SPATIAL SHAPE ANALYSIS OF CELL NUCLEI*

YE JIN CHOI†, SEBASTIAN KURTEK†, AND KARTHIK BHARATH‡

Abstract. In a histopathology image, to assess if the shapes of tumor cell nuclei influence tumor infiltration in surrounding tissues it is important to first quantify and visualize variability in the shape of cell nuclei whilst accounting for their spatial correlations. To avoid loss of shape information when using summaries, we represent the nucleus boundary as a planar curve, and develop a geometric framework to compute local spatially-informed summaries of shapes of planar curves. The framework is based on the definition of a novel shape trace-variogram that captures dependence between shapes of curves. The variogram is used to compute a local spatially-weighted shape average and covariance, using which, shape variation is quantified via principal component analysis applied to the weighted covariance operator. Primary directions of spatial shape variation are subsequently visualized as vector fields (and their pointwise magnitudes) on the spatially-weighted average shape. Our approach serves as a practical exploratory tool for clinicians who routinely assess morphological and spatial heterogeneity of cell nuclei, and also represents the first step toward probabilistic modeling of spatially varying cell shapes. We demonstrate utility of the framework on synthetic data, and on data pertaining to cell nuclei from whole slide images of breast cancer tissues.

Relevance to Life Sciences. Pathologists routinely examine tumor regions in histopathology images to assess severity and progression of cancer. Nuclear shape heterogeneity plays a key role in this process and provides valuable information that guides clinical treatment decisions. To aid in such assessments, there is need for tools that enable statistical quantification, summarization and visualization of cell nucleus shape variation. While such tools exist for independent shape data, they are not applicable in the current context due to the inherent spatial correlation among cell nuclei. Thus, we define local spatially-informed shape summaries which, when computed based on cell nuclei within histopathology image regions, reveal their shape heterogeneity. The proposed framework is applied to whole slide images of breast cancer tissues.

Mathematical Content. Cell nucleus boundaries are represented using planar closed curves. Their shape is a property that is unaffected by translation, rescaling, rotation and reparameterization. To ensure that our framework is invariant to all shape-preserving transformations, we build on the Riemannian geometric elastic shape analysis framework. We first define a shape trace-variogram, which models spatial dependence among the shapes of planar curves. The trace-variogram is then used to define a spatially-weighted shape average and covariance, which enable exploration of shape variability in a local spatial region. Finally, we provide algorithms for estimation of the proposed spatially-informed summaries.

Key words. Shape Trace-variogram; Weighted Shape Average; Weighted Covariance and Principal Component Analysis

MSC codes. 62H11, 62R30, 62R10

1. Introduction.

Statistical shape analysis tools have become indispensable in many scientific domains, including medical imaging [3,6,28] and biology (morphometrics) [19, 21]. In these fields, researchers are often interested in understanding how the geometry or shape of structures observed in images vary across space, usually defined as (a subset of) the image domain, under different conditions. For example, in medical applications, the shape of organs or cells along with their spatial configurations can provide crucial information about physiological conditions, disease progression and treatment outcomes [12, 22, 23]. Pathologists have long relied on nuclear shape abnormalities to distinguish benign from malignant cells [7], demon-

*Submitted to the editors February 10, 2026.

Funding: This work was supported in part by NIH R37-CA214955 (SK and KB), NSF DMS-2015374 (KB), and NSF DMS-2413747 (SK).

†Department of Statistics, The Ohio State University, Columbus, OH (choi.1577@osu.edu, kurtek.1@stat.osu.edu).

‡School of Mathematical Sciences, University of Nottingham, Nottingham, UK (Karthik.Bharath@nottingham.ac.uk).

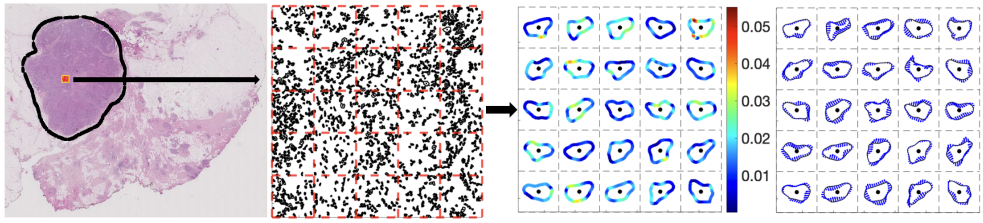


FIG. 1. *Local spatially-informed summarization of cell nucleus shapes in a histopathology image.*

strating the critical role of shape in characterizing cell states. In a tumor, these cells interact with stromal (supportive) cells and other components to form the tumor microenvironment (TME) [8, 13]. The TME exhibits a distinct structure featuring spatial patterns that can guide cancer grading and predict treatment outcomes. For instance, [9] showed that the location of immune cells in colorectal cancer is a key prognostic factor, while [25] discovered a correlation between the local spatial structure of tumor-infiltrating lymphocytes (TILs) and overall survival. These findings suggest that both, shape variation among cells and the spatial arrangement of TME components, hold valuable information for cancer research. Thus, there is a clear need for specialized methods capable of capturing spatially correlated shape data.

Many studies integrate cell-level morphological features with spatial information to characterize tumors. Some are based on nuclear shape/texture descriptors, e.g., size, irregularity or fractal features, combined with spatial measures like cell–cell distances to predict recurrence or treatment response [2,15,31]. [14] combined single-cell morphological features, e.g., area and eccentricity, with spatial interaction networks to identify multicellular communities predictive of breast cancer recurrence and survival. [33] classified individual nuclei into cancer, stromal and immune cell types based on morphology, and quantified their spatial distribution using Ripley's K-function, revealing prognostic patterns of stromal organization. However, all of these methods represent shapes using scalar summaries, which do not capture their full geometric structure.

When boundaries of objects observed in two-dimensional images are represented as parameterized curves, i.e., functions of the form $\beta : \mathcal{D} \to \mathbb{R}^2$ ($\mathcal{D}$ is the parameterization domain of β), their shape is a physical property that remains unchanged under translation, rescaling, rotation and reparameterization. Statistical shape analysis is then concerned with various tasks including (i) quantification of shape differences via a distance, (ii) statistical summarization, i.e., computation of a representative (sample average) shape and exploration of shape variability via principal component analysis (PCA), and (iv) shape modeling and inference. While shape analysis approaches for independent shape data are well-established, see e.g., [18, 26], methods which are appropriate for shapes that exhibit spatial dependence are scarce. Thus, we define a statistical framework for local summarization of spatially dependent shape data. The framework is accompanied by visualization tools, which enable easy interpretability. To model spatial dependence among shapes, we extend the concept of the trace-variogram, which has been effectively used for kriging and hierarchical clustering of spatial functional data [10, 11]. In essence, the proposed shape trace-variogram encodes the dependence of shape variation on spatial distance. It allows efficient estimation of weights for each observed shape that are used to define spatially-informed

shape summaries.

We provide an illustration of our approach in Figure 1. We start with a high-resolution histopathology image and focus on a region of interest (ROI, red square) inside a tumor area (black outline). We extract cell nucleus outlines and divide the ROI into equally sized, disjoint subregions. We then summarize shape variation of cell nuclei, while incorporating their spatial correlation, within each subregion. We compute a spatially-weighted shape average, based on a linear estimator that incorporates weights estimated from the shape trace-variogram. We further quantify local shape variation by performing PCA, using a spatially-weighted covariance matrix, on the tangent space at this average shape. Summarization results are displayed as PC vector fields, and their pointwise magnitudes, on the spatially-weighted mean shape. Our techniques build on elastic shape representations and metrics, wherein parameterization variability is removed via registration [20, 26, 27, 32].

We demonstrate the effectiveness of spatially-informed summarization using synthetic and real data examples. In the synthetic study, we generate spatially correlated planar shapes, and show that the spatially-weighted average shape and associated principal modes of variation are effective at summarizing the data. In the real data application, we analyze cell nucleus outlines extracted from histopathology images of breast cancer [1, 30], which are known to exhibit complex morphological and spatial variability [12]. These nuclear shapes, along with their spatial distribution, can offer insights into how cells interact under various disease conditions. Our results indicate that local spatially-informed summaries can uncover morphological differences across tissue regions, which may be overlooked by global summaries that do not consider spatial dependence. As such, we demonstrate the potential of the developed tools for broader biomedical investigations.

Our contributions are as follows.

1. We define a shape trace-variogram to model spatial dependence between the *shapes* of curves, a property that is invariant to translation, rescaling, rotation and reparameterization. The proposed variogram thus represents a novel extension of the functional trace-variograms proposed in [10, 11].
2. We use the shape trace-variogram to define a spatially-weighted shape average and covariance that can be used to explore shape variation in a local spatial region. The weights ensure that shapes close to a given spatial location have a stronger impact on the computed summaries.
3. We introduce concise, yet effective, visualizations of the spatially-informed summaries that aid in assessment of shape heterogeneity.
4. We apply our approach to histopathology images of breast cancer, resulting in meaningful summaries that effectively capture spatial shape heterogeneity of cell nuclei.

The rest of this article is organized as follows. Section 2 reviews summarization tools from elastic shape analysis under the assumption of independence. Section 3 introduces the shape trace-variogram and describes its use in the definition and computation of (i) a spatially-weighted shape average, and (ii) spatially-weighted covariance. It additionally outlines how the proposed summaries are visualized. Section 4 presents simulation results, while Section 5 describes an application to histopathology images of breast cancer. We close with a brief discussion in Section 6. The supplement includes (i) illustrations of the elastic shape analysis framework (Section 1), (ii) a proof of Proposition 3.1, which is related to the estimation of spatial weights via the shape trace-variogram (Section 2), (iii) empirical convergence analysis for Algorithm 3.1 (Section 3), (iv) additional simulation results and sensitivity analysis (Section 4),

and (v) additional real data results (Section 5).

2. Elastic Shape Analysis. We provide an overview of the elastic shape analysis framework; more details can be found in [26, 27]. Our particular focus is on the case of planar closed curves, i.e., curves without a natural start/end point, as this is the most pertinent case when analyzing cell nucleus shapes in histopathology images. Illustrations of the framework are given in Section 1 in the supplement. Throughout this section, we use $|\cdot|$ and $\langle\cdot,\cdot\rangle$ to denote the Euclidean norm and inner product in $\mathbb{R}^2$, and $\|\cdot\|$ and $\langle\langle\cdot,\cdot\rangle\rangle$ to denote the norm and inner product on $\mathbb{L}^2(\mathbb{S}^1,\mathbb{R}^2)$, hereafter denoted as $\mathbb{L}^2$, where $\mathbb{S}^1$ is the unit circle on the plane.

