JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–13: Applications and Case Studies
https://doi.org/10.1080/01621459.2023.2298031

Bayesian Landmark-Based Shape Analysis of Tumor Pathology Images

Cong Zhang[a], Tejasv Bedi[a], Chul Moon[b], Yang Xie[c], Min Chen[a], and Qiwei Li[a]

[a]Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX; [b]Department of Statistics and Data Science, Southern Methodist University, Dallas, TX; [c]Quantitative Biomedical Research Center, School of Public Health, The University of Texas Southwestern Medical Center, Dallas, TX

ABSTRACT

Medical imaging is a form of technology that has revolutionized the medical field over the past decades. Digital pathology imaging, which captures histological details at the cellular level, is rapidly becoming a routine clinical procedure for cancer diagnosis support and treatment planning. Recent developments in deep-learning methods have facilitated tumor region segmentation from pathology images. The traditional shape descriptors that characterize tumor boundary roughness at the anatomical level are no longer suitable. New statistical approaches to model tumor shapes are in urgent need. In this article, we consider the problem of modeling a tumor boundary as a closed polygonal chain. A Bayesian landmark-based shape analysis model is proposed. The model partitions the polygonal chain into mutually exclusive segments, accounting for boundary roughness. Our Bayesian inference framework provides uncertainty estimations on both the number and locations of landmarks, while outputting metrics that can be used to quantify boundary roughness. The performance of our model is comparable with that of a recently developed landmark detection model for planar elastic curves. In a case study of 143 consecutive patients with stage I to IV lung cancer, we demonstrated the heterogeneity of tumor boundary roughness derived from our model effectively predicted patient prognosis (p-value <0.001). Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2021
Accepted December 2023

KEYWORDS
Artificial intelligence-reconstructed images; Landmark detection; Markov chain Monte Carlo; Shape analysis; Tumor boundary roughness

1. Introduction

Statistical shape analysis is an emerging field due to the necessity of making inferences on shapes, which is an important physical property of objects. It directly impacts medical imaging, computer vision, geographical profiling, and many other fields. Quantitatively describing the shape of an object, such as the tumor tissue from a medical image, has been a long-standing and fundamental challenge in medical imaging.

Since the emergence of radiology imaging technologies, a myriad of shape descriptors have been proposed for analyzing X-ray, computerized tomography (CT) scan, magnetic resonance imaging (MRI), and positron emission tomography (PET) images. These shape descriptors play a vital role in disease screening/staging/surveillance and treatment planning (Kijima et al. 2014; Mohammadzadeh et al. 2015), along with morphological texture descriptors (see. e.g., Haralick, Shanmugam, and Dinstein 1973; Larroza, Bodí, and Moratal 2016). With current advancements in imaging technology, hematoxylin and eosin (H&E)-stained pathology imaging (see an example in Figure 1(a)) is rapidly becoming a routine procedure in clinical diagnosis and prognosis of various malignancies (Niazi, Parwani, and Gurcan 2019). Compared to radiology images, pathology images can capture histological details in much higher resolution. Current pathology image analysis only builds upon morphological texture features (see. e.g., Tabesh et al. 2007; Yuan et al. 2012; Luo et al. 2016; Yu et al. 2016). There is a lack of shape descriptors to characterize tumors in high-resolution and complex medical images.

Recent technology breakthroughs in digital pathology imaging and machine learning have enabled comprehensive and detailed shape characterization of tumors on a large scale. Wang et al. (2018) have developed a deep convolutional neural network to classify image patches in a pathology image into three categories: normal, tumor, and empty (see Figure 1(b)), resulting in an artificial intelligence (AI)-reconstructed image (see Figure 1(c)). Consequently, the tumor boundary can be extracted and represented as a sequence of densely sampled pixel points (see the black solid line in Figure 1(d)). This boundary extraction can be achieved using tools like the R package `SAFARI` (Fernández et al. 2022).

Based on a case study analyzing AI-reconstructed lung cancer pathology images (detailed in Section 5.2), we found that most traditional shape descriptors based on radial distance, such as the zero-crossing count (ZCC) and tumor boundary roughness (TBR) (Kilday, Palmieri, and Fox 1993), developed for analysis of radiology images at the anatomical level performed poorly to characterize tumor boundaries extracted from pathology images at the cellular level. This implies that novel shape descriptors

CONTACT Qiwei Li qiwei.li@utdallas.edu Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

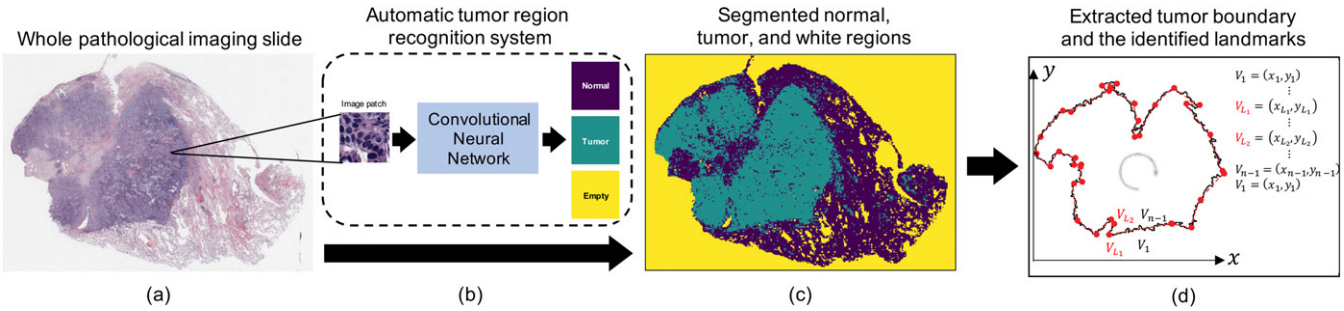


Figure 1. (a)–(c) an illustration of the pipeline developed by Wang et al. (2018): (a) The whole pathological imaging slide from a lung cancer patient (the median size of the slides analyzed in this article is 24,244 × 19,261 pixels); (b) The deep convolutional neural network that classifies each 300 × 300 pixels image patch into three categories: normal, tumor, and empty; (c) The AI-reconstructed image corresponding to the raw pathology image as shown in (a); (d) the tumor boundary extracted from the main tumor region as shown in (c) by Fernández et al. (2022) with the identified landmarks by BayesLASA (in red).

are needed for analyzing new high-resolution medical images, which usually exhibit substantial heterogeneity (Sadimin and Foran 2012). One motivation of this article is to develop a novel landmark identification model to enrich the family of tumor boundary descriptors. The identified landmarks, which approximately reconstruct the tumor shape, such as the red dots in Figure 1(d), partition the whole boundary into mutually exclusive pieces. The distribution of piecewise roughness measurements provides insight into the heterogeneity of tumor boundary roughness. For instance, a histogram with a sharp peak has a low kurtosis value, suggesting a constant roughness along the boundary. In contrast, a flat histogram featuring high kurtosis points to a greater degree of heterogeneity, indicating varied roughness across the tumor boundary.

The fundamental step in quantifying the heterogeneity of tumor boundary roughness is the selection of landmarks that effectively partition the entire boundary into pieces (i.e., segments) based on roughness. Landmark identification problem has been a primary focus in shape analysis. Several methods were developed based on global convexity (Subburaj, Ravi, and Agarwal 2008; Zulqarnain Gilani, Shafait, and Mian 2015) or local curvature (Liu et al. 2012). However, they have been challenged by low robustness and infeasible uncertainty assessment due to a lack of underlying statistical models. In contrast, Domijan and Wilson (2005) presented a model-based approach without considering shape-preserving transformations, while Strait, Chkrebtii, and Kurtek (2019) proposed a Bayesian model to detect the number and locations of landmarks using square-root velocity function (SRVF) representation under the elastic curve paradigm. Functional data are infinite-dimensional. The potential computational challenges may hinder its application in analyzing complex tumor shapes in high-resolution pathology images.

In this article, we consider using a closed polygonal chain to represent the boundary of an object. We develop a Bayesian LAndmark-based Shape Analysis (BayesLASA) model that can quantify the uncertainties of the number and locations of landmarks in a polygonal chain simultaneously. Our landmark estimation is naturally invariant to rotating, translating, reflecting, and scaling polygonal chains. BayesLASA is more efficient in practice compared to alternative approaches. Compared to existing landmark detection methods, BayesLASA also characterizes boundary smoothness through auxiliary parameters. We conduct a case study on a large cohort of lung cancer pathology images. The result shows that the heterogeneity of tumor boundary roughness, based on either the traditional surface profiling or hidden Markov modeling approach, is significantly associated with patient prognosis (p-value $<$ 0.001). This statistical methodology not only presents a new perspective to represent a digitized object's shape by using its landmarks, but also provides a new insight for understanding the role of tumor boundary roughness in cancer progression.

The remainder of this article is organized as follows. Section 2 introduces the proposed BayesLASA and discusses the parameter structure and prior specification. Section 3 briefly describes the Markov chain Monte Carlo (MCMC) algorithm and the resulting posterior inference for the landmark indicators. In Section 4, we evaluate BayesLASA on simulated data, comparing it with two alternative approaches. Section 5 analyzes one benchmark dataset in computer vision and a large cohort of pathology images in a lung cancer case study. Section 6 concludes the article with remarks on future extensions of the model.

