The Annals of Applied Statistics
2021, Vol. 15, No. 2, 619–637
https://doi.org/10.1214/21-AOAS1452

AGGREGATED PAIRWISE CLASSIFICATION OF ELASTIC PLANAR SHAPES

BY MIN HO CHO[*], SEBASTIAN KURTEK[†] AND STEVEN N. MACEACHERN[‡]

Department of Statistics, Ohio State University, [*]cho.829@osu.edu; [†]kurtek.1@stat.osu.edu; [‡]snm@stat.osu.edu

The classification of shapes is of great interest in diverse areas ranging from medical imaging to computer vision and beyond. While many statistical frameworks have been developed for the classification problem, most are strongly tied to early formulations of the problem with an object to be classified described as a vector in a relatively low-dimensional Euclidean space. Statistical shape data have two main properties that suggest a need for a novel approach: (i) shapes are inherently infinite-dimensional with strong dependence among the positions of nearby points, and (ii) shape space is not Euclidean but is fundamentally curved. To accommodate these features of the data, we work with the square-root velocity function of the curves to provide a useful formal description of the shape, pass to tangent spaces of the manifold of shapes at projection points (which effectively separate shapes for pairwise classification in the training data) and use principal components within these tangent spaces to reduce dimensionality. We illustrate the impact of the projection point and choice of subspace on the misclassification rate with a novel method of combining pairwise classifiers.

1. Introduction. Classification of shapes is a fundamental task in many application areas where the primary data object is an image. For example, in medical imaging radiologists and doctors are often interested in classifying patients to disease types based on shapes of anatomical structures. Consequently, the statistical analysis of shape data, and, in particular, the shape classification task is of great interest to the research community. Our focus in this paper is on the multiclass shape classification problem, which presents some unique challenges. To elucidate the main difficulties, we begin by explaining what we mean by "shape data."

The literature on shape analysis has considered many mathematical representations of shape, including finite point sets or landmarks (Dryden and Mardia (1992, 2016), Cootes et al. (1995)), level sets (Malladi, Sethian and Vemuri (1996)), skeletal models (Pizer et al. (2013), Siddiqi and Pizer (2008)) and diffeomorphic transforms or deformable templates (Grenander and Miller (1998), Glaunès et al. (2008)), among others. Stretching the definition of a landmark, consider a set of 2D images of leaves, each marked with a dense set of landmarks. The landmarks provide the outline of a leaf and with them its shape. If the image is shifted, rescaled or rotated, the shape remains unchanged; other transformations change the shape. Kendall (1984) recognized these invariances and defined shape as the geometric information in the set of landmarks that remains when translation, scaling and rotation have been filtered out.

In many applications, such as the leaf example that we return to in Section 4, it seems most natural to study the shape of an object via its entire outline rather than through a finite set of landmarks. In the 2D setting on which we focus, the shape is a planar, closed curve. The functional representation of such a curve replaces the set of landmarks with an alternative description. The curve is parameterized by a starting point on the shape and a mapping from

Received May 2020; revised February 2021.
Key words and phrases. Dimension reduction, LDA, pairwise classification, projection point, QDA, registration.

the unit interval that describes the traversal of the shape, ending the journey at the starting point. Early versions of the functional representation relied on an arc-length parameterization for the traversal (Zahn and Roskies (1972), Klassen et al. (2004)). However, later papers showed that the arc-length parameterization was too rigid (Srivastava et al. (2011), Kurtek et al. (2012), Srivastava and Klassen (2016)) and that statistical analysis of shapes benefits from the more flexible elastic deformations. These elastic parameterizations rely on registration to determine the optimal point-to-point correspondences across objects (we describe this process formally later).

In this work we adopt the popular square-root velocity function (SRVF) representation for elastic shape analysis of planar, closed curves (Joshi et al. (2007), Srivastava et al. (2011)). There are two main advantages associated with this framework: (1) the SRVF simplifies a specific instance of an elastic metric (Mio, Srivastava and Joshi (2007), Younes (1998)) to the simple $\mathbb{L}^2$ metric, facilitating efficient computation, and (2) the SRVF shape space is a quotient space of the unit Hilbert sphere for which many geometric quantities of interest have analytic expressions. These two ingredients allow parameterization-invariant (in addition to the other standard shape preserving transformations) comparisons and statistical models of shape. We exploit this representation for the model-based shape classification task. We provide more mathematical details on the SRVF and the formal definition of the associated shape space in Section 2.1.