2.1. Representation, Pre-shape Space and Distance. Let $\mathcal{B} = \{\beta : \mathbb{S}^1 \to \mathbb{R}^2 \mid \beta \text{ is absolutely continuous}\}$ denote the space of planar closed curves. As briefly discussed in Section 1, the shape of β is unchanged under translation, rescaling, rotation and reparameterization. Thus, we seek a metric structure on $\mathcal{B}$ that is invariant to these shape-preserving transformations. The reparameterization group $\Gamma = \{\gamma : \mathbb{S}^1 \to \mathbb{S}^1 \mid \gamma \text{ is orientation-preserving diffeomorphism}\}$ acts on $\mathcal{B}$ via composition: $(\beta,\gamma) = \beta \circ \gamma$. It is easy to see that the usual $\mathbb{L}^2$ distance between two curves $\beta_1,\ \beta_2 \in \mathcal{B}$ is not invariant to the reparameterization action: $\|\beta_1 \circ \gamma - \beta_2 \circ \gamma\| \neq \|\beta_1 - \beta_2\|$ for all $\gamma \in \Gamma$ except $\gamma_{id}(t) = t$. This, in turn, means that the $\mathbb{L}^2$ distance on $\mathcal{B}$ is not suitable for parameterization-invariant comparison and modeling of shape.

To overcome the aforementioned challenge, we equip $\mathcal{B}$ with an elastic Riemannian metric [20]. Let $\dot{\beta}$ be the coordinatewise derivative of β, which exists almost everywhere on $\mathbb{S}^1$; set $\dot{\beta}(t) = 0$ on the set of all $t \in \mathbb{S}^1$ where it is not defined pointwise. Outside of this set assume $|\dot{\beta}| \neq 0$. Then, $p = |\dot{\beta}|$ and $\theta = \dot{\beta}|\dot{\beta}|^{-1}$ denote the speed and angle functions, which fully characterize $\dot{\beta}$ since $\dot{\beta} = p\theta$. Thus, this defines a bijective map (up to a translation) from $\mathcal{B}$ to the product space $\mathcal{P} \times \Theta$, where $\mathcal{P} = \{p : \mathbb{S}^1 \to \mathbb{R}_+\}$ and $\Theta = \{\theta : \mathbb{S}^1 \to \mathbb{S}^1\}$. Letting $(\delta p_1, \delta\theta_1)$ and $(\delta p_2, \delta\theta_2)$ denote two vectors in the tangent space at (p,θ), $T_{(p,\theta)}(\mathcal{P} \times \Theta)$, the family of elastic Riemannian metrics is defined as

(2.1)
$$\langle\langle(\delta p_1, \delta\theta_1), (\delta p_2, \delta\theta_2)\rangle\rangle_{(p,\theta)} := a \int_{\mathbb{S}^1} \delta p_1(t)\delta p_2(t)\frac{1}{p(t)}dt + b \int_{\mathbb{S}^1} \langle\delta\theta_1(t), \delta\theta_2(t)\rangle p(t)dt,$$

where $a,\ b > 0$. Note that the first term measures variations in the speed function while the second term measures variations in the angle function. The constants a and b provide relative weights for the two terms. An important property of this metric is that it is preserved when two curves are reparameterized in the same way; this is true for any choice of a and b. However, the resulting Riemannian distance does not have a closed-form expression and is difficult to compute.

To simplify the metric structure and retain the desirable property of invariance to reparameterization, we employ the square-root velocity function (SRVF) [27]. The SRVF of a curve $\beta \in \mathcal{B}$ is defined via the mapping $Q : \mathcal{B} \to \mathbb{L}^2$ as

$$q \equiv Q(\beta) := \begin{cases} \frac{\dot{\beta}}{\sqrt{|\dot{\beta}|}}, & |\dot{\beta}| \neq 0 \\ 0, & |\dot{\beta}| = 0 \text{ or } \dot{\beta} \text{ is not defined.} \end{cases} \tag{2.2}$$

A key consequence of this transformation is that the elastic Riemannian metric on $\mathcal{B}$, with $a = 1/4$ and $b = 1$, simplifies to the $\mathbb{L}^2$ metric on the space of SRVFs [27]. The SRVF of a closed curve β satisfies the following constraint: $\int_{\mathbb{S}^1} q(t)|q(t)|dt = 0$. Finally, the inverse mapping, Q^{-1}, can be computed (up to a translation) using

$\int_0^t q(u)|q(u)|du$, where the domain $\mathbb{S}^1$ has been unwrapped to the interval $[0, 2\pi]$; note that this inverse mapping subsumes projection of the SRVF q into the space of closed curves using Algorithm 18 in [26, Ch. 6] to ensure that the aforementioned closure condition is satisfied.

We now return to the problem of removing nuisance variation from the representation space. Translation is automatically removed under the SRVF since its definition only involves the derivative of β. To remove scale, we normalize all curves to have unit length, resulting in SRVFs with unit $\mathbb{L}^2$ norm: $L(\beta) = \int_{\mathbb{S}^1} |\dot{\beta}(t)|dt = \int_{\mathbb{S}^1} |q(t)|^2 dt = \|q\|^2 = 1$. This leads to the pre-shape space for planar closed curves, defined as $\mathcal{C} = \{q \in \mathbb{L}^2 \mid \|q\| = 1,\ \int_{\mathbb{S}^1} q(t)|q(t)|dt = 0\}$, which is a subset of the Hilbert sphere. The space $\mathcal{C}$ is endowed with the ambient $\mathbb{L}^2$ metric, resulting in the extrinsic pre-shape distance $d_{\mathcal{C}}(q_1, q_2) = \|q_1 - q_2\|$ between $q_1,\ q_2 \in \mathcal{C}$[2]. Removal of rotation and parameterization variability cannot be achieved via normalization. Instead, we use the notion of equivalence classes as a formal definition of shape.

2.2. Registration, Shape Space and Distance. Let $O \in SO(2) = \{R \in \mathbb{R}^{2\times 2} \mid RR^\top = R^\top R = I_2,\ \det(R) = 1\}$ represent a rotation matrix. The rotation of a curve $\beta \in \mathcal{B}$ by $O \in SO(2)$ is given by $O\beta$. The corresponding transformation of the SRVF of β, $Q(\beta) = q$, is the same, i.e., Oq. The reparameterization of β via γ, $(\beta, \gamma) = \beta \circ \gamma$, results in the following transformation of its SRVF: $(q, \gamma) = (q \circ \gamma)\sqrt{\dot{\gamma}}$. It is now easy to verify that the distance $d_{\mathcal{C}}$ is preserved if two curves $\beta_1,\ \beta_2 \in \mathcal{B}$ are reparameterized using the same $\gamma \in \Gamma$ and rotated using the same $O \in SO(2)$ $(q_1 = Q(\beta_1),\ q_2 = Q(\beta_2))$:

$$\|O(q_1, \gamma) - O(q_2, \gamma)\| = \|q_1 - q_2\|. \tag{2.3}$$

To account for rotation and parameterization variation, we define equivalence classes of the form $[q] := \{O(q, \gamma) \mid O \in SO(2),\ \gamma \in \Gamma\}$, which serve as the definition of shape. In other words, all possible rotations and reparameterizations of a curve are unified via an equivalence class, which uniquely represents the shape of that curve. This leads to the definition of the shape space as the quotient space $\mathcal{S} = \mathcal{C}/(SO(2) \times \Gamma) = \{[q]\}$. The extrinsic shape distance on $\mathcal{S}$ is defined via the distance $d_{\mathcal{C}}$ as (for $[q_1],\ [q_2] \in \mathcal{S}$)

$$d_{\mathcal{S}}([q_1], [q_2]) = \inf_{(O,\gamma)\in SO(2)\times\Gamma} \|q_1 - O(q_2, \gamma)\|. \tag{2.4}$$

To compute this distance, we have to determine an optimal rotation $\hat{O}$ and an optimal reparameterization $\hat{\gamma}$ of q_2 with respect to q_1. Optimization over $SO(2)$ is performed via Procrustes analysis [16]. Optimization over Γ is split into (i) an exhaustive search over starting points on $\mathbb{S}^1$, and (ii) Dynamic Programming given each starting point [24]. The joint optimization over $SO(2) \times \Gamma$ is referred to as elastic registration, with $\hat{O}(\beta_2 \circ \hat{\gamma})$ being optimally registered to β_1, and ensures that shape comparisons are natural, i.e., they preserve geometric features of the shapes being compared.

2.3. Shape Summaries Under Independence. Let $\beta_1, \ldots, \beta_n \in \mathcal{B}$ denote a sample of *independent* planar closed curves with SRVFs $q_1, \ldots, q_n \in \mathcal{C}$. Their extrinsic

[2]Under the SRVF, a specific instance of the elastic Riemannian metric simplifies to the $\mathbb{L}^2$ metric. When restricted to the pre-shape space of closed curves, $\mathcal{C}$, the resulting *intrinsic* distance does not have a closed-form expression [27]. However, throughout this manuscript, we employ the *extrinsic* distance $d_{\mathcal{C}}$, which does have a closed-form expression and results in simplified computation.

Algorithm 2.1 Extrinsic Karcher Mean Shape

Input: $q_1, \ldots, q_n$.
Output: $\hat{\mu}_q$; $\hat{O}_1(q_1, \hat{\gamma}_1), \ldots, \hat{O}_n(q_n, \hat{\gamma}_n)$.
Initialize $k = 1$, $\epsilon > 0$ (small), $\delta > 0$, $\hat{\mu}_q^{(1)}$, $\|\bar{z}^{(0)}\| > 0$ (large).
while $\|\bar{z}^{(k-1)}\| > \epsilon$ **do**
  Register each q_i to $\hat{\mu}_q^{(k)}$ using (2.4) to obtain $\hat{O}_i^{(k)}(q_i, \hat{\gamma}_i^{(k)})$.
  Compute deviation vector from $\hat{\mu}_q^{(k)}$ for each $\hat{O}_i^{(k)}(q_i, \hat{\gamma}_i^{(k)})$: $z_i = \hat{O}_i^{(k)}(q_i, \hat{\gamma}_i^{(k)}) - \hat{\mu}_q^{(k)}$.
  Compute average direction: $\bar{z}^{(k)} = \frac{1}{n}\sum_{i=1}^n z_i$.
  Update: $\hat{\mu}_q^{(k+1)} \leftarrow \hat{\mu}_q^{(k)} + \delta \bar{z}^{(k)}$.
  Normalize: $\hat{\mu}_q^{(k+1)} \leftarrow \frac{\hat{\mu}_q^{(k+1)}}{\|\hat{\mu}_q^{(k+1)}\|}$.
  Project $\hat{\mu}_q^{(k+1)}$ using Algorithm 18 in [26, Ch. 6] to ensure that closure condition is satisfied.
  Set $k \leftarrow k + 1$.
end while
Register each q_i to $\hat{\mu}_q$ using (2.4) to obtain $\hat{O}_i(q_i, \hat{\gamma}_i)$.

sample Karcher mean shape, $[\hat{\mu}_q]$, is defined as a minimizer of

$$\mathcal{S} \ni [q] \mapsto \sum_{i=1}^{n} d_{\mathcal{S}}([q], [q_i])^2. \tag{2.5}$$

In practice, we select a single element $\hat{\mu}_q \in [\hat{\mu}_q]$ as a representative for further analyses. The mean can be visualized after inversion, $\hat{\mu}_\beta = Q^{-1}(\hat{\mu}_q)$. The procedure for computing the extrinsic Karcher mean shape is presented as Algorithm 2.1.

