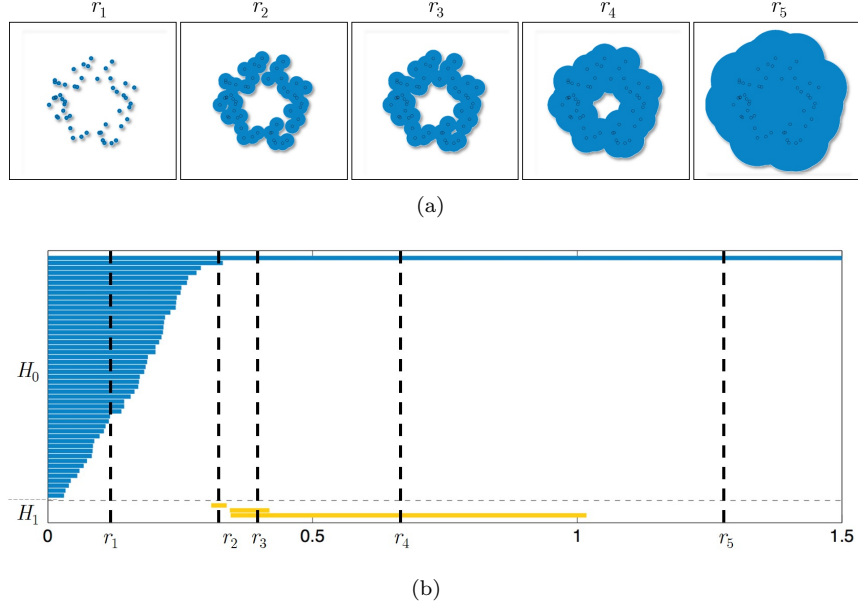


Figure A2: **Illustrative example of persistent homology and corresponding example of a barcode.** (a) X_r is a union of balls of radius r around a random set of $n = 50$ points, generated from a uniform distribution on an annulus in \mathbf{R}^2 . Illustrated are five instances of X_r with different radius sizes. (b) The persistent homology of the filtration $\{X_r\}_{r=0}^\infty$. The x -axis is the radius of the balls, and the bars represent the homology features that are born and die. For H_0 , we observe that at radius zero, the number of components is exactly n and as the radius increases, components merge (or die). The cycles show up later in this process. There are two bars that are significantly longer than the others (one in H_0 and one in H_1). These correspond to the true feature of the annulus. This figure was first published in Bobrowski et al. (2017).

Data Type	Disease Free Survival		Overall Survival	
	RMSEP	Optimal%	RMSEP	Optimal%
Gene Expression	1.001 (0.037)	0.16	1.008 (0.027)	0.18
Morphometrics	1.020 (0.036)	0.13	1.028 (0.027)	0.21
Volume	1.006 (0.036)	0.03	0.977 (0.028)	0.32
SECT	0.992 (0.041)	0.68	1.002 (0.027)	0.29

Table A1: **Results for predicting disease free survival (DFS) and overall survival (OS) using the linear kernel function.** The first and third panels show comparisons of root mean squared errors of prediction (RMSEP) for the four considered data types. The second and fourth panels detail the percentage of the time that a model exhibits the lowest RMSEP. This is denoted as Optimal%. All values in bold represent the method with the lowest RMSEP or the method that most frequently performs best, respectively. These values are based on 100 different random 80-20 splits for each clinical outcome. Standard errors for each model are given the parentheses.

	Disease Free Survival		Overall Survival	
Data Type	RMSEP	Optimal%	RMSEP	Optimal%
Gene Expression	1.026 (0.037)	0.16	1.012 (0.031)	0.20
Morphometrics	1.026 (0.036)	0.07	1.013 (0.031)	0.16
Volume	1.026 (0.037)	0.06	1.013 (0.031)	0.13
SECT	0.963 (0.033)	0.71	1.006 (0.032)	0.51

Table A2: **Results for predicting disease free survival (DFS) and overall survival (OS) using the hyperbolic tangent (sigmoid) kernel function.** The first and third panels show comparisons of root mean squared errors of prediction (RMSEP) for the four considered data types. The second and fourth panels detail the percentage of the time that a model exhibits the lowest RMSEP. This is denoted as Optimal%. All values in bold represent the method with the lowest RMSEP or the method that most frequently performs best, respectively. These values are based on 100 different random 80-20 splits for each clinical outcome. Standard errors for each model are given the parentheses.