1.1. *Motivation*. The complex geometric nature (quotient space of a Hilbert sphere) of the elastic shape space prevents us from directly using standard techniques for classification that are strongly tied to Euclidean geometry. The normal distribution, for example, is defined in $\mathbb{R}^d$, with extension to spaces of infinite dimension provided by the Gaussian process. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), two popular model-based classification techniques which we consider in this work, are described in various fashions but are intrinsically tied to the normal distribution and hence to Euclidean spaces. While alternative classification approaches exist, usually based on shape distances and nearest neighbor-type classification rules (Kurtek et al. (2012), Laga et al. (2012)), we argue that a model-based approach provides more flexibility in the definition of multiclass classification procedures. In particular, the ability to compute likelihoods for shape classes allows us to aggregate multiple pairwise classifiers in a principled manner. Furthermore, multiclass k-nearest neighbors classification often results in many class ties which have to be resolved to reach a final decision. Many tie-breaking rules exist, providing different classification performance, and the choice of the particular tie-breaking rules is often arbitrary.

One standard approach to classification of shapes is based on "linearization" of the shape space (Pal et al. (2017), Srivastava et al. (2011)). This is done by choosing a particular point in the shape space, usually given by the overall sample mean, identifying the (linear) tangent space at this point and projecting the shapes into the tangent space via the inverse-exponential map. Due to the required invariances to rotation and re-parameterization (translation and scale can be removed from the representation space simply by normalization), the inverse-exponential map involves a registration step that solves for an optimal rotation and reparameterization of a shape with respect to the projection point. This makes the choice of this point extremely important for subsequent statistical analysis. Once the shapes are projected into the tangent space, one can apply standard classification techniques. A major drawback of such an approach is that a single tangent space is generally chosen for the pairwise and multiclass problems. In the multiclass case, if one or more populations are very far from the others, projecting all shapes into a single tangent space at the overall mean introduces significant distortion into the tangent space shape coordinates, due to both the nonlinearity of the representation space as well as the misregistration of shapes with respect to the projection point.

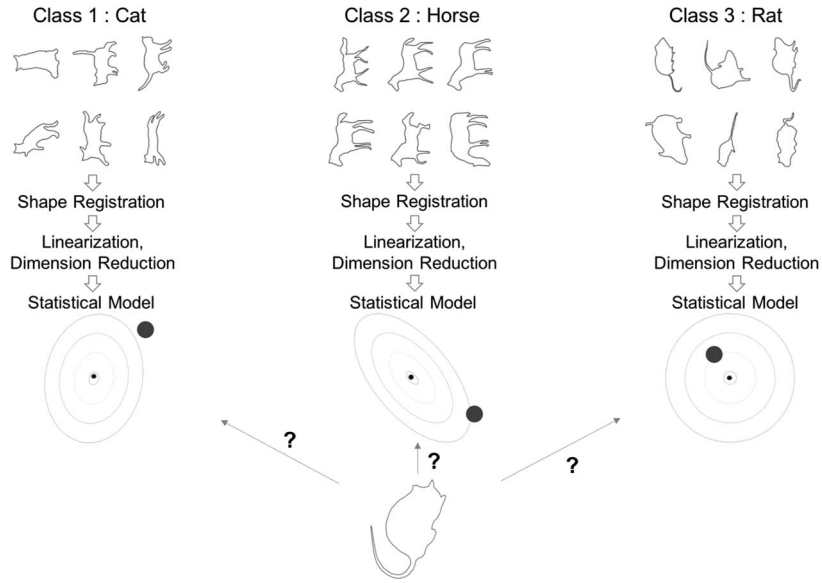

FIG. 1. *A general outline of the multiclass shape classification problem. The shape extraction step requires registration over rotations and reparameterizations, which has a significant effect on classification performance.*

In turn, as seen in later sections, this can have a negative effect on any subsequent statistical task, for example, classification.

A second major challenge in shape classification is the high dimensionality of the shape space. Theoretically, the shape space and corresponding tangent space are infinite-dimensional since we use a functional representation of shape. In practice, the outlines are represented using a fine discretization, typically on the order of hundreds to thousands of points for an individual shape. This discretization leads to a tangent space of large, but finite dimension. The large dimension necessitates modification of LDA and QDA through dimension reduction or regularization. We pursue dimension reduction via a standard form of tangent Principal Component Analysis (tPCA), and modify the LDA and QDA classification procedures accordingly.

Figure 1 summarizes the multiclass shape classification problem. We are given labeled training data in the form of curves. The first task is to extract the shape from these curves. This requires registration, as mentioned earlier, which involves a fairly complex optimization problem. The second task is to linearize the shape space to a Euclidean space, reduce the dimension and then fit a classwise statistical model. The models we consider are based on LDA and QDA. Finally, to classify a new unlabeled object, one has to register its outline to each class and make an assignment based on a classification rule; we consider likelihood-based classification. The cat, horse and rat shapes displayed in Figure 1 are a subset of the animal dataset described in Section 4.