To explore variation in a sample of shapes, we compute the covariance matrix in the tangent space $T_{\hat{\mu}_q}(\mathcal{S})$ as follows. With respect to the extrinsic spherical metric on $\mathcal{C}$, we first project the registered curves (shape representatives) $q_1^* = \hat{O}_1(q_1, \hat{\gamma}_1), \ldots, q_n^* = \hat{O}_n(q_n, \hat{\gamma}_n)$ to $T_{\hat{\mu}_q}(\mathcal{S})$ using the inverse-exponential map [26, Ch. 3]: $v_i := \exp_{\hat{\mu}_q}^{-1}(q_i^*) = \theta \sin(\theta)^{-1}\left(q_i^* - \cos(\theta)\hat{\mu}_q\right)$, where $\theta = \cos^{-1}\left(\langle\langle \hat{\mu}_q, q_i^* \rangle\rangle\right)$. While theoretically shape is an infinite-dimensional object, at the implementation stage, it is approximated using a finite number of points T. Thus, each v_i is a matrix of size $2 \times T$. We further reshape v_i to a vector of size $2T$ resulting in $v_i^* \in \mathbb{R}^{2T}$. The sample shape covariance matrix is then given by

$$\hat{K} := \frac{1}{n-1}\sum_{i=1}^{n} v_i^* v_i^{*\top} \in \mathbb{R}^{2T \times 2T}. \tag{2.6}$$

To perform PCA, we apply singular value decomposition (SVD) to the sample covariance matrix to obtain $\hat{K} = U\Sigma U^\top$. The columns of U contain the principal directions of shape variation; the diagonal elements of Σ are the variances along each direction (ordered from largest to smallest). The total sample shape variance can be computed using the trace of Σ.

To visualize shape variation along the kth principal direction, we compute a displacement, in units of standard deviation, from the mean shape on the tangent space and map it back to the shape space via the exponential map, $q_k^l = \exp_{\hat{\mu}_q}\left(l\sqrt{\Sigma_{kk}}\, U_k\right)$,

where l controls the size of the displacement, Σ_{kk} is the kth diagonal element of Σ, and U_k is the kth column of U reshaped to a matrix of size $2 \times T$. For $v \in T_{\hat{\mu}_q}(\mathcal{S})$, the exponential map is given by $\exp_{\hat{\mu}_q}(v) = \cos(\|v\|)\,\hat{\mu}_q + \sin(\|v\|)\,v\|v\|^{-1}$ [26, Ch. 3]. As before, the principal directions of shape variation can be visualized after inversion, $\beta_k^l = Q^{-1}(q_k^l)$.

3. Spatially-informed Shape Summaries. While the elastic shape analysis framework described in Section 2 is effective in summarizing independent shape data, as illustrated in Section 1 in the supplement, it does not account for potential dependence among observed shapes. Depending on the specific application of interest, shape dependence may arise in many different ways. However, in the current work, we focus on spatial dependence among cell nucleus shapes in histopathology images. Thus, we introduce a new framework, with elastic shape analysis serving as a building block, that leverages local spatial dependence for shape summarization.

Let $X \subset \mathbb{R}^2$ denote the spatial domain; our focus here is on two-dimensional histopathology images, but the concepts presented in this section can be extended to higher-dimensional (and irregular) spatial domains. In spatial statistics, the variogram [5] is a core tool for summarizing dependence among spatially-indexed objects. In particular, [10] proposed the trace-variogram for functional data. Under the assumption that the functional random field $\{f_s,\ s \in X\}$ with values in $\mathbb{L}^2([a,b],\mathbb{R})$ is second-order stationary and isotropic[3], the trace-variogram is defined as

$$V_f(h) := \frac{1}{2}E\Big[\|f_s - f_{s'}\|^2\Big], \tag{3.1}$$

where $h = |s - s'|$ denotes the Euclidean distance between spatial locations s and s'. In this definition, each spatially indexed function f_s is univariate with domain given by the interval $[a,b]$. The trace-variogram V_f is a function of the spatial distance h, and captures variation among the functional random field at different distance lags via the $\mathbb{L}^2$ distance. Then, given a realization of the random field $\{f_{s_i},\ s_i \in X\}$ $(i = 1,\ldots,n)$ at n locations, one can compute the empirical trace-variogram using

$$\hat{V}_f(h) = \frac{1}{2|N(h)|}\sum_{i,j \in N(h)} \|f_{s_i} - f_{s_j}\|^2, \tag{3.2}$$

where $N(h) = \{(s_i, s_j) \mid |s_i - s_j| = h\}$ is a spatial neighborhood. For irregularly spaced functional data, $N(h)$ is often modified to $N(h) = \{(s_i, s_j) \mid |s_i - s_j| \in (h-\epsilon, h+\epsilon)\}$.

If there exists spatial dependence among the functions $\{f_{s_i}\}$, the values of the empirical trace-variogram will be smaller at shorter spatial distance lags. This is due to the fact that functions that are located close to each other in the spatial domain X will exhibit less variation than those located far away from each other. The trace-variogram allows estimation of spatial weights for the sample $\{f_{s_i}\}$, which can be subsequently used for prediction at a particular location and hierarchical clustering of spatial functional data. The definition of the trace-variogram was extended to the case of spatial functional data with phase variation in [11]. In this work, we define a *shape trace-variogram* to construct local shape summaries: (i) a spatially-weighted shape average, and (ii) a spatially-weighted covariance matrix for PCA.

[3]Recall that a second-order stationary and isotropic random field has a constant mean, finite variance, and covariance that only depends on the spatial distance between observations.

3.1. Shape Trace-variogram and its Estimation. We begin with a spatial random field of planar closed curves, $\{\beta_s,\ s \in X\}$ with values in $\mathcal{B}$, which is transformed to a spatial random field of SRVFs, $\{q_s,\ s \in X\}$ with values in $\mathcal{C} \subset \mathbb{L}^2$. In most applications, for example those involving cell nucleus shapes in histopathology images, assuming that the field $\{q_s,\ s \in X\}$ is second-order stationary with constant mean shape is inappropriate. As such, the spatially-weighted average shape we wish to estimate is the shape of the value $q_{s_0} \equiv q_0$ of the SRVF random field at a *fixed, known location* $s_0 \in X$; for all results presented in Sections 4 and 5 we choose s_0 to be the center of the spatial domain X. The estimate will be a local spatially-weighted average of the shapes of observed values of $\{q_s,\ s \in X\}$ in a neighborhood of s_0, and is referred to within the spatial statistics literature as a kriging predictor at s_0.

The shape of the random field $\{q_s,\ s \in X\}$ is encoded within the random field $\{[q_s],\ s \in X\}$ of equivalence classes. This object is difficult to work with in practice: probability distributions on the shape space $\mathcal{S}$ may behave in non-standard ways due to the infinite-dimensionality of the equivalence class $[q_s]$ for each $s \in X$ and that of the non-compact reparameterization group Γ. To avoid such difficulties, we instead work with the transformed field $\{O_s(q_s, \gamma_s) \in [q_s],\ s \in X\}$ that represents, for each $s \in X$, the registration of the field $\{q_s,\ s \in X\}$ to a preferred element $O_{s_0}(q_{s_0}, \gamma_{s_0}) \equiv O_0(q_0, \gamma_0)$ of the shape $[q_0]$, where $\{O_s,\ s \in X\}$ and $\{\gamma_s,\ s \in X\}$ are, respectively, random fields with values in $SO(2)$ and Γ satisfying

$$(O_s, \gamma_s) = \underset{(O,\gamma) \in SO(2)\times\Gamma}{\arg\inf} \|O(q_s, \gamma) - O_0(q_0, \gamma_0)\|^2, \quad s \in X.$$

Abusing terminology, we refer to $\{O_s(q_s, \gamma_s) \in [q_s],\ s \in X\}$, consisting of shape representatives for $\{q_s,\ s \in X\}$, as the *shape random field*[4], which implicitly depends on the representative $O_0(q_0, \gamma_0)$ of the shape $[q_0]$ at location s_0. At a different location, $\tilde{s}_0$, we obtain a different shape SRVF random field defined for the purpose of estimating a different spatially-weighted average shape $O_{\tilde{s}_0}(q_{\tilde{s}_0}, \gamma_{\tilde{s}_0})$.

As done for the functional random field in [10], we assume that the shape random field is second-order stationary and isotropic, and accordingly define the *shape trace-variogram*

$$V_{sh}(h) := \frac{1}{2} E\Big[\|O_s(q_s, \gamma_s) - O_{s'}(q_{s'}, \gamma_{s'})\|^2\Big], \tag{3.3}$$

where $h = |s - s'|$. It is important to note that the definition of the shape trace-variogram V_{sh} is novel and can be viewed as a modification of the functional trace-variogram in (3.1) that incorporates registration: the presence of shape-preserving nuisance variation in $\{q_s,\ s \in X\}$ is handled via the random fields $\{O_s,\ s \in X\}$ and $\{\gamma_s,\ s \in X\}$, which determine the shape random field via the registration of $\{q_s,\ s \in X\}$ to the target $O_0(q_0, \gamma_0)$. Thus, importantly, the definition of the shape trace-variogram implicitly depends on the chosen spatial location, s_0, through the estimation target $O_0(q_0, \gamma_0)$. As such, one can define a family of shape trace-variograms by varying the location s_0 resulting in different shape random fields and spatially-weighted average shapes.

Similarly to the functional trace-variogram, the shape trace-variogram captures variation among the shape random field at different spatial distance lags. Evidently,

[4]Note that the original random field $\{q_s\}$ and the shape random field $\{O_s(q_s, \gamma_s)\}$ are both random fields of SRVFs. Their difference lies in the fact that the latter depends on the registration of $\{q_s\}$ to $O_0(q_0, \gamma_0)$, and models the variation in $\{q_s\}$ relative to the spatially-weighted average $O_0(q_0, \gamma_0)$.

the shape trace-variogram is invariant to simultaneous rotation and reparameterization of $\{q_s,\ s \in X\}$ using any $(O,\gamma) \in SO(2) \times \Gamma$, which is a direct consequence of the isometry property verified in (2.3): it captures pure spatial shape variation.

We now define an empirical version of the shape trace-variogram, needed to compute an estimator, $\tilde{q}_0$, of the spatially-weighted shape average $O_0(q_0, \gamma_0)$. Let $\{q_{s_i},\ s_i \in X\}$ denote an observed sample of n curves represented by their SRVFs. The definition requires the estimator $\tilde{q}_0$, which will be defined in the sequel. To simplify notation, we use the subscript i for the spatial location s_i. Then, the empirical shape trace-variogram is defined as

$$\hat{V}_{sh}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \|\hat{O}_i(q_i, \hat{\gamma}_i) - \hat{O}_j(q_j, \hat{\gamma}_j)\|^2, \tag{3.4}$$

where $N(h) = \{(s_i, s_j) \mid |s_i - s_j| \in (h - \epsilon, h + \epsilon)\}$ and $(\hat{O}_i, \hat{\gamma}_i)$ minimize $(O,\gamma) \mapsto \|\tilde{q}_0 - O(q_i, \gamma)\|$ for $i = 1, \ldots, n$. To ensure that the estimated variogram is conditionally negative definite [5], we additionally fit a Matérn model (with a predetermined smoothness) to $\hat{V}_{sh}$ using ordinary least squares. The resulting Matérn fit is used in subsequent analyses, and with a slight abuse in notation, will henceforth be denoted by $\hat{V}_{sh}$.