	Disease Free Survival		Overall Survival	
Data Type	RMSEP	Optimal%	RMSEP	Optimal%
Gene Expression	0.963 (0.036)	0.15	1.010 (0.03)	0.21
Morphometrics	0.970 (0.037)	0.14	1.034 (0.03)	0.24
Volume	0.984 (0.037)	0.04	1.034 (0.03)	0.11
SECT	0.884 (0.035)	0.67	1.019 (0.03)	0.44

Table A3: **Results for predicting disease free survival (DFS) and overall survival (OS) using the log kernel function.** The first and third panels show comparisons of root mean squared errors of prediction (RMSEP) for the four considered data types. The second and fourth panels detail the percentage of the time that a model exhibits the lowest RMSEP. This is denoted as Optimal%. All values in bold represent the method with the lowest RMSEP or the method that most frequently performs best, respectively. These values are based on 100 different random 80-20 splits for each clinical outcome. Standard errors for each model are given the parentheses.

References

- Adin, M. E., L. Kleinberg, D. Vaidya, E. Zan, S. Mirbagheri, and D. M. Yousem (2015). Hyperintense dentate nuclei on t1-weighted mri: Relation to repeat gadolinium administration. *American Journal of Neuroradiology* 36(10), 1859–1865.
- Al-Aifari, R., I. Daubechies, and Y. Lipman (2013). Continuous Procrustes Distance Between Two Surfaces. *Communications on Pure and Applied Mathematics* 66(6), 934–964.
- Birgisson, H., U. Wallin, L. Holmberg, and B. Glimelius (2011). Survival endpoints in colorectal cancer and the effect of second primary other cancer on disease free survival. *BMC Cancer* 11(1), 438.
- Bobrowski, O., S. Mukherjee, and J. E. Taylor (2017). Topological consistency via kernel estimation. *Bernoulli* 23(1), 288–328.
- Bookstein, F. L. (1997). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.
- Borsuk, K. (1948). On the imbedding of systems of compacta in simplicial complexes. *Fundamenta Mathematicae* 35(1), 217–234.
- Boyer, D. M., Y. Lipman, E. St. Clair, J. Puente, B. A. Patel, T. Funkhouser, J. Jernvall, and I. Daubechies (2011). Algorithms to Automatically Quantify the Geometric Similarity of Anatomical Surfaces. *Proceedings of the National Academy of Sciences* 108(45), 18221–18226.
- Boyer, D. M., J. Puente, J. T. Gladman, C. Glynn, S. Mukherjee, G. S. Yapuncich, and I. Daubechies (2015). A new fully automated approach for aligning and comparing shapes. *The Anatomical Record* 298(1), 249–276.
- Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* 16(1), 77–102.
- Burago, Y., M. Gromov, and G. Perel'man (1992). A.D. Alexandrov spaces with curvature bounded below. *Russian Mathematical Surveys* 47(2), 1–58.
- Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)* 46(2), 255–308.
- Carlsson, G. (2014). Topological pattern recognition for point cloud data. *Acta Numerica* 23, 289–368.
- Carrière, M., S. Y. Oudot, and M. Ovsjanikov (2015). Stable topological signatures for points on 3d shapes. In *Proceedings of the Eurographics Symposium on Geometry Processing, SGP '15*, Aire-la-Ville, Switzerland, Switzerland, pp. 1–12. Eurographics Association.
- Chang, H., G. V. Fontenay, J. Han, G. Cong, F. L. Baehner, J. W. Gray, P. T. Spellman, and B. Parvin (2011). Morphometric analysis of tcga glioblastoma multiforme. *BMC Bioinformatics* 12(1), 484.
- Chen, A. X. and R. Rabadan (2017). Fast and accurate semi-automatic segmentation tool for brain tumor mris. arXiv:1705.06823.
- Clark, K., B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior (2013). The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging* 26(6), 1045–1057.
- Crawford, L., K. C. Wood, X. Zhou, and S. Mukherjee (2017). Bayesian approximate kernel regression with variable selection. *Journal of the American Statistical Association*.
- Dryden, I. and K. Mardia (1998). *Statistical shape analysis*. Wiley Series in Probability and Statistics. Wiley.
- Edelsbrunner, H., D. Letscher, and A. Zomorodian (2000). Topological persistence and simplification. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science, FOCS'00*, Washington, DC, USA. IEEE Computer Society.

- Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.
- Friston, K., J. Phillips, D. Chawla, and C. Buchel (2000). Nonlinear pca: characterizing interactions between modes of brain activity. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 355(1393), 135–146.
- Gao, J., B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cbiportal. *Science Signaling* 6(269), p11.
- Gao, T. (2015). *Hypoelliptic Diffusion Maps and Their Applications in Automated Geometric Morphometrics*. Ph. D. thesis, Duke University.
- Gao, T., G. S. Yapuncich, I. Daubechies, S. Mukherjee, and D. M. Boyer (2016). Development and assessment of fully automated and globally transitive geometric morphometric methods, with application to a biological comparative dataset with high interspecific variation. *bioRxiv*. 086280.
- Gevaert, O., L. a. Mitchell, A. S. Achrol, J. Xu, S. Echegaray, G. K. Steinberg, S. H. Cheshier, S. Napel, G. Zaharchuk, and S. K. Plevritis (2014). Glioblastoma Multiforme: Exploratory Radiogenomic Analysis by Using Quantitative Image Features. *Radiology* 273(1), 131731.
- Grhist, R. (2008). Barcodes: the persistent topology of data. *Bull. Amer. Math. Soc. (N.S.)* 45(1), 61–75.
- Gounder, M. M., L. Nayak, S. Sahebjam, A. Muzikansky, A. J. Sanchez, S. Desideri, X. Ye, S. P. Ivy, L. B. Nabors, M. Prados, S. Grossman, L. M. DeAngelis, and P. Y. Wen (2015). Evaluation of the safety and benefit of phase i oncology trials for patients with primary cns tumors. *Journal of Clinical Oncology* 33(28), 3186–3192.
- Gutman, D. A., L. A. D. Cooper, S. N. Hwang, C. A. Holder, J. Gao, T. D. Aurora, J. William D. Dunn, L. Scarpace, T. Mikkelsen, R. Jain, M. Wintermark, M. Jilwan, P. Raghavan, E. Huang, R. J. Clifford, P. Mongkolwat, V. Kleper, J. Freymann, J. Kirby, P. O. Zinn, C. S. Moreno, C. Jaffe, R. Colen, D. L. Rubin, J. Saltz, A. Flanders, and D. J. Brat (2013). Mr imaging predictors of molecular profile and survival: Multi-institutional study of the tcga glioblastoma data set. *Radiology* 267(2), 560–569. PMID: 23392431.
- Han, J., H. Chang, K. Andarawewa, P. Yaswen, M. H. Barcellos-Hoff, and B. Parvin (2010). Multidimensional profiling of cell surface proteins and nuclear markers. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 7(1), 80–90.
- Irizarry, R., C. Hobbs, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2(4), 249–64.
- Jiang, Y. and J. C. Reif (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768.
- Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* 91(433), 401–407.
- Keerthi, S. S. and C.-J. Lin (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation* 15(7), 1667–1689.
- Kendall, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society* 16(2), 81–121.
- Kusano, G., K. Fukumizu, and Y. Hiraoka (2017). Kernel method for persistence diagrams via kernel embedding and weight factor. *arXiv:1706.03472*.