2. The BayesLASA Model

2.1. Observed Data: A Polygonal Chain

Although the outline of a planar object is an absolutely continuous curve, it can also be represented as a sequence of discretization points (i.e., a closed polygonal chain). In geometry, a polygonal chain is a connected series of line segments (i.e., edges), each of which is a part of a line bounded by two distinct endpoints. Mathematically, a polygonal chain P is a discretized curve specified by a sequence of vertices $\{V_1, \ldots, V_n\}$ in a two-dimensional Cartesian plane. We use the ordered pair $(x_i, y_i) \in \mathbb{R}^2$ to denote the coordinates of each vertex $V_i, i = 1, \ldots, n$ (see an example in Figure 1(d)). This article only focuses on planar polygonal chains. However, the proposed method can be extended to a general case of $\mathbb{R}^k, k \geq 3$. A *simple polygonal chain* is one in which only consecutive segments intersect at their endpoints, while its opposite is a *self-intersecting polygonal chain*. For any simple polygonal chain, if the first vertex coincides with the last one $V_1 = V_n$ (i.e., their coordinates $(x_1, y_1) = (x_n, y_n)$), then it is a *closed polygonal chain* (CPC); otherwise, it is an *open polygonal chain* (OPC). In a simple polygon, two line segments meeting at a corner are usually required to form an angle that is not straight. However, we relax this constraint in the proposed

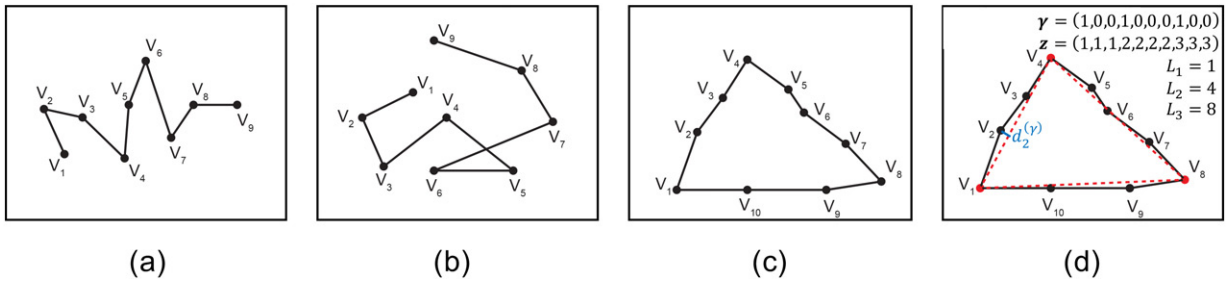


Figure 2. (a)–(c): examples of different types of polygonal chains: (a) an open polygonal chain (OPC); (b) a self-intersecting polygonal chain; (c) a closed polygonal chain (CPC); (d) a landmark chain (the red dashed line) of the CPC as shown in (c) with its three parameterizations: the landmark indicator vector $\boldsymbol{\gamma}$, the cluster allocation vector z, and the collection of landmark indices $\{L_1, \ldots, L_k\}$.

model. Figure 2 shows examples of open, self-intersecting, and closed polygonal chains, respectively.

We mainly consider modeling a CPC in this article, although this approach can also model an OPC with minor adjustments. The length of a CPC is defined as the sum of all line segments' lengths, while the center is defined as the arithmetic mean position of all vertices. Without loss of generality, we assume that the CPC has a unit length and is centered at the origin $(0, 0)$. This can be done by scaling and translating the polygon by altering each coordinate (x_i, y_i) as

$$\begin{cases} \left(x_i - \frac{1}{n-1}\sum_{i=1}^{n-1} x_i\right) \Big/ \sum_{i=1}^{n-1} \sqrt{(x_{i+1}-x_i)^2 + (y_{i+1}-y_i)^2} & \mapsto \quad x_i \\ \left(y_i - \frac{1}{n-1}\sum_{i=1}^{n-1} y_i\right) \Big/ \sum_{i=1}^{n-1} \sqrt{(x_{i+1}-x_i)^2 + (y_{i+1}-y_i)^2} & \mapsto \quad y_i \end{cases}. \tag{1}$$

2.2. Parameter Structure: A Gaussian Process

BayesLASA depends on two parameters. The first parameter, denoted as $\boldsymbol{\gamma}$, indicates the landmark locations. The second parameter, denoted as $\boldsymbol{d}^{(\gamma)}$, characterizes the discrepancy between the original CPC and the CPC formed by the landmarks.

2.2.1. Identifying the Landmarks

We define *landmarks* as those mathematically or structurally meaningful vertices in the boundary of a simple polygon, ignoring the remaining outline information. As the set of landmarks is a subset of $\{V_1, \ldots, V_{n-1}\}$, we use a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{n-1})^\top$ to indicate which vertices are landmarks, with $\gamma_i = 1, i = 1, \ldots, n-1$ if vertex i is a landmark and $\gamma_i = 0$ otherwise. The number of ones in $\boldsymbol{\gamma}$ is the number of landmarks, denoted by $K = \sum_{i=1}^{n-1} \gamma_i$. Those K landmarks form a polygonal chain; namely, the *landmark chain*. We use $P^{(\gamma)} = \{V_{L_1}, \ldots, V_{L_k}, \ldots, V_{L_K}, V_{L_1}\}$ to represent the landmark chain with L_k being the index of the kth landmark in P. Once $\boldsymbol{\gamma}$ is specified, we can calculate each L_k as

$$L_k = \sum_{i=1}^{n-1} \delta\left(\sum_{i'=1}^{i} \gamma_{i'} = k\right) \delta(\gamma_i = 1), \tag{2}$$

where $\delta(\cdot)$ is the indicator function. This formulation leads us to represent the landmark chain as $P^{(\gamma)}$, with the superscript (γ) denoting the indices of the landmarks characterized by $\gamma_i = 1$. Figure 2(d) shows an example of $P^{(\gamma)}$, along with different parameterizations. It is important to note that each configuration of $\boldsymbol{\gamma}$ corresponds to a specific landmark chain $P^{(\gamma)}$, provided P is defined with a known starting vertex and the direction. Consequently, this clarity in definition ensures the absence of identifiability issues.

To complete our model specification, we impose a product Bernoulli prior on $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma}|\boldsymbol{\omega} \sim \prod_{i=1}^{n-1} \text{Bern}(\omega_i)$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{n-1})^\top$ and $\omega_i \in (0, 1)$ for $i = 1, \ldots, n-1$. If no prior information is available, we could choose $\omega_1 = \cdots = \omega_{n-1} = \omega_0$ (e.g., a considerably small value), where ω_0 is interpreted as the probability of any vertex being a landmark a priori. For preferential landmarking of certain regions, we could consider a relatively large value of ω_i for those vertices within. Alternatively, we can model the uncertainty of ω_i by assuming $\omega_i \sim \text{Beta}(a_\omega, b_\omega)$. Consequently, we have

$$\begin{aligned} \pi(\boldsymbol{\gamma}) &= \prod_{i=1}^{n-1} \pi(\gamma_i) = \prod_{i=1}^{n-1} \int \pi(\gamma_i|\omega_i)\pi(\omega_i)d\omega_i \\ &= \frac{\Gamma(a_\omega + b_\omega)}{\Gamma(a_\omega)\Gamma(b_\omega)} \frac{\Gamma(a_\omega + K)\Gamma(b_\omega + n - 1 - K)}{\Gamma(a_\omega + b_\omega + n - 1)}. \end{aligned} \tag{3}$$

In practice, we suggest a constraint of $a_\omega + b_\omega = 2$ for a vague setting (Tadesse, Sha, and Vannucci 2005). Since the number of landmarks K equals to the sum of ones in $\boldsymbol{\gamma}$, we have $K \sim \text{BetaBin}(n-1, a_\omega, b_\omega)$ with an expected mean of $(n-1)a_\omega/(a_\omega + b_\omega) = (n-1)a_\omega/2$. Therefore, if there are K_E landmarks expected, then we suggest setting $a_\omega = 2K_\text{E}/(n-1)$ and $b_\omega = 2[1 - K_\text{E}/(n-1)]$. However, our sensitivity analysis in Section S4.2 of the supplementary materials shows that K_E has a minimal impact on posterior inference of $\boldsymbol{\gamma}$. Finally, we require that the landmark chain $P^{(\gamma)}$ be a CPC. This can be achieved by forcing $\pi(\boldsymbol{\gamma}) = 0$ if there are fewer than three landmarks or if $P^{(\gamma)}$ is a self-intersecting polygonal chain.

As vertices between two adjacent landmarks can be viewed as belonging to the same cluster or segment, the landmark identification is equivalent to a clustering or segmentation problem (i.e., partitioning $n-1$ vertices into K mutually exclusive

segments). To that end, we introduce an auxiliary set of cluster allocation variables $\boldsymbol{z} = (z_1, \ldots, z_{n-1})^\top, z_i \in \{1, \ldots, K\}$ to reparameterize $\boldsymbol{\gamma}$, where $z_i = k$ if vertex i is between the k and $(k+1)$th landmarks. Note that the index arithmetic of k will be taken modulo K if it is greater than K throughout the article, implying a cyclic ordering of landmarks with the first landmark following the last one. Mathematically, $\boldsymbol{z}$ is the cumulative sum of $\boldsymbol{\gamma}$, where $z_i = \sum_{i=1}^{i} \gamma_i$, with all zeros replaced by K, while $\boldsymbol{\gamma}$ is the lag-one difference of $\boldsymbol{z}$, where $\gamma_i = z_i - z_{i-1}$, with all negative entries replaced by one. Figure 2(d) shows the $\boldsymbol{z}$ induced by $\boldsymbol{\gamma}$ in the given example. It is worth noting that $\boldsymbol{\gamma}$ and $\boldsymbol{z}$ are interchangeable in that both reveal the same information about landmark locations. Our goal is to find the landmark chain $P^{(\gamma)}$ defined by $\boldsymbol{\gamma}$ or $\boldsymbol{z}$, given the observed CPC P.