3.2. Estimation of Spatially-weighted Average Shape. In (3.4), $\hat{O}_i(q_i, \hat{\gamma}_i)$ serves the role of a representative of the shape equivalence class $[q_i]$. Given $(\hat{O}_i, \hat{\gamma}_i)$ $(i = 1, \ldots, n)$, the empirical shape trace-variogram (and the Matérn fit) is computed once the estimator $\tilde{q}_0$ is prescribed. On the other hand, definition of the kriging estimator $\tilde{q}_0$ *requires* using $\hat{V}_{sh}$. We now detail the issue and present an algorithm that enables computation of $\hat{V}_{sh}$ while simultaneously allowing for the computation of $\tilde{q}_0$ and $(\hat{O}_i, \hat{\gamma}_i)$ $(i = 1, \ldots, n)$ via registration to $\tilde{q}_0$.

For a given vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top \in \Delta := \{\boldsymbol{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1\}$, the spatially-weighted average shape estimator, $\tilde{q}_0 \in [\tilde{q}_0]$, is defined as a minimizer of

$$q \mapsto \sum_{i=1}^{n} \eta_i d_{\mathcal{S}}([q], [q_i])^2, \tag{3.5}$$

which is directly available as

$$\tilde{q}_0 = \frac{\sum_{i=1}^n \eta_i \hat{O}_i(q_i, \hat{\gamma}_i)}{\| \sum_{i=1}^n \eta_i \hat{O}_i(q_i, \hat{\gamma}_i)\|}. \tag{3.6}$$

For clarity of exposition, for just this section, we make explicit the dependence of the estimator $\tilde{q}_0$ on $\boldsymbol{\eta}$ and denote it by $\tilde{q}_0(\boldsymbol{\eta})$. Note that $\tilde{q}_0(\boldsymbol{\eta})$ is not a weighted sample Karcher mean in a strict sense since some of the elements of $\boldsymbol{\eta}$ may be negative. On the other hand, given the nuisance transformations $(\hat{O}_i, \hat{\gamma}_i)$ $(i = 1, \ldots, n)$, the coefficient vector $\boldsymbol{\eta}$ is obtained by minimizing the expected shape prediction error functional:

$$\boldsymbol{\eta} = \underset{\boldsymbol{x} \in \Delta}{\operatorname{argmin}}\ E\left[\|\tilde{q}_0(\boldsymbol{x}) - O_0(q_0, \gamma_0)\|^2\right]. \tag{3.7}$$

If $(\hat{O}_i, \hat{\gamma}_i)$ $(i = 1, \ldots, n)$ can be recovered exactly, the optimization problem in (3.7) is equivalent to minimizing $\boldsymbol{x} \mapsto \boldsymbol{x}^\top V_{sh} \boldsymbol{x}$, where $V_{sh} = V_{sh}(h_{i0}) + V_{sh}(h_{0j}) - V_{sh}(h_{ij})$, $h_{ij} = |s_i - s_j|$ is the population shape trace-variogram in (3.3). Hence, under this setting, the spatially-weighted average shape estimator depends only on the shape

Algorithm 3.1 Spatially-weighted Average Shape

Input: $(s_1, q_1), \ldots, (s_n, q_n)$; $s_0 \in X$.
Output: $\tilde{q}_0$; $\hat{V}_{sh}$; $\boldsymbol{\eta}$; $\hat{O}_1(q_1, \hat{\gamma}_1), \ldots, \hat{O}_n(q_n, \hat{\gamma}_n)$.
Initialize $k = 1$, $\epsilon > 0$ (small), $\delta > 0$, $\tilde{q}_0^{(0)}$ as q_j where $j = \underset{i \in \{1,\ldots,n\}}{\operatorname{argmin}} |s_0 - s_i|$, $\|\bar{z}^{(0)}\| > 0$ (large).
while $\|\bar{z}^{(k-1)}\| > \epsilon$ **do**
    Register each q_i to $\tilde{q}_0^{(k)}$ using (2.4) to obtain $\hat{O}_i^{(k)}(q_i, \hat{\gamma}_i^{(k)})$.
    Compute $\hat{V}_{sh}^{(k)}$, using $\{\hat{O}_i^{(k)}(q_i, \hat{\gamma}_i^{(k)})\}$ and (3.4), and corresponding Matérn fit.
    Estimate weight vector $\boldsymbol{\eta}^{(k)}$ using $\hat{V}_{sh}^{(k)}$.
    Compute deviation vector from $\tilde{q}_0^{(k)}$ for each $\hat{O}_i^{(k)}(q_i, \hat{\gamma}_i^{(k)})$: $z_i = \hat{O}_i^{(k)}(q_i, \hat{\gamma}_i^{(k)}) - \tilde{q}_0^{(k)}$.
    Compute spatially-weighted average direction: $\bar{z}^{(k)} = \frac{1}{n} \sum_{i=1}^n \eta_i^{(k)} z_i$.
    Update: $\tilde{q}_0^{(k+1)} \leftarrow \tilde{q}_0^{(k)} + \delta \bar{z}^{(k)}$.
    Normalize: $\tilde{q}_0^{(k+1)} \leftarrow \frac{\tilde{q}_0^{(k+1)}}{\|\tilde{q}_0^{(k+1)}\|}$.
    Project $\tilde{q}_0^{(k+1)}$ using Algorithm 18 in [26, Ch. 6] to ensure that closure condition is satisfied.
    Set $k \leftarrow k + 1$.
end while
Register each q_i to $\tilde{q}_0$ using (2.4) to obtain $\hat{O}_i(q_i, \hat{\gamma}_i)$.
Compute $\hat{V}_{sh}$, using $\{\hat{O}_i(q_i, \hat{\gamma}_i)\}$ and (3.4), and corresponding Matérn fit.
Estimate weight vector $\boldsymbol{\eta}$ using $\hat{V}_{sh}$.

trace-variogram V_{sh}. In practice, the minimization problem is solved using the empirical trace-variogram $\hat{V}_{sh}$. This is also discussed in detail in [11] for functional data with phase variation. A formal statement of this result is provided as Proposition 3.1; the proof is included in Section 2 in the supplement.

PROPOSITION 3.1. *Under the assumption that (O_i, γ_i) can be estimated exactly such that $\hat{O}_i(q_i, \hat{\gamma}_i) = O_i(q_i, \gamma_i)$, the vector of coefficients $\boldsymbol{\eta} \in \Delta$ that minimizes $\boldsymbol{x} \mapsto E\left[\|\tilde{q}_0(\boldsymbol{x}) - O_0(q_0, \gamma_0)\|^2\right]$ also minimizes $\boldsymbol{x}^\top V_{sh} \boldsymbol{x}$, where the matrix $V_{sh} \in \mathbb{R}^{n \times n}$ is composed of elements $V_{sh} = V_{sh}(h_{i0}) + V_{sh}(h_{0j}) - V_{sh}(h_{ij})$, $h_{ij} = |s_i - s_j|$ $(i, j = 1, \ldots, n)$.*

Remark. Evidently, the spatially-weighted average shape estimation problem cannot be defined using a single cost function. Estimation of $\tilde{q}_0(\boldsymbol{\eta})$ and $(\hat{O}_i, \hat{\gamma}_i)$ $(i = 1, \ldots, n)$ is only possible if the coefficient vector $\boldsymbol{\eta}$ is known. Similarly, estimation of $\boldsymbol{\eta}$ requires the empirical shape trace-variogram, which in turn depends on $\tilde{q}_0(\boldsymbol{\eta})$ and $(\hat{O}_i, \hat{\gamma}_i)$ $(i = 1, \ldots, n)$. Thus, as discussed next, we utilize an iterative algorithm that alternates between the two optimization problems.

Unfortunately, exact recovery of rotation or reparameterization is difficult when the observed curves exhibit complex noise; see [4, 17] for more details in the case of functional data with phase variation, which is similar to the shape setting we consider. In addition, accurate estimation of these nuisance transformations depends on accurate recovery of the underlying spatially-weighted average shape. We propose an iterative procedure, detailed in Algorithm 3.1, which simultaneously estimates the (i) spatially-weighted average shape, and (ii) the shape trace-variogram and associated

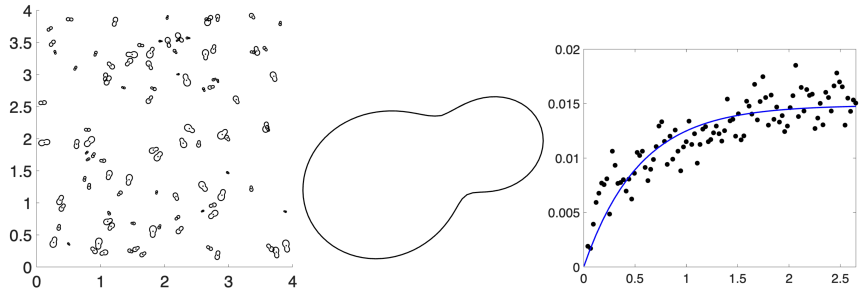


FIG. 2. *(a) Simulated curves with spatially varying shapes. (b) Spatially-weighted average shape. (c) Empirical shape trace-variogram (black) and fitted Matérn model (blue).*

vector of spatial weights $\boldsymbol{\eta}$. First, we initialize the shape average $\tilde{q}_0(\boldsymbol{\eta})$ as the SRVF observed at a spatial location that is closest to s_0. Empirically, we have found that the final estimate of the spatially-weighted average shape is not sensitive to initialization of Algorithm 3.1. We then iterate between the following steps until convergence: (i) registration of the sample SRVFs $\{q_i\}$ to the current shape average using (2.4), (ii) estimation of the shape trace-variogram using (3.4) and corresponding Matérn fit, (iii) computation of the ceofficient vector $\boldsymbol{\eta}$, and (iv) incremental update of the shape average using weighted deviations from the current estimate.

Recall that, under independence, computation of the Karcher mean shape of the sample SRVFs $q_1, \ldots, q_n$ (Section 2.3; Algorithm 2.1) uses equal weights for each sample. Under spatial dependence, we estimate a spatially-weighted average, where the weight (coefficient) for each sample is determined by the shape trace-variogram $\hat{V}_{sh}$. In particular, spatial dependence among the shapes of $\{q_i\}$ is incorporated via the trace-variogram estimate at each iteration k, $\hat{V}_{sh}^{(k)}$, to determine the weight vector $\boldsymbol{\eta}^{(k)}$. Thus, the algorithm performs spatially local registration of curves, with respect to a spatially-weighted average, that results in their shape representatives. Empirically, we have not observed any issues with convergence of Algorithm 3.1; we leave the formal convergence analysis as future work. Section 3 in the supplement presents convergence plots for a few simulated examples.