- Kwitt, R., S. Huber, M. Niethammer, W. Lin, and U. Bauer (2015). Statistical topological data analysis - a kernel perspective. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28*, pp. 3070–3078. Curran Associates, Inc.
- Lee, J.-K., J. Wang, J. K. Sa, E. Ladewig, H.-O. Lee, I.-H. Lee, H. J. Kang, D. S. Rosenbloom, P. G. Camara, Z. Liu, P. van Nieuwenhuizen, S. W. Jung, S. W. Choi, J. Kim, A. Chen, K.-T. Kim, S. Shin, Y. J. Seo, J.-M. Oh, Y. J. Shin, C.-K. Park, D.-S. Kong, H. J. Seol, A. Blumberg, J.-I. Lee, A. Iavarone, W.-Y. Park, R. Rabadan, and D.-H. Nam (2017). Spatiotemporal genomic architecture informs precision oncology in glioblastoma. *Nat Genet* 49(4), 594–599.
- Liang, F., K. Mao, S. Mukherjee, and M. West (2009). Nonparametric Bayesian kernel models. Technical report, Department of Statistical Science, Duke University.
- Lin, H.-T. and C.-J. Lin (2003). A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. Technical report, Department of Computer Science, National Taiwan University.
- Lipman, Y. and I. Daubechies (2011). Conformal Wasserstein Distances: Comparing Surfaces in Polynomial Time. *Advances in Mathematics* 227(3), 1047–1077.
- Lippert, C., J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods* 8(10), 833–834.
- Macyszyn, L., H. Akbari, J. M. Pisapia, X. Da, M. Attiah, V. Pigrish, Y. Bi, S. Pal, R. V. Davuluri, L. Roccograndi, N. Dahmane, M. Martinez-Lage, G. Biros, R. L. Wolf, M. Bilello, D. M. O’Rourke, and C. Davatzikos (2016). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-Oncology* 18(3), 417–425.
- Mazurowski, M. A., A. Desjardins, and J. M. Malof (2013). Imaging descriptors improve the predictive power of survival models for glioblastoma patients. *Neuro-Oncology* 15(10), 1389–1394.
- Mileyko, Y., S. Mukherjee, and J. Harer (2011). Probability measures on the space of persistence diagrams. *Inverse Problems* 27(12), 124007.
- Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application* 2, 321–359.
- Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *Ann. Statist.* 33(2), 774–805.
- Pérez, P. and G. de los Campos (2014). Genome-wide regression and prediction with the bgrr statistical package. *Genetics* 198(2), 483.
- Pillai, N. S., Q. Wu, F. Liang, S. Mukherjee, and R. Wolpert (2007). Characterizing the function space for Bayesian kernel models. *Journal of Machine Learning Research* 8, 1769–1797.
- Preda, C. (2007). Regression models for functional data by reproducing kernel hilbert spaces methods. *Journal of Statistical Planning and Inference* 137(3), 829–840.
- Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.
- Reininghaus, J., S. Huber, U. Bauer, and R. Kwitt (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4741–4748.
- Sargent, D., H. Wieand, D. Haller, R. Gray, J. Benedetti, M. Buyse, R. Labianca, J. Seitz, C. O’Callaghan, G. Francini, A. Grothey, M. O’Connell, P. Catalano, C. Blanke, D. Kerr, E. Green, N. Wolmark, T. Andre, R. Goldberg, and A. De Gramont (2005). Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. *Journal of Clinical Oncology* 23(34), 8664–8670.
- Servin, B. and M. Stephens (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet* 3(7), e114–.

- Souza, C. R. (2010). Kernel functions for machine learning applications. *Creative Commons Attribution-Noncommercial-Share Alike* 3, 29.
- Speed, D. and D. J. Balding (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research* 24, 1550–1557.
- The Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455(7216), 1061–1068.
- Turner, K., Y. Mileyko, S. Mukherjee, and J. Harer (2014). Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry* 52(1), 44–70.
- Turner, K., S. Mukherjee, and D. M. Boyer (2014). Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA* 3(4), 310–344.
- Verhaak, R. G. W., K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O’Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *pdgfra*, *idh1*, *egfr*, and *nf1*. *Cancer Cell* 17(1), 98–110.
- Wahba, G. (1990). *Splines Models for Observational Data*, Volume 59 of *Series in Applied Mathematics*. Philadelphia, PA: SIAM.
- Wahba, G. (1997). Support Vector Machines, reproducing kernel Hilbert spaces and the randomized GACV. *Neural Information Processing Systems (NIPS)* 6, 69–87.
- Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual Review of Statistics and Its Application* 3, 257–295.
- Wedeen, V. J., D. L. Rosene, R. Wang, G. Dai, F. Mortazavi, P. Hagmann, J. H. Kaas, and W.-Y. I. Tseng (2012). The geometric structure of the brain fiber pathways. *Science* 335(6076), 1628.
- Yuan, M. and T. T. Cai (2010). A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics* 38(6), 3412–3444.
- Zhou, X. and M. Stephens (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44(7), 821–825.
- Zhou, X. and M. Stephens (2014). Efficient multivariate linear mixed model algorithms for genomewide association studies. *Nature Methods* 11(4), 407–409.