2.2.2. Modeling the Discrepancy between the Polygonal and Landmark Chains

Here we discuss the probabilistic dependency between the observed P and its latent landmark chain $P^{(\gamma)}$. We write the full likelihood of $P = \{V_1, \ldots, V_{n-1}\}$ as a product over the K segments defined by its underlying landmarks,

$$f(V_1, \ldots, V_{n-1}|\boldsymbol{\gamma}, \cdot) = f(V_1, \ldots, V_{n-1}|\boldsymbol{z}, \cdot) = \prod_{k=1}^{K} f(P_k^*|\cdot), \text{ where}$$
$$P_k^* = \begin{cases} \{V_{L_k}, \ldots, V_{L_{k+1}-1}\} & \text{if } k < K \\ \{V_{L_K}, \ldots, V_{n-1}, V_1, \ldots, V_{L_1-1}\} & \text{if } k = K \text{ and } L_1 \neq 1 \\ \{V_{L_K}, \ldots, V_{n-1}\} & \text{if } k = K \text{ and } L_1 = 1 \end{cases}. \tag{4}$$

Note that although both the k and $(k+1)$th landmarks define segment k, we only place the former into segment k while the latter into the following segment by default. Next, we discuss how to specify $f(P_k^*|\cdot)$ in (6).

A non-landmark vertex whose $\gamma_i = 0$ should not be distant from the line segment defined by its two landmarks; otherwise, it might be considered a landmark itself. Therefore, we assume the shortest distance, denoted by $d_i^{(\gamma)}$, between vertex V_i and its associated line segment in $P^{(\gamma)}$ follows a distribution whose pdf is a monotonically decreasing function. Figure 2(d) shows the $d_i^{(\gamma)}$ (i.e., the blue solid line) of the given example. Section S2.1 and Figure S1 of the supplement describe how to derive $d_i^{(\gamma)}$ in detail. In particular, suppose vertex V_i at location (x_i, y_i) belongs to the kth segment (i.e., $z_i = k$), then the line passes through the k and $(k+1)$th landmarks at locations (x_{L_k}, y_{L_k}) and $(x_{L_{k+1}}, y_{L_{k+1}})$, respectively. We can compute the point-to-line distance as

$$d_i^{(\gamma)} = \pm \frac{|(x_{L_{k+1}} - x_{L_k})(y_{L_k} - y_i) - (x_{L_k} - x_i)(y_{L_{k+1}} - y_{L_k})|}{\sqrt{(x_{L_{k+1}} - x_{L_k})^2 + (y_{L_{k+1}} - y_{L_k})^2}}. \tag{5}$$

The numerator is twice the area of the triangle with vertices V_{L_k}, $V_{L_{k+1}}$, and V_i, while the denominator is the length of the line segment between V_{L_k} and $V_{L_{k+1}}$. If we view $d_i^{(\gamma)}$ as the height of the triangle, then (5) is just a rearrangement of the standard formula for the area of a triangle. We further define the sign of $d_i^{(\gamma)}$ as follows. The positive sign indicates that V_i is outside of the boundary of the landmark polygon $P^{(\gamma)}$, while the negative sign suggests the opposite. We could use the crossing number algorithm (Shimrat 1962) to find whether a point is inside or outside a simple polygon. In practice, we use the topology-based dimensionally extended nine-intersection model implemented by the function `st_within` in the R package `sf` (Pebesma 2018).

Given the landmark locations defined by $\boldsymbol{\gamma}$ or $\boldsymbol{z}$, we assume that the shortest distances $d_i^{(\gamma)}$'s that belong to the same segment are generated from a zero-mean stationary Gaussian process (GP). The spatial dependency among local vertices is modeled through the covariance structure in a multivariate normal (MVN) distribution,

$$f(P_k^*|\sigma_k^2, \cdot) = \text{MVN}\left(\boldsymbol{d}_k^*; \boldsymbol{0}, \sigma_k^2 \boldsymbol{G}\right), \text{ where}$$
$$\boldsymbol{d}_k^* = \begin{cases} \left(d_{L_k}^{(\gamma)}, \ldots, d_{L_{k+1}-1}^{(\gamma)}\right)^\top & \text{if } k < K \\ \left(d_{L_K}^{(\gamma)}, \ldots, d_{n-1}^{(\gamma)}, d_1^{(\gamma)}, \ldots, d_{L_1-1}^{(\gamma)}\right)^\top & \text{if } k = K \text{ and } L_1 \neq 1 \\ \left(d_{L_K}^{(\gamma)}, \ldots, d_{n-1}^{(\gamma)}\right)^\top & \text{if } k = K \text{ and } L_1 = 1 \end{cases}. \tag{6}$$

Here $\boldsymbol{d}_k^*$ is the distance vector of all non-landmark vertices assigned to segment k, $\boldsymbol{0}$ is a n_k-by-1 all zero column vector, σ_k^2 is the scaling factor, and the kernel $\boldsymbol{G}$ is a n_k-by-n_k positive definite matrix with each diagonal entry being one and each off-diagonal entry being a function of the relative position (i.e., Euclidean distance) between each pair of those non-landmark vertices in segment k. Here $n_k = \sum_{i=1}^{n-1} \delta(z_i = k) - 1$ is defined as the number of non-landmark vertices between the k and $(k+1)$th landmarks. For the sake of simplicity, we use the white-noise kernel $\boldsymbol{G} = \boldsymbol{I}$, where $\boldsymbol{I}$ is a n_k-by-n_k identity matrix that assumes each pair of d_i's is uncorrelated. Generalization of $\boldsymbol{G}$ to incorporate a certain spatial

dependence structure or desired smoothness is left as future work.

Employing a conjugate Bayesian approach, we impose an inverse-gamma (IG) hyperprior on each σ_k^2, which is expressed as $\sigma_k^2 \sim \text{IG}(a_\sigma, b_\sigma)$. This parameterization is standard in most Bayesian normal models. It allows for creating a computationally efficient algorithm by integrating out the variance component, which is usually a nuisance parameter. The integration leads to a marginal non-standardized t-distribution on each d_i (i.e., $d_i^{(\gamma)} \sim \text{t}_{2a_\sigma}(0, \sqrt{b_\sigma/a_\sigma})$ with $2a_\sigma$ degrees of freedom and a scale parameter of $\sqrt{b_\sigma/a_\sigma}$), and

$$f\left(P_k^*\right) = \int f\left(P_k^* | \sigma_k^2\right) \pi\left(\sigma_k^2\right) d\sigma_k^2 = (2\pi)^{-n_k/2} \frac{\Gamma\left(a_\sigma + n_k/2\right)}{\Gamma(a_\sigma)} \frac{b_\sigma^{a_\sigma}}{\left(b_\sigma + \boldsymbol{d}_k^{*\top} \boldsymbol{d}_k^*/2\right)^{a_\sigma + n_k/2}}. \tag{7}$$

In the case of σ_k^2 being considered a parameter of interest, we can easily sample this parameter from an IG distribution, $\sigma_k^2 | P_k^* \sim \text{IG}\left(a_\sigma + n_k/2, b_\sigma + \boldsymbol{d}_k^{*\top} \boldsymbol{d}_k^*/2\right)$ because of conjugacy.

To specify the IG hyperparameters, we first suggest setting $a_\sigma = 3$, which is the minimum integer value defining the IG variance. Then, we have $d_i^{(\gamma)} \sim \text{t}_6(0, \sqrt{b_\sigma/3})$, indicating that 95% of the point-to-line distances $|d_i^{(\gamma)}|$'s should be less than $\text{t}_{6,0.975}\sqrt{b_\sigma/3} \approx 1.4\sqrt{b_\sigma}$ a priori. If b_σ is set to be small, then this setting places high mass on small values of $|d_i^{(\gamma)}|$, encouraging "smoother" piecewise boundaries formed by more landmarks. Through our sensitivity analysis in Section S4.2 of the supplementary materials, we found that b_σ played a more important role than other hyperparameters on posterior inference of $\boldsymbol{\gamma}$. Therefore, we consider b_σ the main regularization parameter in our model. As shown in the deer example in Section 5.1 and the U.S. state shape examples in Section S5.1 of the supplement, we found that decreasing b_σ usually yielded more landmarks to capture fine-scale boundary structures. To avoid over-fitting, we suggest setting $b_\sigma = 1/(n-1)$ once the CPC has been scaled to have a unit length. With this setting, 95% of $|d_i^{(\gamma)}|$'s are less than 6%–14% of the unit length a priori, when the number of vertices $n-1$ ranges from 100 to 500 in the simulation study. This choice performed well on those scaled CPCs from simulated and various real datasets. However, it is noteworthy that we should choose b_σ with some degree of caution for unscaled CPCs or irregularly and sparsely-spaced CPCs.

2.3. Landmark Shape-Preserving Transformations

The proposed distance $d_i^{(\gamma)}$ is invariant to shape-preserving transformations such as rotation, translation, and reflection. Here we define the rotation of a polygon as multiplying each vertex (x_i, y_i) by the rotation matrix $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$, where $\theta \in [-\pi, \pi]$ is the rotation angle with respect to the positive x-axis. Translation is defined as the addition of a fixed vector, termed the translation vector, $(x_0, y_0) \in \mathbb{R}^2$ to each vertex (x_i, y_i). This operation shifts all vertices by x_0 and y_0 units along the x and y-direction, respectively. The reflection across the x and y-axis is defined as multiplying each vertex (x_i, y_i) by the matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$, respectively. The distance metric $d_i^{(\gamma)}$ (conditional on $\boldsymbol{\gamma}$) will remain the same if we alter each coordinate in (5) based on any of the above transformations or their combinations. Thus, our landmark estimation is invariant to those transformations.

As for the scaling, it is defined as multiplying each vertex (x_i, y_i) by the uniform scaling matrix $\begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}$, where $s \in (0, \infty)$ is the normalizing factor. Accordingly, the distance $d_i^{(\gamma)}$ can be enlarged ($s > 1$) or shrunk ($s < 1$). To ensure that the estimation on $\boldsymbol{\gamma}$ or $\boldsymbol{z}$ is invariant to changes in the scale, we should modify the prior setting of the model. In particular, if we assume $\sigma_k^2 \sim \text{IG}(a_\sigma, b_\sigma)$ before the scaling, then we need to set $\sigma_k^2 \sim \text{IG}(a_\sigma, s^2 b_\sigma)$ to preserve the distribution of $f\left(P_k^*\right)$ in (7).