Figure 2 shows an example of estimation of a spatially-weighted average shape, and the associated shape trace-variogram, using Algorithm 3.1. Simulated spatially-indexed planar closed curves are shown in (a), where $X = [0, 4] \times [0, 4]$; here, the sample size is $n = 100$. Panel (b) displays the estimated spatially-weighted average shape $\tilde{q}_0$ at spatial location $s_0 = (2, 2)$. In (c), we show the final empirical shape trace-variogram (black points) and the Matérn fit with smoothness equal to 0.5.

3.3. Spatially-weighted Covariance and PCA. In addition to estimating a spatially-weighted average shape, we want to characterize and visualize shape variation in the sample. The vector $\boldsymbol{\eta}$ estimated by Algorithm 3.1 provides spatial weights associated with shapes $\{q_i^* = \hat{O}_i(q_i, \hat{\gamma}_i)\}$, which can be used to incorporate the contribution of each shape to the local variation around the spatially-weighted average via a weighted covariance matrix.

As in Section 2.3, the weighted covariance matrix is computed on the tangent space at the spatially-weighted average shape, $T_{\tilde{q}_0}(\mathcal{S})$. We project $\{q_i^*\}$ to $T_{\tilde{q}_0}(\mathcal{S})$ via the inverse-exponential map, resulting in $\{v_i\}$. At the implementation stage, each v_i is a $2 \times T$ matrix, which is reshaped to a vector of size $2T$, resulting in v_i^*. The

weighted covariance matrix is defined as

$$\hat{K}_{\boldsymbol{\eta}} = \frac{n}{n-1} \sum_{i=1}^{n} |\eta_i| v_i^* v_i^{*\top} \in \mathbb{R}^{2T \times 2T}. \tag{3.8}$$

In this definition, shape q_i^* is assigned a weight of $n|\eta_i|$. The reasoning behind this weighting scheme is as follows. First, to ensure that the weighted covariance matrix is non-negative definite we require each weight to be positive, but in practice, some shapes may be assigned a small negative weight. Hence, we use the absolute value of the weights. Second, scaling each η_i by n ensures that the weights sum to n, which matches the unweighted scenario described in Section 2.3. One can alter the form of the weighted covariance matrix to accommodate other weighting schemes, e.g., weights proportional to η_i^2. As in the unweighted case, PCA is performed via SVD of $\hat{K}_\eta$, resulting in spatially-informed principal directions of variation and the variances along each direction. Additionally, the overall spatially-weighted shape variance can be computed using the sum of the singular values of $\hat{K}_\eta$.

3.4. Visualization of Spatial Shape Variation. In Section 5 that considers real data, we first select an ROI within a histopathology image. Motivated by the assumption of second-order stationarity and isotropy of the underlying shape random field, which is more reasonable on a smaller spatial domain, we further partition this region into disjoint, equally sized subregions. The size/number of subregions is user and application specific. Spatially-informed summarization of cell nucleus shapes is then performed independently within each subregion. We first compute a spatially-weighted average shape for each subregion. Because these averages are computed independently across subregions, they may exhibit nuisance rotation and reparameterization variation. Thus, we additionally optimally register them with respect to their shape Karcher mean (Algorithm 2.1); while this step is not necessary it allows for easier interpretation of the local shape summaries. We then estimate the weighted covariance matrix and perform PCA. Instead of visualizing deformation paths along each principal direction, we instead display the deformation vector field associated with each direction, and its pointwise magnitude, on the spatially-weighted average. This provides concise displays of spatially-informed shape variation in each subregion.

Let U_k (reshaped to a matrix of size $2 \times T$) and Σ_{kk} denote the kth spatially-weighted principal direction of shape variation and the variance along this direction, respectively, resulting from SVD of $\hat{K}_{\boldsymbol{\eta}}$. We compute $q_k^1 = \exp_{\tilde{q}_0}(\sqrt{\Sigma_{kk}} U_k)$, i.e., the SRVF that is one standard deviation away from the spatially-weighted shape average $\tilde{q}_0$ along the kth direction. Next, we invert both $\tilde{q}_0$ and q_k^1 resulting in $\tilde{\beta}_0 = Q^{-1}(\tilde{q}_0)$ and $\beta_k^1 = Q^{-1}(q_k^1)$, respectively. Since the inverse mapping is only determined up to a translation, we additionally center them such that their centroids are at the origin in $\mathbb{R}^2$. Then, the vector field is defined as $\beta_k^1 - \tilde{\beta}_0$, i.e., the difference between the centered shapes; for display purposes and easy interpretation, the vector field is normalized. In addition, we compute the magnitude of the resulting vector field at each point along the domain: $|\beta_k^1(t) - \tilde{\beta}_0(t)| \; \forall \, t \in \mathbb{S}^1$. The vector field and its pointwise magnitude are finally displayed on the estimated spatially-weighted average shape $\tilde{\beta}_0$.

4. Simulation Studies. We evaluate the proposed spatially-informed approach for summarization of shape variation using two simulation studies. In the first, we show that if the underlying shapes are indeed correlated in space, then the spatially-weighted shape average is a more accurate estimate of the underlying ground truth

shape average than the Karcher mean shape. In the second, we provide qualitative and quantitative comparisons of spatially-weighted shape variation and variation computed under the assumption of independence.

In both simulations, we generate dependent shapes as follows. We first define the spatial domain $X = [0, D] \times [0, D]$ and sample n spatial locations uniformly over this domain. We generate spatially correlated shapes at the sampled locations using the Fourier basis functions $\{\cos(lt),\ \sin(lt),\ l = 0, 1, 2,\ t \in [-\pi, \pi]\}$. Letting $\{b_j,\ j = 1, \ldots, 6\}$ denote the basis functions, the radial distance as a function of the angle for shape i is defined as $\Omega_i(t) = \sum_{j=1}^{6} \nu_{ij} b_j(t)$, where $\{\nu_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, 6\}$ are Fourier coefficients. Each shape is obtained by transforming the radial distances to Cartesian coordinates $(x_i(t) = \Omega_i'(t)\cos(t),\ y_i(t) = \Omega_i'(t)\sin(t))$, where $\Omega_i'(t) = \Omega_i(t) + |\min_t \Omega_i(t)| + |\max_t \Omega_i(t)|$ is an adjusted radial function that ensures that the resulting shapes are non-self-intersecting. Spatial dependence between the shapes is imposed using the Fourier coefficients ν_{ij}. Given a mean vector for the coefficients, $\boldsymbol{m} = (m_1, \ldots, m_6)^\top$, we independently sample them for each j from the multivariate normal distribution, $\text{MVN}(m_j 1_n^\top, C_{mat})$, where $1_n^\top$ is an n-dimensional vector of ones, and C_{mat} is the Matérn covariance. The Matérn covariance has three parameters. In all simulations, we fix the scale parameter to 1 and the smoothness parameter to 0.5. The range parameter will depend on the simulation setting.

We further add nuisance variation to the simulated shapes. We generate random rescalings c_i independently from the $\text{Unif}(0.2, 0.6)$ distribution. Random rotations are represented by angles ϕ_i, which are sampled independently from the $\text{Unif}(0, 2\pi)$ distribution. The resulting spatially dependent shapes, with additional random rotations and rescalings, are given by $\beta_i(t) = (x_i^*(t) = c_i[x_i(t)\cos(\phi_i) - y_i(t)\sin(\phi_i)],\ y_i^*(t) = c_i[x_i(t)\sin(\phi_i) + y_i(t)\cos(\phi_i)]$.

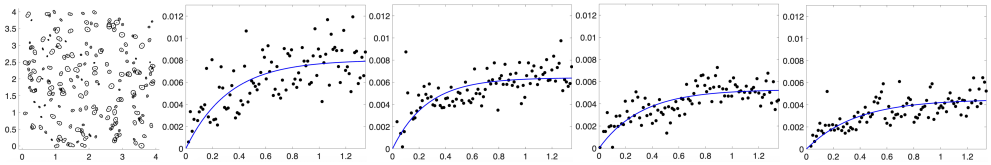
4.1. Comparison of Spatially-weighted Shape Average and Karcher Mean Shape. We set the spatial domain to $X = [0, 2] \times [0, 2]$ and sample a mean Fourier coefficient vector $\boldsymbol{m} = (m_1, \ldots, m_6)$ from $\text{MVN}(0_6^\top, 3I_6)$, where $0_6^\top$ is a six-dimensional vector of zeros and I_6 is the 6×6 identity matrix. We use $\boldsymbol{m}$ to define the radial distance function for the ground truth mean shape, defined at $s_0 = (1, 1)$, and transform to Cartesian coordinates resulting in β_{gt} and its SRVF q_{gt}. We set the sample size to $n = 50$ and generate spatially dependent shapes as described earlier. We consider three settings for the range parameter in the Matérn covariance: 2, 1 and 0.5. We then use the generated data to estimate the Karcher mean shape $\hat{\mu}_q$ (under assumption of independence) using Algorithm 2.1, and the spatially-weighted shape average $\tilde{q}_0$ using Algorithm 3.1 with $s_0 = (1, 1)$. Accuracy of estimation is assessed using the shape distance $d_{\mathcal{S}}([q_{gt}], [\hat{\mu}_q])$ (defined in (2.4)) for the Karcher mean and $d_{\mathcal{S}}([q_{gt}], [\tilde{q}_0])$ for the spatially-weighted average. We repeat this process 100 times.

Table 1 reports the means and standard deviations of the resulting estimation errors. On average, for all values of the range parameter, the spatially-weighted shape average is more similar to the ground truth than the Karcher mean shape; the standard deviations of estimation errors are also smaller for the spatially-weighted average shape. Further, the spatially-weighted shape average becomes more accurate in estimating the underlying ground truth shape when the range parameter in the data generating process decreases. We also repeated this simulation for $n = 100$ and $n = 200$ (see Section 4 in the supplement for results). In all cases, the spatially-weighted shape average was more accurate in estimating the ground truth shape. However, the trend of increasing accuracy as the range parameter decreased is not as clear in those settings. This is likely due to the data generating process, which results

Range	Method	Mean	SD
0.5	Proposed	0.0947	0.0361
	Karcher Mean	0.1118	0.0488
1	Proposed	0.0982	0.0372
	Karcher Mean	0.1099	0.0521
2	Proposed	0.1035	0.0426
	Karcher Mean	0.1075	0.0447

TABLE 1

Means and standard deviations (SDs) of estimation errors for the spatially-weighted average shape and the Karcher mean shape across 100 replicates.