While our focus in this manuscript is on classification of elastic planar shapes represented by their SRVFs, the proposed method can be applied in many other shape analysis settings, for example, shapes of higher dimensional curves, landmark shapes, shapes of surfaces, skeletal shapes, shapes represented by diffeomorphic transformations termed Large Deformation Diffeomorphic Metric Mapping (LDDMM) (Glaunès et al. (2008)), etc. In general, our approach is applicable to multiclass classification problems for high-dimensional, non-Euclidean data. However, after exhaustive simulations we have observed, empirically, that the proposed approach provides the largest gains in classification performance when nonlinear registration is required during local linearization of the shape space, as in the elastic shape analysis framework.

1.2. *Contributions*. Our investigations show the need to modify the standard, single projection point-based linearization approach to classification of shapes. The main reason for this is the arbitrariness in the choice of the tangent space where the full procedure is carried out. This issue is exacerbated by the fact that, prior to classification, all shapes are first registered to this point of projection, which has a significant impact on the chosen shape coordinates in the tangent space and, as a result, classification performance. In other words, since the lower-dimensional Euclidean coordinates of the shapes (after tPCA dimension reduction) are intimately tied to the tangent space used to define them, the chosen point of projection can have a major impact on classification. This is especially true when intra- and/or interclass variability is large, making a single tangent space approximation unsuitable for statistical analysis. In view of this, we list our main contributions:

- For multiclass shape classification we provide a heuristic that suggests different projection points for pairwise problems. We develop a method to combine the pairwise results and compare its performance to a classifier based on a single projection point. The new procedure has substantially better performance than the currently-used single projection methods.
- For aggregating the pairwise problems, we first propose a one-shot method that chooses the class with the highest likelihood (based on the LDA or QDA models). Additionally, we define a recursive approach that drops the class with the lowest likelihood and then recomputes all relevant quantities without this class present in the data. This procedure is especially effective when there are several classes that differ greatly from most groups in the data.
- Finally, we suggest an intermediate method that is also based on aggregation of pairwise problems but only uses a single tangent space. This alternative procedure performs better than the one-shot method in a single tangent space. As expected, it does not perform as well as the new pairwise procedure described above. The main motivation for this intermediate approach comes from the large computational cost of working with all pairwise tangent spaces (this requires the computation of all pairwise means), when the number of classes is large.

The rest of this paper is organized as follows. Section 2 briefly reviews the SRVF framework for elastic shape analysis of planar curves and defines tools for computing all relevant statistics. Section 3 begins by describing the currently used one-shot classification approach in a single tangent space. It then defines the three novel procedures which rely on pairwise statistics and dimension reduction to different degrees. Section 4 includes detailed empirical studies of a popular plant leaf dataset as well as an animal dataset with very diverse shapes. Section 5 provides a short discussion and lays out some directions for future work. Supplementary Material A (Cho, Kurtek and MacEachern (2021a)) includes: (1) a toy one-dimensional simulation study that motivates the use of pairwise procedures in classification, (2) results of classification of data on spheres and the landmark shape space, (3) results of classification using a nonelastic curve framework, (4) results of classification using classwise LDA/QDA models with and without parallel transport, (5) detailed classification results for the recursive method and (6) additional results based on k-nearest neighbors classification.

2. Elastic shape analysis preliminaries. We briefly review the elastic shape analysis framework, based on the square-root velocity function representation, as detailed in Srivastava and Klassen (2016).

2.1. *Square-root velocity function shape space and distance*. Let $\beta : D \to \mathbb{R}^2$ represent a parameterized, planar curve. For an open curve $D = [0, 1]$, while for a closed curve $D = \mathbb{S}^1$.

We restrict our analysis to the set of absolutely continuous curves. The shape of a curve is then invariant to translation, scaling, rotation and reparameterization. Two-dimensional rotation matrices, O, are elements of the special orthogonal group $SO(2) = \{O \in \mathbb{R}^{2\times 2} \mid O^T O = OO^T = I, \det(O) = 1\}$, while reparameterization functions γ are elements of the set of orientation preserving diffeomorphisms of D, denoted by Γ. Comparison of shapes of different objects is fundamental to shape analysis. This task as well as subsequent statistical tasks benefits from the definition of a distance between shapes. The $\mathbb{L}^2$ norm, given by $\|\beta_1 - \beta_2\| = \sqrt{\int_D |\beta_1(t) - \beta_2(t)|^2\, dt}$, where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^2$, seems a natural choice; unfortunately, this distance is not parameterization invariant (Srivastava et al. (2011)). This suggests the need for another distance on shapes.