Another important shape transformation is warping (Charpiat, Faugeras, and Keriven 2003; Kneip and Ramsay 2008), which is the continuous deformation of a given shape into another. A warping function is defined on functional representations of shape such as SRVF where the underlying shape spaces are infinite-dimensional. In this article, we represent the outline of a shape using a discretized curve. Unfortunately, warping is not achievable unless we create an approximating function that attempts to capture important patterns, while leaving out those fine-scale structures that are substantially present in tumor pathology images. Furthermore, warping is mainly used to analyze shapes that have similar underlying structures. However, tumor shapes from pathology images do not have similar structures and cannot be registered. Therefore, accounting for warping is undesirable under our framework.

3. Model Fitting

3.1. MCMC Algorithm

We now briefly describe the MCMC algorithm for posterior inference, while a detailed description is provided in Section S3.1 of the supplement. Our inferential strategy, which is based on Metropolis search variable selection algorithms (George and McCulloch 1997), allows us to simultaneously infer the number and locations of landmarks *via* $\boldsymbol{\gamma}$.

3.2. Posterior Estimation

We explore posterior inference of the landmark indicator $\boldsymbol{\gamma}$ or $\boldsymbol{z}$ by summarizing the set of MCMC samples after burn-in, denoted by $\{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(T)}\}$ and $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(T)}\}$, respectively, where T denotes the number of iterations after burn-in. We could choose the $\boldsymbol{\gamma}$ corresponding to the *maximum-a-posteriori* (MAP), that is, $\hat{\boldsymbol{\gamma}}^{\text{MAP}} = \underset{t}{\text{argmax}}\, \pi\left(\boldsymbol{\gamma}^{(t)}\right)$. Note that $\hat{\boldsymbol{z}}^{\text{MAP}}$ can be obtained by taking the cumulative sum of $\hat{\boldsymbol{\gamma}}^{\text{MAP}}$. An alternative estimate relies on the computation of the pairwise probability matrix (PPM), which is a $(n-1)$-by-$(n-1)$ symmetric matrix denoted by $\boldsymbol{C}$. Each entry indicates the posterior probability of each pair of vertices i and i' being in the same segment and

can be estimated by $c_{ii'} = \frac{1}{T}\sum_{t=1}^{T}\delta\left(z_i^{(t)} = z_{i'}^{(t)}\right)$. This estimate uses the information from all MCMC samples after burn-in. A point estimate of $\boldsymbol{z}$ can be approximated by minimizing the sum of squared deviations of its association matrix from the PPM (Dahl 2006), $\hat{\boldsymbol{z}}^{\text{PPM}} = \underset{\boldsymbol{z}}{\text{argmin}} \sum_{i<i'}(\delta(z_i = z_{i'}) - c_{ii'})^2$. Note that $\hat{\boldsymbol{\gamma}}^{\text{PPM}}$ can be obtained by differencing consecutive entries in $\hat{\boldsymbol{z}}^{\text{PPM}}$.

Once $\hat{\boldsymbol{\gamma}}^{\text{MAP}}$ or $\hat{\boldsymbol{\gamma}}^{\text{PPM}}$ is determined, we can immediately obtain the landmark indices, $\{L_1, \ldots, L_K\}$, *via* (2). We follow Jiang et al. (2021) to construct a credible interval for each landmark, using the local dependency structure from all MCMC samples on $\boldsymbol{\gamma}$ (see details in Section S3.2 of the supplement or a brief illustration in Figure S2).

4. Simulation

We used simulated data to demonstrate the performance of BayesLASA. We also conducted a sensitivity analysis on the choice of hyperparameters. The full details of the sensitivity analysis and scalability test can be found in Section S4.2 (including Figures S6–S8) and S4.3 of the supplement, respectively.

Simulated data were generated *via* the following steps. We first randomly generated an equilateral or non-equilateral simple polygon with $K = 4$, 5, or 6 vertices (considered as true landmarks) in a planar space, corresponding to a quadrilateral, pentagon, or hexagon, respectively. The perimeter of simple polygon uniformly ranged from $50K$ to $100K$. Next, we "binned" the landmark chain into a series of $n-1 = 100, 150, 200, 300$, or 500 equally sized intervals. Then, for each underlying interval, a non-landmark vertex was generated with its perpendicular distance to the interval sampled from $\text{N}(0, \sigma^2)$, where σ was chosen from $\{0.5, 1, 2\}$ with equal probability. We sequentially connected all vertices, including the K landmarks, to form a CPC. Last, we scaled and translated the generated CPC so that it had a unit length and was centered at $(0, 0)$, according to (1). Since we had three choices of K and five choices of n, there were $3 \times 5 = 15$ scenarios in total. For each scenario, we repeated the above steps to generate 100 replicated datasets.

For the beta prior on the landmark selection parameter ω, we set the two hyperparameters $a_\omega = 2K_E/(n-1)$ and $b_\omega = 2[1 - K_E/(n-1)]$ as discussed in Section 2.2.1, where $K_E = 3$ indicates that three landmarks were expected. For the IG prior on the variance component σ_k^2, we set $a_\sigma = 3$ and $b_\sigma = 1/(n-1)$ as discussed in Section 2.2.2. We ran $200n$ MCMC iterations with the first half treated as burn-in (i.e., $T = 100n$). Each chain was started from a model with three randomly chosen vertices as initial landmarks.

We evaluated model performance using two well-known metrics, both of which are given in Section S4.1 of the supplement. We first evaluated the landmark identification accuracy *via* the landmark indicator vector $\boldsymbol{\gamma}$. Since landmark and non-landmark vertices are usually of very different sizes (i.e., landmarks are assumed to be a small fraction of all vertices), most binary classification metrics are not suitable. We chose the Matthews correlation coefficient (MCC) (Matthews 1975). The MCC value ranges from -1 to 1. Larger values of MCC indicate better accuracy in landmark identification. Next, we chose the adjusted Rand index (ARI) (Hubert and Arabie 1985) to further evaluate model performance through the cluster allocation vector $\boldsymbol{z}$. The ARI is the corrected-for-chance version of the Rand index (Rand 1971), which is a similarity measure between two cluster allocation vectors. The ARI usually yields values between 0 and 1, although it can yield negative values (Santos and Embrechts 2009). Larger values of ARI indicate more similarities between $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$; thus, the identification of landmarks is more accurate as well.

To conduct a comparison study, we considered a recently developed algorithm named automatic landmark detection model for planar shape data (ALDUQ) (Strait, Chkrebtii, and Kurtek 2019). ALDUQ detects the number and locations of landmarks using SRVF representation under the elastic curve paradigm. It is, thus, necessary to convert the discrete polygonal chain into a continuously differentiable curve using the Gaussian kernel smoother with an appropriate length scale parameter. The ALDUQ output includes the relative locations of landmarks and their credible intervals represented as arcs. In some of our simulated datasets, we found that the reported credible interval covered more than half of the whole boundary. To make a feasible comparison, we considered a landmark as correctly identified if its estimated location was within a local window of the true position. In particular, let L_k denote the index of the kth true landmark. If the index of the landmark identified by BayesLASA was within $\{L_k - 5, \ldots, L_k, \ldots, L_k + 5\}$, or the location of a landmark detected by ALDUQ was on the curve bounded between V_{L_k-5} and V_{L_k+5}, then we regarded that the identified landmark hit the kth true landmark. We chose a window size of 11 vertices because the average number of non-landmark vertices between two true landmarks ranges from 17 to 125 in the simulated datasets. Note that the coverage needs to be customized based on the total number of vertices and their density for analyzing other datasets. As for the ALDUQ's setting, we used the default number of MCMC iterations (i.e., 100,000) and the default choice of $\lambda = 0.0001$ (i.e., a key regularization parameter in ALDUQ) since, according to the original paper, a smaller value of λ works better for simpler shapes. Some applications considered using the vertices of the convex hull of a shape as its landmarks. Thus, we also included this curvature-based approach implemented by the `R` function `chull`.

We first present the identified landmarks and their uncertainties by each method on one randomly selected dataset from three scenarios where K varied from 4 to 6 and $n - 1 = 150$. Figure 3 shows that BayesLASA successfully detected all true landmarks, while ALDUQ performed similarly but with some false positives and the convex hull approach had too many false positives. We also explored the posterior inference of the variance component σ_k^2 on the above three datasets. Although BayesLASA integrates each σ_k^2 out, we could reconstruct their posterior distributions on the original scale by sampling ${\sigma_k^{(t)}}^2 | \cdot \sim \text{IG}\left(a_\sigma + n_k^{(t)}/2, b_\sigma/s^2 + \boldsymbol{d}_k^{(t)*\top}\boldsymbol{d}_k^{(t)*}/2\right)$ at each MCMC iteration t, integrating the discussions in Sections 2.2.2 and 2.3. Here s is the normalizing factor and equal to the total length of the CPC on the original scale (i.e., $s = \sum_{i=1}^{n-1}\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$). Figure S3 of the supplement shows the 95% credible intervals $\hat{\sigma}_k^2$ against the