(a) (b) $s_0 = (1,1)$ (c) $s_0 = (1,3)$ (d) $s_0 = (3,1)$ (e) $s_0 = (3,3)$

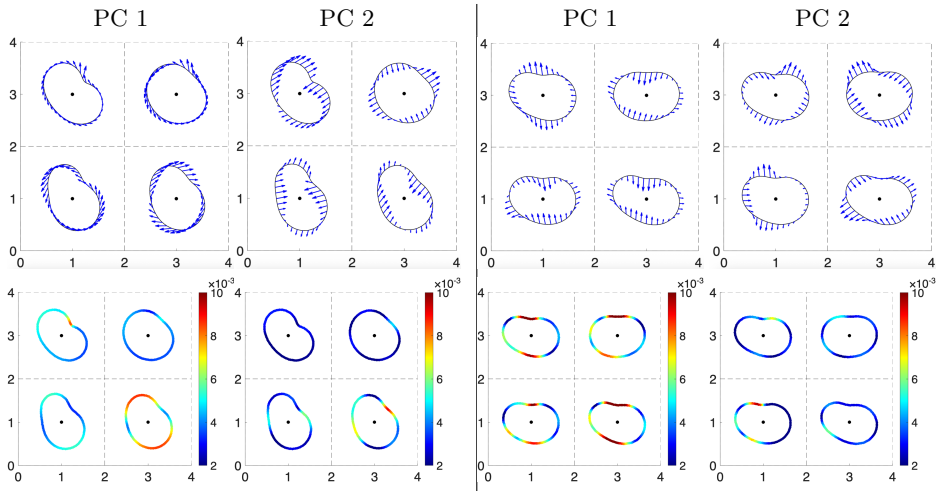
FIG. 3. *(a) Simulated spatially dependent shapes. (b)-(e) Empirical shape trace-variograms (black points) and fitted Matérn models (blue curve) for subregions X_1-X_4, respectively.*

in more shape heterogeneity with increasing n. This simulation illustrates the benefit of incorporating spatial shape dependence through the shape trace-variogram when summarizing spatially correlated shape data.

4.2. Comparison of Independent and Spatially-weighted Shape Variation. In this simulation, we fix the spatial domain to $X = [0,4] \times [0,4]$. To generate $n = 200$ spatially correlated shapes, we use $\boldsymbol{m} = 0_6^\top$ and a range of 2 in the Matérn covariance. To reflect the summarization procedure that will be used for histopathology images, we further partition the spatial domain into four subregions, defined as $X_1 = [0,2] \times [0,2]$, $X_2 = [0,2] \times [2,4]$, $X_3 = [2,4] \times [0,2]$ and $X_4 = [2,4] \times [2,4]$, and perform local summarization based on data within each subregion. For spatially-informed summarization, the location s_0 is chosen as $(1,1)$, $(1,3)$, $(3,1)$ and $(3,3)$ for each subregion, respectively.

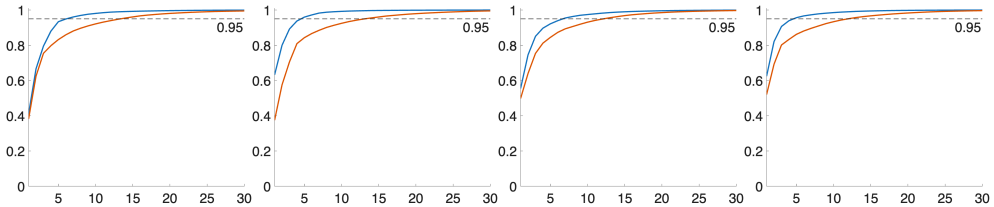
The data is shown in Figure 3(a). In Figure 3(b)-(e), we show the estimated shape trace-variograms within each subregion obtained via Algorithm 3.1. The empirical variogram is shown as black points with the Matérn model fit, with smoothness parameter of 0.5, in blue. Based on the structure of these trace-variograms, it is clear that the shapes within each subregion are spatially dependent, i.e., there is less variation at smaller spatial distance lags; this is especially clear in regions X_1 and X_2 (panels (b) and (c)).

Figure 4 compares local spatially-informed shape summarization (panel (a)), as described in Section 3, to local summarization under independence (panel (b)), as described in Section 2.3. The columns show variation along the first two principal directions of shape variation. First, there is a clear difference in the estimated average shapes across the four subregions. The spatially-weighted shape averages appear to have more geometric features than the Karcher mean shapes, especially in subregions X_1 and X_2. Second, the estimated patterns of variation are markedly different across



(a) Spatially-informed Summarization (b) Summarization Under Independence

FIG. 4. *(a) Spatially-informed summarization of shape variation. (b) Summarization of shape variation under independence. Columns correspond to the first and second principal directions of shape variation. We show variation as a vector field (top row), and its pointwise magnitude (bottom row), on the estimated average.*



(a) (b) (c) (d)

FIG. 5. *Cumulative explained variance as a function of the number of PCs based on spatially-weighted PCA (blue) and unweighted PCA (red) for subregions (a) X_1, (b) X_2, (c) X_3, and (d) X_4.*

the two summarization approaches as evident through the vector fields (top row) and their pointwise magnitudes (bottom row). Finally, PCA carried out using the spatially-weighted covariance captures primary variation more efficiently than PCA carried out on the covariance matrix computed without spatial weighting. In particular, for regions X_1-X_4, 7, 5, 7 and 5 spatially-weighted PCs explain at least 95% of shape variation, respectively. For unweighted PCA, in the same four subregions, 14, 14, 13 and 12 shape PCs are needed to explain at least 95% of the variation. This trend is consistent for other variation thresholds as seen in Figure 5. For all four subregions, the blue cumulative explained variance curve, corresponding to spatially-weighted PCA, is above the red one, which corresponds to unweighted PCA.

5. Application to Histopathology Images. We apply our framework to cell nucleus outlines extracted from whole slide images (WSIs) of breast cancer tissue. The images were obtained from the TiGER challenge [30], which focused on HER2-positive and triple-negative breast cancers; both types are linked to notably poor prognosis. We first selected two images from this database; these are shown in Figure 6(a), where the tumor tissue is outlined in black. For each image, we selected two ROIs

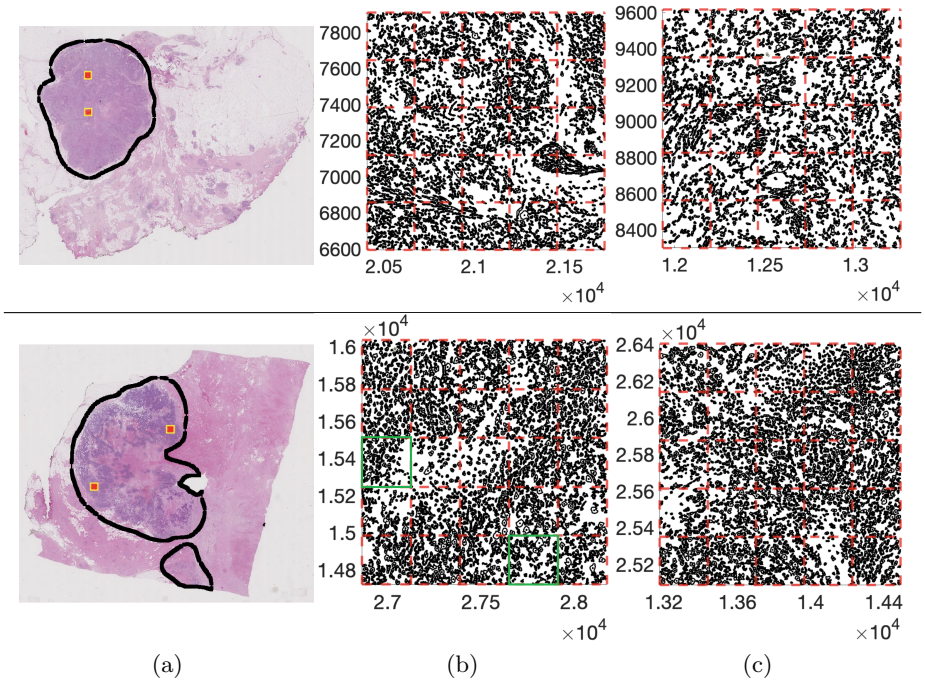
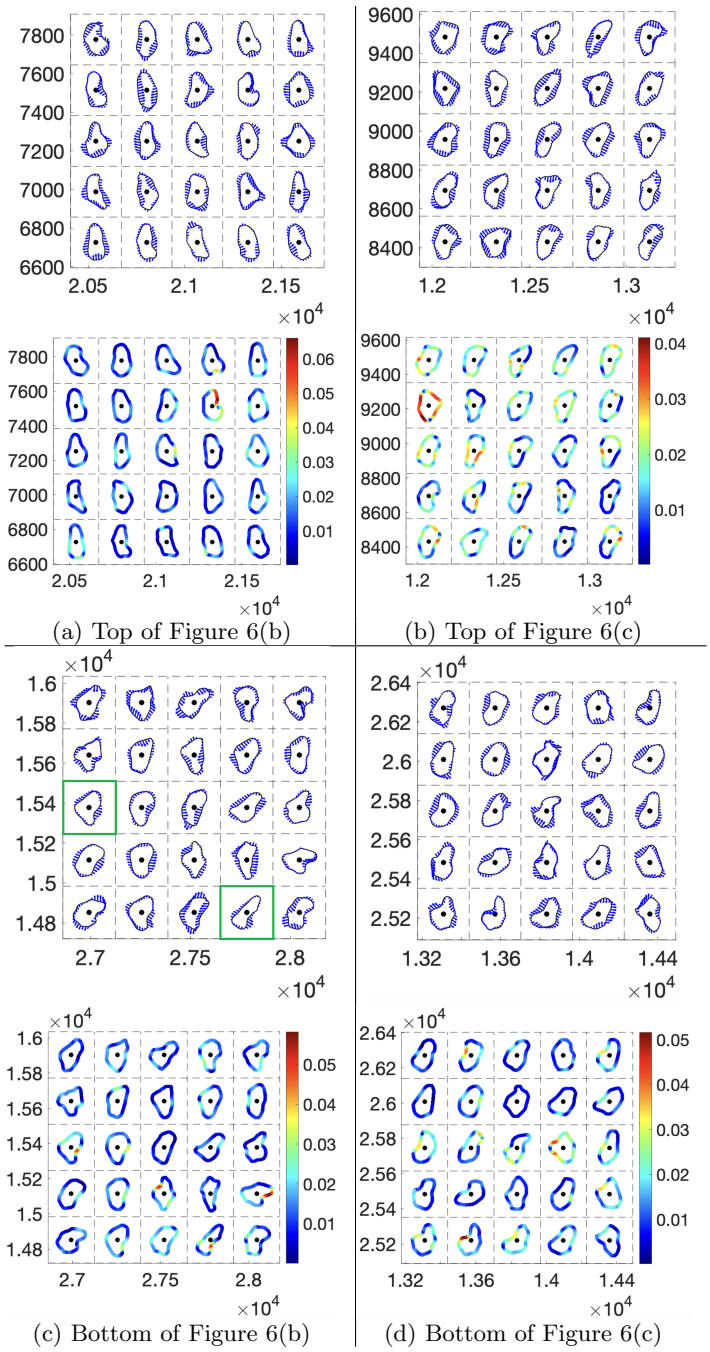


FIG. 6. *(a) Breast cancer tissue images; tumor outlined in black. (b)&(c) Cell nucleus outlines in ROIs (red squares) from (a); subregions are denoted by red dashed lines. The two subregions highlighted in green in the bottom row in (b) are visualized in more detail in Figure 9(a).*

within tumor tissue, shown as red squares in (a). Using QuPath [1], we segmented cell nucleus outlines from these ROIs. The extracted outlines are displayed in (b) and (c) for each of the two images and ROIs. Each ROI was subsequently partitioned into 25 equally sized subregions for local spatially-informed summarization and visualization of cell nucleus shape variation. The two subregions highlighted in green in the bottom row of Figure 6(b) are displayed in more detail in Figure 9(a).