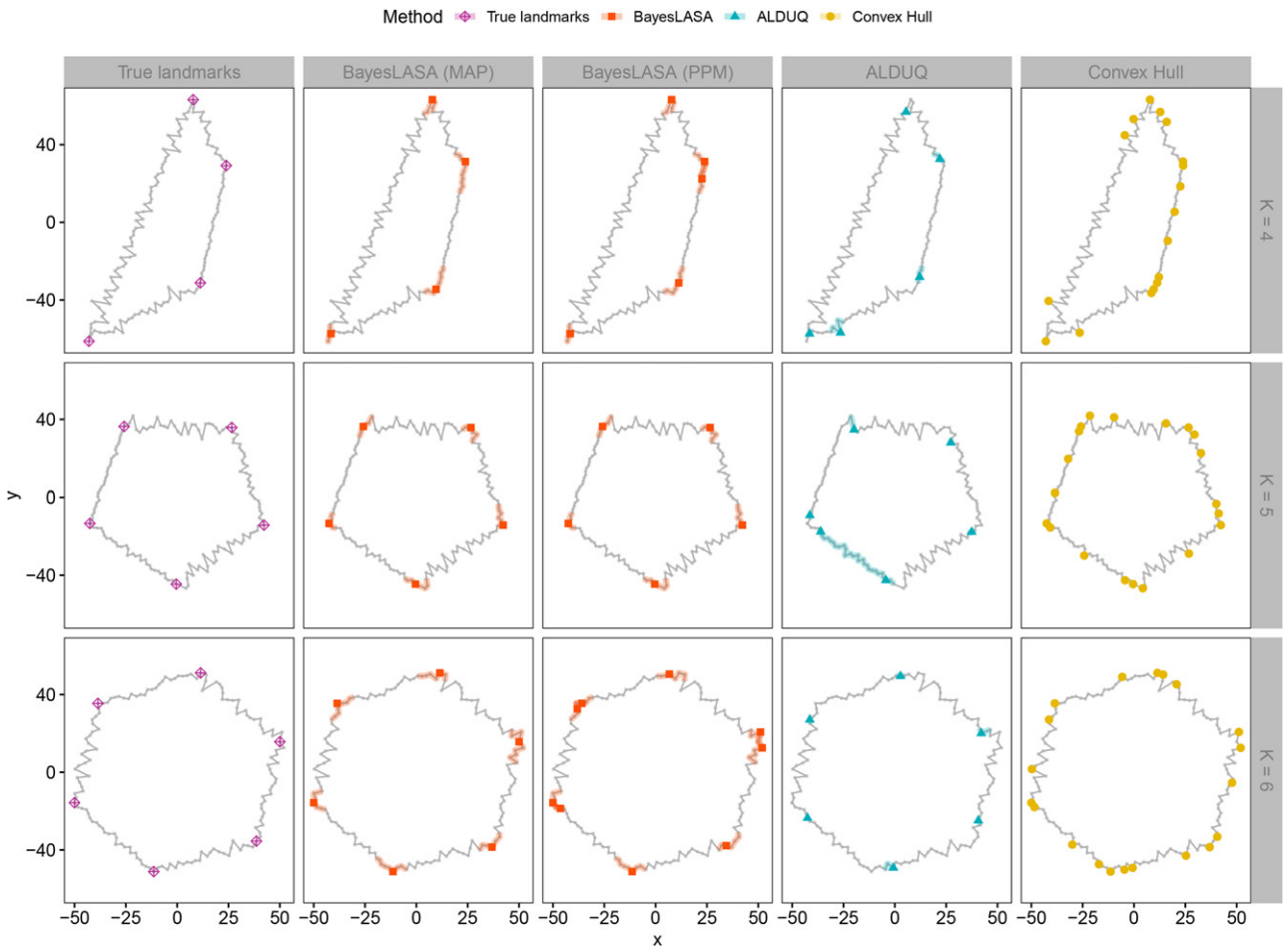


Figure 3. Simulation study: the identified landmarks by BayesLASA, ALDUQ, and convex hull on three randomly selected datasets under different scenarios in terms of the number of true landmarks K. For BayesLASA and ALDUQ, 95% credible intervals are provided, depicted as shaded regions, to indicate the level of uncertainty in landmark identification.

vertex indices. All true values were within their corresponding intervals. Furthermore, we performed a more comprehensive evaluation on all simulated datasets. Figures S4–S5 of the supplement exhibit the MCC and ARI distribution, respectively, under each scenario in terms of K and n. Regardless of the choice between $\hat{\boldsymbol{\gamma}}^{\text{MAP}}$ and $\hat{\boldsymbol{\gamma}}^{\text{PPM}}$, BayesLASA had a notable advantage over the competing methods in all scenarios. We also note that the performance of all methods decreased as either K or n increased. It is noteworthy that ALDUQ was mainly developed for landmarking multiple curves. It is expected to notice improved performance if multiple samples from the same shape are provided. Additionally, Figure S9 of the supplement shows that the computational cost of BayesLASA was significant lower than ALDUQ and it increased approximately linearly in the number of vertices of the CPC.

5. Applications

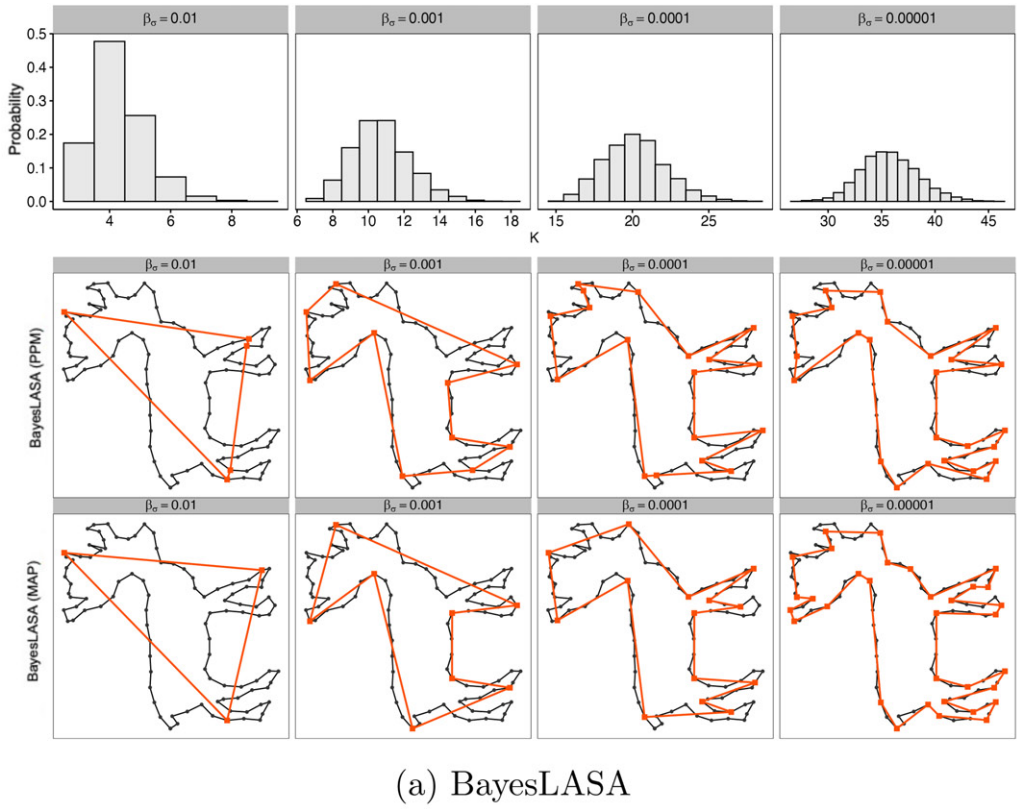
In this section, we first evaluated the performance of our methodology using a benchmark dataset in computer vision. Then, we applied the model to the AI-reconstructed pathology images from a large cohort of lung cancer patients. The results revealed novel potential tumor shape-based imaging biomarkers for lung cancer prognosis. Note that in the above two case studies, the vertices of each CPC were regularly and densely sampled from the related image. To show that BayesLASA is applicable to irregularly and sparsely sampled discretized curves, we demonstrated an additional case study on the contiguous 48 U.S. state shapes in Section S5.1 and Figures S11–S16 of the supplementary material.

5.1. Case Study on a Complex Shape in Computer Vision

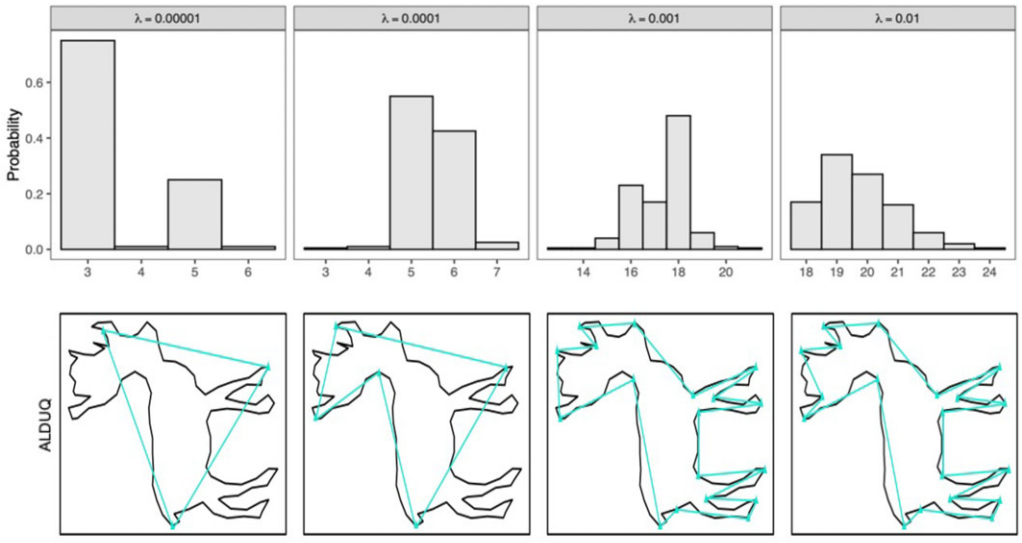
The well-known MPEG-7 dataset[1] is commonly used for benchmarking shape matching algorithms. The dataset includes 1400 binary images of 70 objects, all of which involve closed curves. To demonstrate how BayesLASA can lead to sharper inferences, we focus on complex shapes such as deer, which were also analyzed in Strait, Chkrebtii, and Kurtek (2019), because it is unclear where or how the landmarks should be selected. It is, thus, more prudent to apply BayesLASA on those complex objects.

The CPC representing the deer outline contains evenly spaced $n - 1 = 100$ vertices. Figure S10 of the supplement illustrates the beta-binomial prior distributions of the number of landmarks K under various prior configuration. We applied

[1] https://dabi.temple.edu/external/shape/MPEG7/dataset.html



(a) BayesLASA



(b) ALDUQ

Figure 4. Computer vision case study: the marginal posterior distribution of the number of landmarks $\pi(K|\cdot)$ and the estimated landmark chain $P^{(\gamma)}$ based on $\hat{\boldsymbol{\gamma}}^{\text{MAP}}$ and $\hat{\boldsymbol{\gamma}}^{\text{PPM}}$ under different prior choices by (a) BayesLASA and (b) ALDUQ.

BayesLASA with a constant $K_E = 1$ (refer to the red line in Figure S10) and different values of $b_\sigma \in \{\frac{1}{n}, \frac{1}{10n}, \frac{1}{100n}, \frac{1}{1000n}\}$, while the remaining prior settings were kept the same as described in Section 4. Notably, sensitivity analysis on simulated data revealed that BayesLASA's performance is considerably insensitive to the choice of K_E. For each setting, we ran four independent MCMC chains. We checked MCMC algorithm convergence based on the $(n-1)$-dimensional marginal posterior probabilities of inclusion (PPI) vectors, where the ith entry is $\sum_{t=1}^{T} \delta\left(\gamma_i^{(t)} = 1\right)/T$ that indicates the marginal posterior probability of vertex i being a landmark. We calculated the PPIs for all four chains and found that their pairwise Pearson correlation coefficients ranged from 0.954 to 0.964 across all settings, which suggested good MCMC convergence. We then aggregated the outputs of all chains.

Figure 4(a) shows the marginal posterior distribution of the number of landmarks and the landmark chains estimated by both MAP and PPM under the four choices of b_σ. The plot clearly illustrates how the prior information influenced landmark selection. More landmarks were identified when b_σ is small, which results in a finer reconstruction of the original shape. The "elbow" points did not become clearly evident until $K \approx 20$. In comparison, we also applied ALDUQ to the deer shape after converting the discrete polygonal chain into a continuously differentiable curve using the Gaussian kernel smoother. Similar to b_σ in BayesLASA, ALDUQ also has a parameter λ that acts as a tuning parameter for regularization. The results are summarized in Figure 4(b). We found that larger values of λ increased the posterior mean of K resulting in better shape reconstructions for substantially small values of K. Both methods allow users to control the number of landmarks selected on the shape of interest through the prior setting. Both methods performed similarly on the deer example when their resulting K values were close.

5.2. Case Study on Lung Cancer Pathology Images

Lung cancer has been ranked as the leading cause of death from cancer, with non-small-cell lung cancer (NSCLC) accounting for about 85% of lung cancer deaths. Current guidelines for diagnosing and treating cancer are largely based on pathological examination of tissue section slides. A deep-learning approach has been developed to perform the tumor segmentation of pathology images (Wang et al. 2018). Specifically, a convolutional neural network (CNN)-based classifier was trained using a large cohort of lung cancer pathology images. This approach classifies each 300 × 300 pixels image patch at 40× magnification into one of the three categories: normal, tumor, or background. It is required that at least 20 cells were within each image patch. Tumor and non-tumor image patches were randomly extracted from tumor regions and nonmalignant regions labeled by a well-experienced pathologist. The patches were classified as background if the mean intensity of all pixel values was larger than a threshold. After the three-class AI-reconstructed image was generated, we used a recently developed R package `SAFARI` (Fernández et al. 2022) to extract the connected tumor regions and their boundaries.

In this case study, we used 246 H&E-stained pathology imaging slides from 143 consecutive patients with stage I to IV NSCLC in the National Lung Screening Trial (NLST). All patients had undergone surgical procedures for treatment. The median size of the pathology slides is 24,244 × 19,261 pixels. Tumor segmentation was done by the CNN classifier. Each AI-reconstructed image was further enlarged three times to avoid single-pixel boundary lines or singularities. The median size of the resulting three-class images used for tumor boundary tracing (i.e., CPC generation by `SAFARI`) is 1011 × 806 pixels. Only the tumor region with the largest area in each image was considered. The number of CPC vertices (in pixels) ranges from $n - 1 = 360$ to 15,931, with a median of 3836. Note that we did not scale CPCs to have a unit length in this case study because we need to characterize tumor boundary roughness at the same spatial resolution across all images so that comparable image-wise analysis and patient-wise survival analysis are achievable.

We applied BayesLASA to each of the 246 CPCs independently with the same setting described in Section 4 except for choosing a standardized $b_\sigma = 500$. This choice makes 95% of $|d_i^{(\gamma)}|$'s (i.e., the shortest distances between non-landmark vertices and their associated line segments in the landmark chain $P^{(\gamma)}$) are less than $1.4\sqrt{b_\sigma} \approx 30$ pixels a priori (see the derivation in Section 2.2.2), corresponding to approximately $30/360 = 8\%$ of the smallest CPC's length. Such a setting ensured fine and proper boundary roughness characterization even for the smallest CPC in this study and generated satisfactory landmarking results. A total of four MCMC chains were run and averaged. $\hat{\boldsymbol{z}}^{\mathrm{PPM}}$ or $\hat{\boldsymbol{\gamma}}^{\mathrm{PPM}}$ was used to determine the landmarks. We performed the same convergence diagnosis as described in Section 5.1. Figure 5(a) and (b) shows two examples of tumor boundaries and their identified landmarks by BayesLASA from a patient with good prognosis (i.e., alive over 2537 days after the surgery) and another patient with poor prognosis (i.e., died on the 29th day after the surgery). Notably, the tumor from the patient with shorter survival time exhibited a more spiculated shape compared to the patient with good prognosis, indicating the invasion of tumor cells into surrounding tissues. Although the two tumor regions had distinctive boundaries, the roughness and its heterogeneity were much more subtle in many other examples. Thus, BayesLASA can play a role in supplementing human visual inspection.

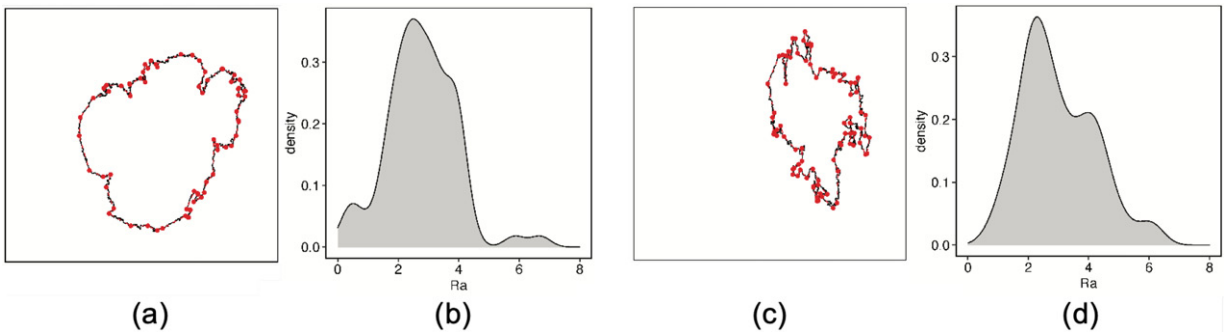


Figure 5. Lung cancer case study: two examples of the extracted tumor boundaries and their landmarks (in red) identified by BayesLASA in the NLST dataset, where (a) was from a patient who was still alive over 2537 days after the surgery and (c) was from a patient who died on the 29th day after the surgery; (b) and (d): The density plots of distance-based roughness measurement Ra corresponding to (a) and (c).

Next, we demonstrate how to characterize the heterogeneity of boundary roughness based on the landmarks. The landmark chain $P^{(\gamma)}$ can be considered as a skeleton reference of the object outline P. The distance vector $\boldsymbol{d}^{(\gamma)} = (d_1^{(\gamma)}, \ldots, d_{n-1}^{(\gamma)})^\top$ reveals boundary roughness. We now introduce two types of boundary roughness measurements: distance and model-based. The piecewise boundary roughness is represented by a $K \times 1$ vector with the kth entry being the distance or model-based roughness measurement calculated or estimated by all $d_i^{(\gamma)}$ in the same segment k.

Distance-based roughness measurements: Surface roughness measurements computed by simple math equations were adopted to quantify the irregularity of tumor boundary. For instance, the arithmetical mean deviation (denoted by Ra) of each segment is defined as $\sum_{i:z_i=k} |d_i^{(\gamma)}|/n_k$ while the other seven measurements, including Rq, Rv, Rp, Rz, Rsk, Rku, and RzJIS, are summarized in Table S1 of the supplement.

Model-based roughness measurements: Since the distances $d_i^{(\gamma)}$'s are sequentially indexed, their changing frequencies indicates the fluctuation degree of surface roughness. A hidden Markov model (HMM) was fitted. In particular, we assumed $d_i^{(\gamma)}$'s for those vertices from the same segment were from a two-component Gaussian mixture model. Two hidden states corresponding to the two components, were defined to illustrate the negative '−' (inside of the landmark chain) and positive '+' (outside of the landmark chain) sign of each $d_i^{(\gamma)}$. The transition probabilities control the way that the hidden state of $d_{i+1}^{(\gamma)}$ is chosen given the hidden state of $d_i^{(\gamma)}$ within each segment in $\boldsymbol{z}$, reflecting the segment-specified roughness. The transition probabilities, denoted by q_{++}, q_{+-}, q_{-+}, and q_{--}, respectively, were estimated for each segment by using the related functions in the `R` package `depmixS4`.

5.2.1. Association Study

With the identified landmarks for all tumor regions, we conducted a downstream analysis to scrutinize tumor shape-related prognostic factors that predict survival, where the survival status of the patients was monitored from the time of the surgery to the end of the clinical trial. Specifically, a Cox proportional hazard (CoxPH) model (Cox 1972) was fitted with the summary statistics such as the mean (X_1), standard deviation (X_2), skewness (X_3), and kurtosis (X_4), which measure the center, spread, asymmetry, and tailedness, respectively, of the distribution of piecewise roughness measurements after adjusting for the number of identified landmarks (X_5), tumor size (in pixels, X_6), cancer stage (X_7, X_8, X_9 for stages II, III, IV vs. stage I, respectively), gender (X_{10}) and tobacco history (X_{11}),

$$h(t) = h_0(t) \exp\left(\sum_{j=1}^{11} \beta_j X_j\right), \tag{8}$$

where $h(t)$ is the expected hazard at time t and $h_0(t)$ is the baseline hazard when all predictors are equal to zero or reference level. Multiple sample images from the same patient were modeled as correlated observations in the CoxPH model to compute a robust variance for each coefficient using the R package `survival`.