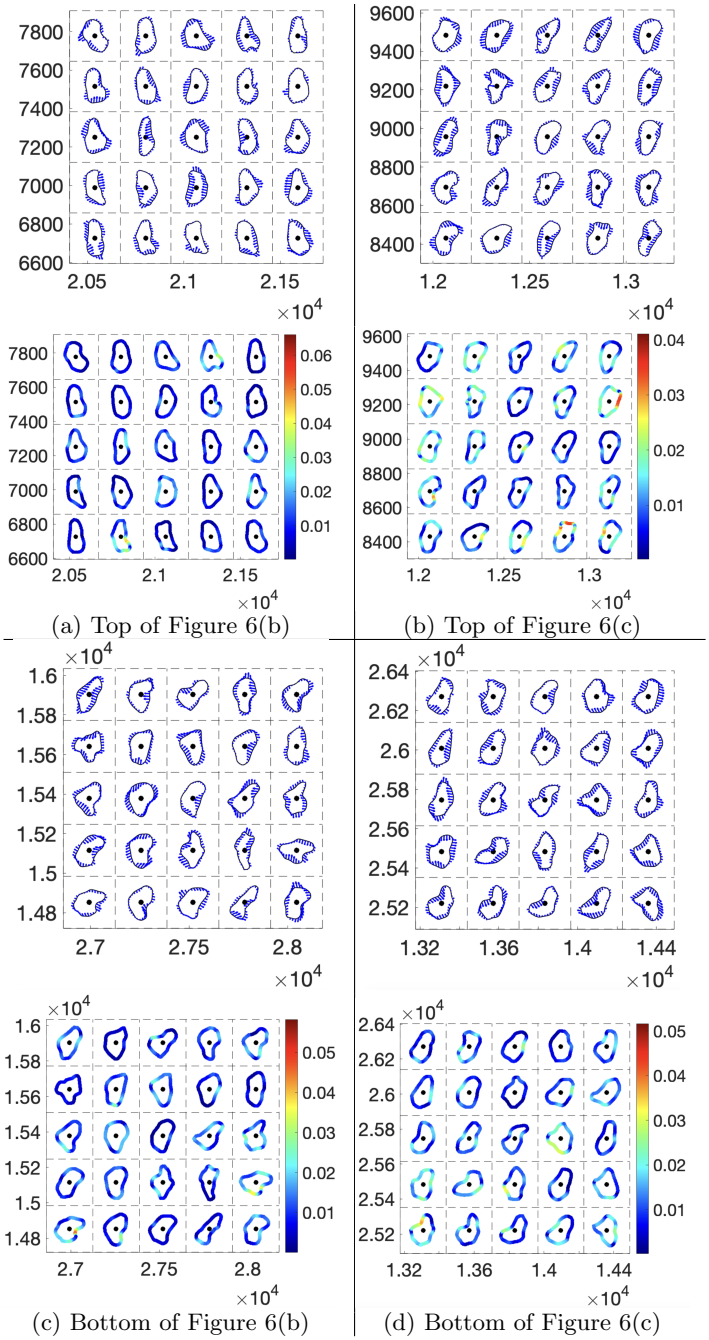
The empirical shape trace-variograms within each subregion exhibited structure that led us to fit a Matérn model with a smoothness parameter of 0.2. Depending on the application and spatial dependence pattern, the smoothness parameter should be tuned to obtain a satisfactory Matérn fit to the empirical shape trace-variogram. However, as we show in Section 4 in the supplement, spatially-informed summarization is not sensitive to the choice of the smoothness parameter, or the density of cell nucleus shapes in a fixed spatial subregion. Figures 7 and 8 show the resulting local spatially-informed summarizations of cell nucleus shape variation. We display the first principal direction of shape variation in Figure 7; the second direction is shown in Figure 8. Panels (a)-(d) consider the ROIs shown in the top row of Figure 6(b)&(c) and bottom row of Figure 6(b)&(c), respectively. The results highlighted in green in the top row of Figure 7(c) are shown in more detail in Figure 9(b). Clearly, the spatially-weighted shape averages can be quite different across subregions within the same ROI, i.e., there appears to be quite a bit of spatial cell nucleus shape heterogeneity. This is also true for the principal directions of shape variation. Certain subregions exhibit much less variation than others. The proposed visualizations represent practical exploratory tools for clinicians to assess cell nucleus shape variation in histopathology images. In



(a) Top of Figure 6(b) (b) Top of Figure 6(c)

(c) Bottom of Figure 6(b) (d) Bottom of Figure 6(c)

FIG. 7. *Spatially-informed summarization of cell nucleus shape variation. Columns correspond to the first principal direction of shape variation for ROIs in top and bottom rows of Figure 6(b)&(c). We show variation as a vector field (top), and its pointwise magnitude (bottom), on the spatially-weighted average shape. The two results highlighted in green in the top row in (c) are visualized in more detail in Figure 9(b).*



(a) Top of Figure 6(b) (b) Top of Figure 6(c)

(c) Bottom of Figure 6(b) (d) Bottom of Figure 6(c)

FIG. 8. *Spatially-informed summarization of cell nucleus shape variation. Columns correspond to the second principal direction of shape variation for ROIs in top and bottom rows of Figure 6(b)&(c). We show variation as a vector field (top), and its pointwise magnitude (bottom), on the spatially-informed average shape.*

particular, the estimated averages and principal directions of variation can be used to identify heterogeneous tissue properties within different ROIs. Importantly, in future

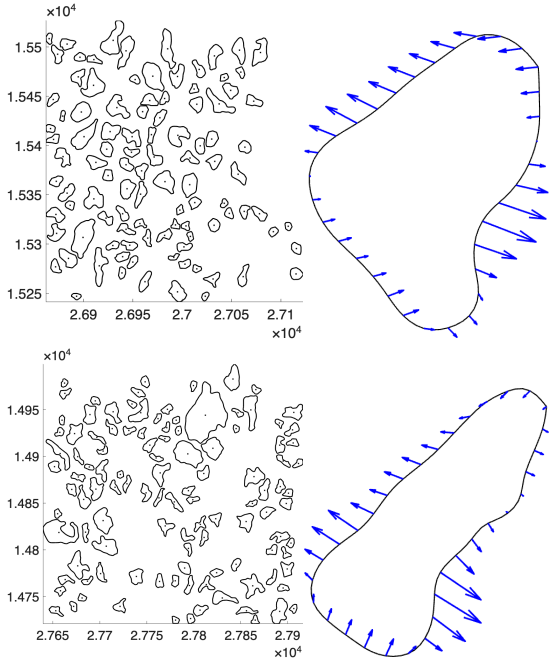


FIG. 9. *(a) Zoomed-in visualizations of the subregions highlighted in green in the bottom row of Figure 6(b). (b) Detailed visualizations of the corresponding spatially-weighted shape averages with the first principal directions of shape variation shown as vector fields.*

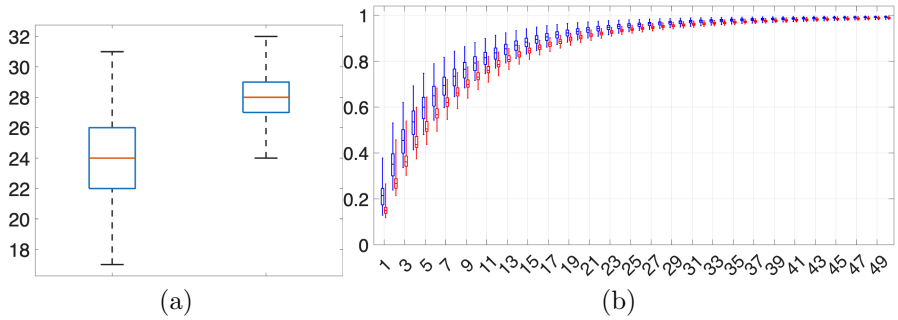


FIG. 10. *(a) Boxplots of the number of PCs required to explain at least 95% of the total shape variance across 100 tumor tissue subregions. Left: Spatially-weighted PCA. Right: Unweighted PCA. (b) Cumulative explained variance as a function of the number of PCs, aggregated over the same 100 tumor tissue subregions: spatially-weighted in blue and unweighted in red.*

studies, such summaries along with further visual aides can be used to characterize subtypes and severity of tumors. Additional results on real data for different ROIs are provided in Section 5 in the supplement.

Comparison of spatially-weighted PCA and unweighted PCA. We next compare efficiency of spatially-weighted PCA and unweighted PCA, in terms of capturing cell nucleus shape variation, based on all 100 subregions (four ROIs shown in Figure 6, each split into 25 subregions). Within each subregion, we carried out spatially-weighted PCA and unweighted PCA. Figure 10(a) presents boxplots of the

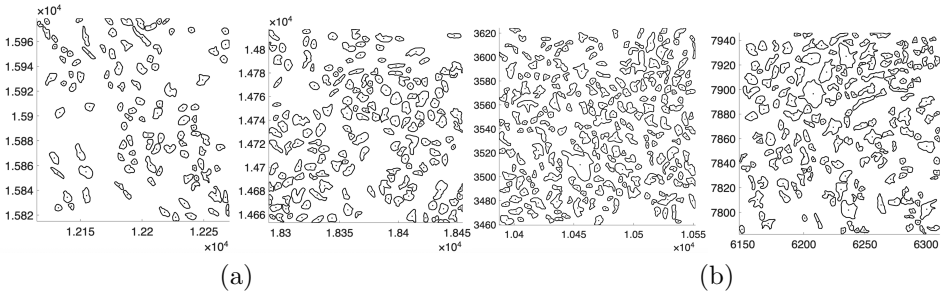


FIG. 11. *Illustration of two randomly sampled regions within (a) normal and (b) tumor tissues.*

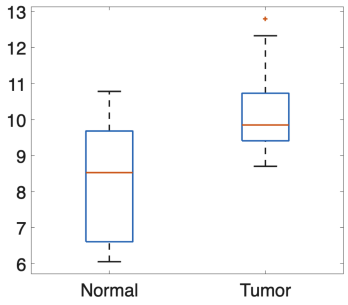


FIG. 12. *Boxplots of spatially-weighted variance of cell nucleus shapes within ten randomly sampled normal and tumor tissue regions of the same size.*

number of PCs required by spatially-weighted PCA (left) and unweighted PCA (right) to capture at least 95% of total shape variation. The proposed method consistently requires fewer PCs to reach the 95% threshold, indicating a more compact representation of shape variability. For 89 of the subregions, spatially-weighted PCA resulted in fewer PCs required to capture at least 95% of the variation; for six of the subregions, both PCA approaches resulted in an equal number of PCs required to capture at least 95% of the variation.