Table 1. Lung cancer case study: the output of the CoxPH model with the distance-based roughness measurement Ra.

Notation	Predictor	Coef	exp(Coef)	SE	p-value
X_1	Mean of Ra's	−0.757	0.469	0.604	0.210
X_2	Standard deviation of Ra's	−0.428	0.652	0.627	0.494
X_3	Kurtosis of Ra's	−0.368	0.692	0.106	$\mathbf{5.4 \times 10^{-4}}$
X_4	Skewness of Ra's	1.675	5.336	0.469	$\mathbf{1.8 \times 10^{-4}}$
X_5	Number of landmarks K	0.014	1.014	0.007	**0.047**
X_6	Tumor size	0.000	1.000	0.000	0.828
X_7	Cancer stage II vs. I	0.343	1.410	0.619	0.579
X_8	Cancer stage III vs. I	1.168	3.214	0.371	**0.002**
X_9	Cancer stage IV vs. I	1.802	6.064	0.463	$\mathbf{9.9 \times 10^{-5}}$
X_{10}	Smoking vs. nonsmoking	−0.119	0.888	0.330	0.718
X_{11}	Female vs. male	−0.127	0.881	0.322	0.694

p-values below the significance threshold of 0.05 are highlighted in bold.

The output of the CoxPH model using one of the distance-based roughness measurements, Ra, is shown in Table 1. Advanced stage lung cancer is significantly correlated with poorer prognosis, which is in agreement with previous knowledge. It is noteworthy that kurtosis and skewness had significant negative and positive effects, respectively. The same result were observed for the other choices of distance-based roughness measurements including Rq, Rp, Rv, Rx, and RzJIS (shown in the Tables S2–S6 of the supplement). The CoxPH model fitted with Ra obtained an overall p-value of 0.0001 by the Wald test. The CoxPH model fitted with the moments of Ra found that kurtosis and skewness were significant prognostic factors (both p-values < 0.001), which measure the peakedness and asymmetry of the probability distribution, respectively. The negative coefficient of kurtosis and positive coefficient of skewness suggested that tumor with smaller kurtosis (flat spreading) and larger skewness (left-centered) are more heterogeneous in surface roughness and, thus, indicate a worse prognosis (as illustrated in Figure 5(e) and (f)). These results are consistent with the biology literature in that high spatial heterogeneity is a pivotal feature of cancer at both the cellular and histological levels resulting from the distinct patterns of different cancer cell subpopulations in terms of dysregulation of proliferation, mobility, and metabolism pathways (Meacham and Morrison 2013; Dagogo-Jack and Shaw 2018). The underlying biological mechanism of a heterogeneous tumor boundary could be attributed to heterogeneous regulation of gene expression by abnormally activated Rho GTPase pathways among cancer cell subpopulations and consequent dissimilarity in the downstream actin cytoskeleton and stress fibers (Pascual-Vargas et al. 2017).

To validate that our landmark-based shape analysis is robust to different roughness measurements, we repeated the above steps to fit another CoxPH model with HMM-based roughness measurements as predictors. The overall p-value of CoxPH model fitted with moments of negative to positive transition probability q_{-+} is 0.0008 (Wald test) and the coefficients and p-values for each variable are shown in Table 2. The CoxPH model for the transition probability q_{+-} was summarized in Table S7 of the supplement, showing a similar result. Again, kurtosis and skewness are significant factors associated with patient prognosis. Furthermore, standard deviation of q_{-+}'s had a p-value $= 0.0009$ and a large positive coefficient, which indicates

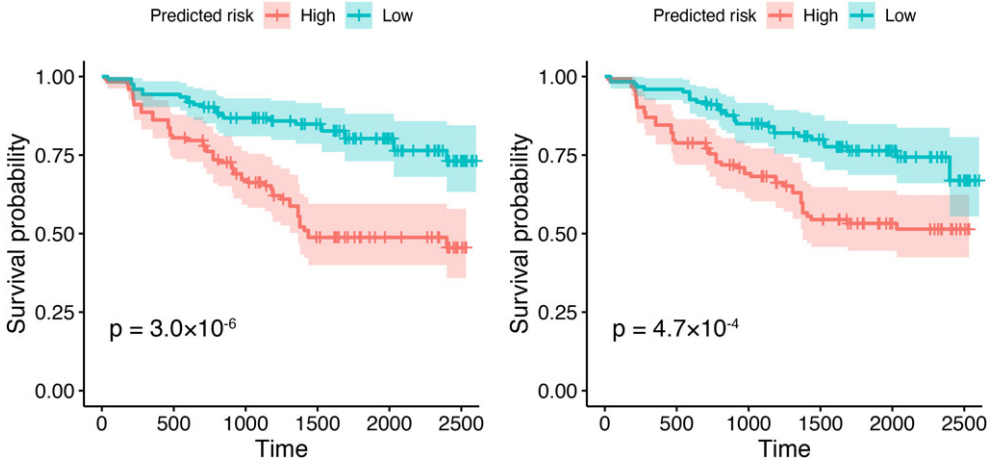


Figure 6. Lung cancer case study: the Kaplan-Meier plots for the low and high-risk groups predicted by the LOOCV *via* the CoxPH model with the (a) distance-based and (b) HMM-based roughness measurements.

Table 2. Lung cancer case study: the output of the CoxPH model with the HMM-based roughness measurement q_{-+}.

Notation	Predictor	Coef	exp(Coef)	SE	*p*-value
X_1	Mean of q_{-+}'s	5.797	329.4	9.541	0.543
X_2	Standard deviation of q_{-+}'s	18.10	7.3×10^7	6.915	**0.009**
X_3	Kurtosis of q_{-+}'s	0.091	1.096	0.046	**0.046**
X_4	Skewness of q_{-+}'s	−0.958	0.384	0.369	**0.009**
X_5	Number of landmarks K	0.014	1.014	0.008	0.088
X_6	Tumor size	0.000	1.000	0.000	0.783
X_7	Cancer stage II vs. I	0.502	1.651	0.593	0.397
X_8	Cancer stage III vs. I	1.195	3.303	0.393	**0.002**
X_9	Cancer stage IV vs. I	1.791	5.997	0.490	$\mathbf{2.6 \times 10^{-4}}$
X_{10}	Smoking vs. nonsmoking	−0.042	0.959	0.322	0.896
X_{11}	Female vs. male	−0.122	0.885	0.307	0.692

p-values below the significance threshold of 0.05 are highlighted in bold.

that patient prognosis worsens as tumor boundary roughness heterogeneity increases.

In contrast, we fitted a similar CoxPH model using the radial distance-based shape features as predictors, including the aforementioned ZCC and TBR (Kilday, Palmieri, and Fox 1993). We first introduce the radial length r_i between vertex i and the polygon center at location $(\bar{x}, \bar{y}) = (\sum_{i=1}^{n-1} x_i/(n-1), \sum_{i=1}^{n-1} y_i/(n-1)))$, defined as $r_i = \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}$. Let $\boldsymbol{r} = (r_1, \ldots, r_{n-1})^\top$ denote the radial length of all vertices. The ZCC is the number of times that radial length crosses the mean value $\bar{r} = \sum_{i=1}^{n-1} r_i/(n-1)$, that is, $\mathrm{ZCC}(\boldsymbol{r}) = \sum_{i=1}^{n-1} \delta\left((r_i - \bar{r})(r_{i+1} - \bar{r}) < 0\right)$. The TBR is calculated by averaging the roughness index (RI) for a window with length L over the entire tumor boundary where the RI for window j is defined as $\mathrm{RI}_j(\boldsymbol{r}) = \sum_{i=(j-1)L+1}^{jL-1} |r_{i+1} - r_i|$ for $j = 1, \ldots, \lceil (n-1)/L \rceil$, and so $\mathrm{TBR}(\boldsymbol{r}) = \sum_{j=1}^{\lceil (n-1)/L \rceil} \mathrm{RI}_j(\boldsymbol{r}) / \lceil (n-1)/L \rceil$. Here $\lceil \cdot \rceil$ denotes the ceiling function. The results imply an insignificant association between the ZCC and clinical outcomes (p-value = 0.128). For the TBR, we attempted to vary the window size L from 5 to 200. Figure S17 of the supplement shows the TBR p-values against L. The obtained p-values ranged from 0.142 to 0.925. Unfortunately, we could not identify any association between these two radial distance-based roughness measurements and the patient survival outcome from the NLST dataset. The comparison demonstrates that the proposed model-based shape analysis can lead to enhanced statistical power on tumor boundary roughness with a more robust detection of associations than ordinary exploratory analyses.

5.2.2. Predictive Performance by Cross-Validation

Finally, we employed the leave-one-out cross-validation (LOOCV) to evaluate the predictive performance of the above two CoxPH models. In particular, we first trained a CoxPH model using all images from all patients excluding the left-out one. Next, we obtained a survival risk score for each test image from the left-out patient. The survival risk score for this patient was then calculated as the average survival risk score over all test images. After repeating these steps for each of the 143 NSCLC patients, we divided the patients into two equally sized groups (i.e., low and high-risk), choosing the median of patient-specific risk scores as the cutoff. Their corresponding Kaplan-Meier survival curves are displayed in Figure 6(a) and (b), where the predictors were the summary statistics of Ra's and q_{-+}'s, respectively. Both log-rank tests showed a significant difference between the two groups (i.e., p-value $= 3.0 \times 10^{-6}$ and 4.7×10^{-4}, respectively).