Figure 10(b) further displays cumulative explained variance curves (as boxplots) aggregated across the 100 subregions for both methods (spatially-weighted in blue and unweighted in red). The boxplots corresponding to the proposed approach rise more rapidly as a function of the number of PCs compared to unweighted PCA. This demonstrates that spatial weighting results in a larger proportion of shape variability being captured by the leading PCs. Together, these results show that the spatially-weighted covariance is more efficient in capturing shape variation than the unweighted one, and thus has the potential to provide clearer insight into regional cell nucleus shape heterogeneity.

Comparison of cell nucleus shape heterogeneity within tumor and normal tissue regions. Clinically, it is well known that tumor cells are more variable than healthy cells due to their rapid proliferation and mutation. To confirm this, we use the proposed spatially-informed summarization framework to compare overall shape variation across different regions sampled within tumor and normal tissues. In particular, we randomly sampled ten small regions within tumor tissue and ten size matched regions within normal tissue; see Figure 11 for a visualization of some of the sampled regions. We then computed the total spatially-weighted shape vari-

ance in each of the regions. Figure 12 displays boxplots of the variances. It is clear that spatially-weighted cell nucleus shape variation is larger in tumor tissue than in normal tissue. The average variance within tumor regions is 10.25, whereas the average variance within normal regions is 8.36. This shows that, as expected, cell nuclei exhibit substantially greater morphological heterogeneity within the tumor as compared to normal tissue, which provides quantitative support for increased structural heterogeneity in tumor microenvironments.

6. Discussion and Future Work. We introduced a novel framework for local spatially-informed summarization and visualization of planar curve shapes. By leveraging tools from elastic shape analysis and incorporating spatial dependence via a shape trace-variogram, the method estimates a spatially-weighted shape average and covariance matrix, which can be used to explore spatially-informed shape variation via PCA. Our approach addresses limitations of traditional global summaries by accounting for spatial correlation among shapes, which is especially pertinent in assessing heterogeneity of cell nucleus shapes in histopathology images in the context of cancer. Simulation studies demonstrate that the spatially-weighted shape average is a more accurate estimator of a ground truth shape average than the conventional Karcher mean shape when observed shapes exhibit spatial dependence. Further, spatially-weighted PCA is more efficient in capturing primary modes of shape variation.

We have identified a few directions for future work. First, the natural next step is to generalize the proposed framework to the case of 3D shapes. In many medical imaging applications, the outlines of structures of interest are surfaces rather than curves. Thus, we will define an appropriate analogue of the shape trace-variogram for the case of surfaces; a key aspect of this is to determine how to account for all relevant nuisance transformations. Additionally, we will develop novel visualization techniques for summaries of shape variation in this case. This advance can significantly broaden the framework's applicability, particularly in fields where 3D morphological analysis is critical.

Second, the current approach relies on the assumption of isotropy, i.e., that spatial dependence is uniform in every direction. However, in many practical scenarios, strength of spatial correlation may not be isotropic. We will explore new models to account for directional dependence, potentially leading to better estimates of the spatially-weighted shape average and associated shape variation.

Third, the proposed framework provides point estimates for the spatially-weighted shape average and PCA. In many contexts, uncertainty associated with these estimates is also of interest. While a bootstrap procedure can be utilized to quantify uncertainty in the independent case, it does not apply in the current setting. Thus, we will explore alternative approaches to assess uncertainty when shape data are spatially dependent.

Finally, to translate this framework into clinical practice, we will study the correlation between the spatially-informed shape summaries and clinical outcomes, e.g., tumor grade or survival, to evaluate how well these shape summaries perform as biomarkers in the clinical setting, potentially improving personalized treatment planning and disease monitoring. In particular, breast cancer grading is based on the Nottingham Histopathology Grading system, which includes nuclear pleomorphism as a key component. Nuclear pleomorphism refers to variation in the size, shape and appearance of cell nuclei. As such, the presented methodology can be used to objectively quantify nuclear pleomorphism to aid in tumor grading. Currently, most studies use a potentially small number of shape features to quantify morphological

irregularity of cell nuclei; see, e.g., [29]. However, this can result in significant loss of information, and thus provide misleading results. In contrast, the proposed framework offers spatially-weighted measures of variation estimated from the full geometric shape of cell nuclei.

Authorship and Contributorship Statement. All authors have made substantial intellectual contributions to the study conception, execution, and design of the work. All authors have read and approved the final manuscript. In addition, the following contributions occurred: Conceptualization: Ye Jin Choi, Sebastian Kurtek, Karthik Bharath; Methodology: Ye Jin Choi, Sebastian Kurtek, Karthik Bharath; Formal analysis and investigation: Ye Jin Choi; Writing - original draft preparation: Ye Jin Choi, Sebastian Kurtek, Karthik Bharath; Writing - review and editing: Ye Jin Choi, Sebastian Kurtek, Karthik Bharath; Funding acquisition: Sebastian Kurtek, Karthik Bharath; Supervision: Sebastian Kurtek, Karthik Bharath.

Conflicts of Interest. The authors declare that there are no conflicts of interest.

REFERENCES

[1] P. Bankhead, M. Loughrey, J. Fernández, Y. Dombrowski, D. McArt, et al., *Qupath: Open source software for digital pathology image analysis*, Scientific Reports, 7 (2017), pp. 1–7.

[2] A. Beck, A. Sangoi, S. Leung, R. Marinelli, T. Nielsen, et al., *Systematic analysis of breast cancer morphology uncovers stromal features associated with survival*, Science Translational Medicine, 3 (2011).

[3] F. Bookstein, *Shape and the information in medical images: A decade of the morphometric synthesis*, Computer Vision and Image Understanding, 66 (1997), pp. 97–118.

[4] A. Chakraborty and V. Panaretos, *Functional registration and local variations: Identifiability, rank, and tuning*, Bernoulli, 27 (2021), pp. 1103–1130.

[5] N. Cressie, *Statistics for Spatial Data*, John Wiley & Sons, 2015.

[6] C. de Andrea, A. Petrilli, R. Jesus-Garcia, L. Bleggi-Torres, and M. Alves, *Large and round tumor nuclei in osteosarcoma: Good clinical outcome*, International Journal of Clinical and Experimental Pathology, 4 (2011), p. 169.

[7] E. Fischer, *Nuclear morphology and the biology of cancer cells*, Acta Cytologica, 64 (2020), pp. 511–519.

[8] W. Fridman, F. Pagès, C. Sautès-Fridman, and J. Galon, *The immune contexture in human tumours: Impact on clinical outcome*, Nature Reviews Cancer, 12 (2012), pp. 298–306.

[9] J. Galon, A. Costes, F. Sanchez-Cabo, A. Kirilovsky, B. Mlecnik, et al., *Type, density, and location of immune cells within human colorectal tumors predict clinical outcome*, Science, 313 (2006), pp. 1960–1964.

[10] R. Giraldo, P. Delicado, and J. Mateu, *Ordinary kriging for function-valued spatial data*, Environmental and Ecological Statistics, 18 (2011), pp. 411–426.

[11] X. Guo, S. Kurtek, and K. Bharath, *Variograms for kriging and clustering of spatial functional data with phase variation*, Spatial Statistics, 51 (2022), p. 100687.

[12] M. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, et al., *Histopathological image analysis: A review*, IEEE Reviews in Biomedical Engineering, 2 (2009), pp. 147–171.

[13] D. Hanahan and L. Coussens, *Accessories to the crime: Functions of cells recruited to the tumor microenvironment*, Cancer Cell, 21 (2012), pp. 309–322.

[14] H. Jackson, J. Fischer, V. Zanotelli, H. Ali, R. Mechera, et al., *The single-cell pathology landscape of breast cancer*, Nature, 578 (2020), pp. 615–620.

[15] M.-Y. Ji, L. Yuan, X.-D. Jiang, Z. Zeng, N. Zhan, et al., *Nuclear shape, architecture and orientation features from H&E images are able to predict recurrence in node-negative gastric adenocarcinoma*, Journal of Translational Medicine, 17 (2019), pp. 1–12.

[16] D. Kendall, *Shape manifolds, Procrustean metrics, and complex projective spaces*, Bulletin of the London Mathematical Society, 16 (1984), pp. 81–121.

[17] S. Kurtek, A. Srivastava, and W. Wu, *Signal estimation under random time-warpings and nonlinear signal alignment*, in Proc. of Neural Information Processing Systems, vol. 24, 2011.

[18] S. Kurtek, J. Su, C. Grimm, M. Vaughan, R. Sowell, et al., *Statistical analysis of manual*

segmentations of structures in medical images, Computer Vision and Image Understanding, 117 (2013), pp. 1036 – 1050.

[19] E. Márquez, R. Cabeen, R. Woods, and D. Houle, *The measurement of local variation in shape*, Evolutionary Biology, 39 (2012), pp. 419–439.

[20] W. Mio, A. Srivastava, and S. Joshi, *On shape of plane elastic curves*, International Journal of Computer Vision, 73 (2007), pp. 307–324.

[21] P. Mitteroecker and P. Gunz, *Advances in geometric morphometrics*, Evolutionary Biology, 36 (2009), pp. 235–247.

[22] R. Nafe, K. Franz, W. Schlote, and B. Schneider, *Morphology of tumor cell nuclei is significantly related with survival time of patients with glioblastomas*, Clinical Cancer Research, 11 (2005), pp. 2141–2148.

[23] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, et al., *Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology*, in Proc. of IEEE International Symposium on Biomedical Imaging, 2008, pp. 284–287.

[24] D. Robinson, *Functional Data Analysis and Partial Shape Matching in the Square Root Velocity Framework*, The Florida State University, 2012.

[25] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, et al., *Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images*, Cell Reports, 23 (2018), pp. 181–193.

[26] A. Srivastava and E. Klassen, *Functional and Shape Data Analysis*, Springer Series in Statistics, Springer, 2016.

[27] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn, *Shape analysis of elastic curves in Euclidean spaces*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 33 (2010), pp. 1415–1428.

[28] S. Sternberg, *Biomedical image processing*, Computer, 16 (1983), pp. 22–34.

[29] C. Teoh, X. Tan, K. Ab Rahman, I. Bakrin, K. Goh, J. Siet, and W. Wan Muhamad, *A quantitative measurement method for nuclear-pleomorphism scoring in breast cancer*, Diagnostics, 14 (2024).

[30] Tiger Challenge, *https://tiger.grand-challenge.org/*.

[31] X. Wang, C. Barrera, K. Bera, V. Viswanathan, S. Azarianpour-Esfahani, et al., *Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (TILs) predict clinical benefit for immune checkpoint inhibitors*, Science Advances, 8 (2022), p. eabn3966.

[32] L. Younes, *Computable elastic distances between shapes*, SIAM Journal on Applied Mathematics, 58 (1998), pp. 565–586.

[33] Y. Yuan, H. Failmezger, O. Rueda, H. Ali, S. Gräf, et al., *Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling*, Science Translational Medicine, 4 (2012).