6. Conclusion

A large amount of complex and comprehensive information about tumor aggressiveness and malignancy is harbored in the tumor shapes captured by pathology imaging. Recent advances in deep-learning methods have provided plausible approaches for automatic tumor segmentation in medical images on a large scale. Shape features are proven with success in radiomics, however, they are no longer satisfactory in pathology imaging. To discover more clinically meaningful biomarkers in high-resolution medical images, we propose BayesLASA to better characterize tumor shapes and boundaries. The primary contribution of this work is the development of a more accurate and efficient landmark detection method under the discrete polygonal chain paradigm, in contrast to methods based on elastic curves. BayesLASA can also be extended to applications in various scenarios where a sequence of discretization points could represent the outline of a shape. Furthermore, we propose

several types of new landmark-based features to characterize boundary roughness. Our study demonstrates the prognostic value of those features in two downstream analyses of lung cancer pathology images. The results show that boundary roughness heterogeneity was significantly associated with patient prognosis. The boundary roughness can be easily measured through BayesLASA and be used as a potential biomarker for patient prognosis. This novel imaging biomarker can be conceived as a real clinical tool at low cost because it is based only on tumor pathology slides, which are available in standard clinical care.

For our future research, several extensions of our model are worth investigating. First, we could generalize the kernel used to measure the discrepancy between a polygonal chain and its landmark chain. For instance, using a squared exponential, Matérn, or rational quadratic kernel will help us incorporate spatial dependence or desired smoothness. Landmark identification and smoothness quantification could be jointly inferred. Moreover, we would like to extend our framework to high-resolution pathology images of other cancer types, which would be a promising direction in clinical science research. Lastly, our work only focuses on landmark detection for a single polygonal chain. Landmarking of boundaries from multiple polygonal chains is greatly needed to increase the accuracy and decrease the uncertainty of landmark estimation. It could help in estimating the scaling factors σ_k's, especially for irregularly and sparsely sampled discretized curves. Strait, Chkrebtii, and Kurtek (2019) have studied landmarking of multiple shapes when their boundaries are treated as elastic curves. However, a modeling approach under the polygonal chain paradigm is still a work in progress. Our future direction includes each of the aforementioned extensions.

Supplementary Materials

The supplementary materials for Sections 2–5 are available online, including the derivation of the point-to-line distance, a detailed description of the MCMC algorithms, explicit definitions of evaluation metrics, reports on sensitivity analysis and scalability test, supplementary tables and figures from the lung case study, and an additional U.S. state shape case study. We provide software in the form of `R`/`C++` code on GitHub (*https://github.com/bougetsu/BayesLASA*). To reproduce the figures presented in the paper, refer to the corresponding `R` scripts in the 'manuscript_reproducibility' directory at the same repository.

Acknowledgments

The authors would like to thank the editor, associate editor, and the three committed reviewers for their careful and constructive review. The authors' thanks also go to Kevin C. Lutz and Kevin W. Jin for their help proofreading the paper.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by NSF DMS 2210912, NSF DMS 2113674, and NIH 1R01GM141519.

ORCID

Tejasv Bedi https://orcid.org/0000-0001-7532-4075
Chul Moon https://orcid.org/0000-0002-2892-5449
Yang Xie https://orcid.org/0000-0001-9456-1762
Min Chen https://orcid.org/0000-0002-0676-3143
Qiwei Li https://orcid.org/0000-0002-1020-3050

References

Charpiat, G., Faugeras, O., and Keriven, R. (2003), "Shape Metrics, Warping and Statistics," in *Proceedings 2003 International Conference on Image Processing* (Vol. 2), IEEE, pp. II–627. [5]

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, Series B, 34, 187–202. [10]

Dagogo-Jack, I., and Shaw, A. T. (2018), "Tumour Heterogeneity and Resistance to Cancer Therapies," *Nature Reviews Clinical Oncology*, 15, 81–94. [10]

Dahl, D. B. (2006), "Model-based Clustering for Expression Data via a Dirichlet Process Mixture Model," *Bayesian Inference for Gene Expression and Proteomics*, 4, 201–218. [6]

Domijan, K., and Wilson, S. (2005), "A Bayesian Method for Automatic Landmark Detection in Segmented Images," in *Learning Techniques for Processing Multimedia Content*, p. 69. [2]

Fernández, E., Yang, S., Chiou, S. H., Moon, C., Zhang, C., Yao, B., Xiao, G., and Li, Q. (2022), "SAFARI: Shape Analysis for AI-Segmented Images," *BMC Medical Imaging*, 22, 1–7. [1,2,9]

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [5]

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973), "Textural Features for Image Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, 6, 610–621. [1]

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [6]

Jiang, S., Zhou, Q., Zhan, X., and Li, Q. (2021), "BayesSMILES: Bayesian Segmentation Modeling for Longitudinal Epidemiological Studies," *Journal of Data Science*, 19, 365–389. [6]

Kijima, S., Sasaki, T., Nagata, K., Utano, K., Lefor, A. T., and Sugimoto, H. (2014), "Preoperative Evaluation of Colorectal Cancer Using CT Colonography, MRI, and PET/CT," *World Journal of Gastroenterology*, 20, 16964. [1]

Kilday, J., Palmieri, F., and Fox, M. D. (1993), "Classifying Mammographic Lesions Using Computerized Image Analysis," *IEEE Transactions on Medical Imaging*, 12, 664–669. [1,11]

Kneip, A., and Ramsay, J. O. (2008), "Combining Registration and Fitting for Functional Models," *Journal of the American Statistical Association*, 103, 1155–1165. [5]

Larroza, A., Bodí, V., and Moratal, D. (2016), "Texture Analysis in Magnetic Resonance Imaging: Review and Considerations for Future Applications," in *Assessment of Cellular and Organ Function and Dysfunction using Direct and Derived MRI Methodologies*, ed. C. Constantinides, pp. 75–106, London: IntechOpen. [1]

Liu, Y., Sajja, B. R., Uberti, M. G., Gendelman, H. E., Kielian, T., and Boska, M. D. (2012), "Landmark Optimization Using Local Curvature for Point-based Nonlinear Rodent Brain Image Registration," *International Journal of Biomedical Imaging*, 2012, 635207. [2]

Luo, X., Zang, X., Yang, L., Huang, J., Liang, F., Canales, J. R., Wistuba, I. I., Gazdar, A., Xie, Y., and Xiao, G. (2016), "Comprehensive Computational Pathological Image Analysis Predicts Lung Cancer Prognosis," *Journal of Thoracic Oncology*. 12, 501–509. [1]

Matthews, B. W. (1975), "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405, 442–451. [6]

Meacham, C. E., and Morrison, S. J. (2013), "Tumour Heterogeneity and Cancer Cell Plasticity," *Nature*, 501, 328–337. [10]

Mohammadzadeh, Z., Safdari, R., Ghazisaeidi, M., Davoodi, S., and Azadmanjir, Z. (2015), "Advances in Optimal Detection of Cancer by Image Processing: Experience with Lung and Breast Cancers," *Asian Pacific Journal of Cancer Prevention*, 16, 5613–5618. [1]

Niazi, M. K. K., Parwani, A. V., and Gurcan, M. N. (2019), "Digital Pathology and Artificial Intelligence," *The Lancet Oncology* 20, e253–e261. [1]

Pascual-Vargas, P., Cooper, S., Sero, J., Bousgouni, V., Arias-Garcia, M., and Bakal, C. (2017), "RNAi Screens for Rho GTPase Regulators of Cell Shape and YAP/TAZ Localisation in Triple Negative Breast Cancer," *Scientific Data*, 4, 1–13. [10]

Pebesma, E. (2018), "Simple Features for R: Standardized Support for Spatial Vector Data," *The R Journal* 10, 439–446. DOI:10.32614/RJ-2018-009 [4]

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850. [6]

Sadimin, E. T., and Foran, D. J. (2012), "Pathology Imaging Informatics for Clinical Practice and Investigative and Translational Research," *North American Journal of Medicine and Science*, 5, 103–109. [2]

Santos, J. M., and Embrechts, M. (2009), "On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification," in *International Conference on Artificial Neural Networks*, pp. 175–184, Springer. [6]

Shimrat, M. (1962), "Algorithm 112: Position of Point Relative to Polygon," *Communications of the ACM*, 5, 434. [4]

Strait, J., Chkrebtii, O., and Kurtek, S. (2019), "Automatic Detection and Uncertainty Quantification of Landmarks on Elastic Curves," *Journal of the American Statistical Association*, 114, 1002–1017. [2,6,7,12]

Subburaj, K., Ravi, B., and Agarwal, M. (2008), "3D Shape Reasoning for Identifying Anatomical Landmarks," *Computer-Aided Design and Applications*, 5, 153–160. [2]

Tabesh, A., Teverovskiy, M., Pang, H.-Y., Kumar, V. P., Verbel, D., Kotsianti, A., and Saidi, O. (2007), "Multifeature Prostate Cancer Diagnosis and Gleason Grading of Histological Images," *IEEE Transactions on Medical Imaging*, 26, 1366–1378. [1]

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian Variable Selection in Clustering High-Dimensional Data," *Journal of the American Statistical Association*, 100, 602–617. [3]

Wang, S., Chen, A., Yang, L., Cai, L., Xie, Y., Fujimoto, J., Gazdar, A., and Xiao, G. (2018), "Comprehensive Analysis of Lung Cancer Pathology Images to Discover Tumor Shape and Boundary Features that Predict Survival Outcome," *Scientific Reports*, 8, 1–9. [1,2,9]

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., and Snyder, M. (2016), "Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features," *Nature Communications*, 7, 12474. [1]

Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., Gräf, S., Chin, S.-F., Schwarz, R. F., Curtis, C., Dunning, M. J., Bardwell, H. et al. (2012), "Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling," *Science Translational Medicine*, 4, 157ra143. [1]

Zulqarnain Gilani, S., Shafait, F., and Mian, A. (2015), "Shape-based Automatic Detection of a Large Number of 3D Facial Landmarks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4639–4648. [2]