Mio, Srivastava and Joshi (2007) defined a family of elastic Riemannian metrics that is invariant to all of the aforementioned shape preserving transformations, including reparameterization. These metrics are called elastic because they provide a natural interpretation of shape deformations in terms of their bending and stretching/compression. However, despite these nice mathematical properties, their practical use in shape analysis was limited due to computational difficulties until Joshi et al. (2007) and Srivastava et al. (2011) introduced the square-root velocity function (SRVF)

$$q(t) \equiv \begin{cases} \dot\beta(t)/\sqrt{|\dot\beta(t)|} & \text{if } |\dot\beta(t)| \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\dot\beta$ is the derivative of β with respect to t. The SRVF simplifies the elastic metric to the $\mathbb{L}^2$ metric, thereby facilitating easy computation. If β is absolutely continuous, then its SRVF q is an element of $\mathbb{L}^2(D, \mathbb{R}^2)$, henceforth referred to simply as $\mathbb{L}^2$ (Robinson (2012)). Further, one can uniquely recover the curve β from its SRVF q, up to a translation, via the relation $\beta(t) = \int_0^t q(s)|q(s)|\, ds$, where $t = 0$ is the start point of the parameterization. For the remainder of this paper, we focus on shape analysis of curves facilitated by the SRVF transform.

The translation of a curve is automatically filtered out under the SRVF representation, as it is based on the derivative of the curve. Restricting the set of curves to those that have unit length results in a unit $\mathbb{L}^2$ norm constraint on the associated SRVFs. Thus, we define the preshape space as $\mathbb{S}_\infty = \{q \mid \|q\|^2 = 1\}$, the unit sphere in $\mathbb{L}^2$. Under the $\mathbb{L}^2$ metric the distance between $q_1, q_2 \in \mathbb{S}_\infty$ is given by $d(q_1, q_2) = \cos^{-1}(\langle q_1, q_2\rangle)$, where $\langle q_1, q_2\rangle = \int_D q_1(t)^T q_2(t)\, dt$.

Next, we unify the representation of all SRVFs that are within a rotation and/or reparameterization of each other. Noting that the SRVF of a reparameterized curve, $\beta \circ \gamma$, is given by $(q, \gamma) = (q \circ \gamma)\sqrt{\dot\gamma}$, we define equivalence classes $[q] = \{O(q, \gamma) \mid O \in SO(2), \gamma \in \Gamma\}$ and deem all SRVFs within an equivalence class to have the same shape. The resulting shape space, given by all such equivalence classes, is $\mathcal{S} = \mathbb{S}_\infty/(SO(2) \times \Gamma) = \{[q] \mid q \in \mathbb{S}_\infty\}$. The distance between two shapes is defined as the distance between their equivalence classes as

$$d_{\mathcal{S}}([q_1], [q_2]) = \inf_{O\in SO(2), \gamma\in\Gamma} \cos^{-1}\big(\langle q_1, O(q_2, \gamma)\rangle\big). \tag{2.1}$$

In practice, shape analysis requires a computational implementation. Care with coding, representation of the shape at interior stages of the computation and convergence criteria minimizes numerical error while retaining good speed. For this work we have used standard Procrustes analysis, solved using singular value decomposition (SVD; Section 5.7 in Srivastava and Klassen (2016)) to find optimal rotations, and the "Dynamic Programming" algorithm of Robinson (2012) to find optimal reparameterizations. To implement this approach for closed

curves, they are cut at some point to create an open curve, and the shapes are analyzed as open curves. Our implementation includes an exhaustive search over cut points (also called seeds) on each shape. The optimal rotation and reparameterization (minimizers of equation (2.1)) solve the registration problem. After registration one can construct a geodesic path between two shapes by connecting them via a great circle on the preshape space $\mathbb{S}_\infty$ (see Sections 5.7.2 and 6.8 in Srivastava and Klassen (2016) for several examples).

2.2. *Projection onto tangent space and dimension reduction.* To facilitate classification methods naturally designed for linear spaces, such as LDA or QDA, we use the exponential map and its inverse to linearize the elastic shape space. Since the preshape space is a unit sphere, the mathematical expressions for these mappings are analytic; we omit them here for brevity. Most commonly, the tangent space chosen for statistical shape analysis is defined at the sample Karcher mean shape (Grove and Karcher (1973)), which is defined using the shape distance $d_{\mathcal{S}}$ given in Equation (2.1) (for data $q_1, \ldots, q_n \in \mathbb{S}_\infty$),

$$[\bar{q}] = \arg\min_{[q]\in\mathcal{S}} \sum_{i=1}^{n} d_{\mathcal{S}}([q],[q_i])^2 = \arg\min_{[q]\in\mathcal{S}} \sum_{i=1}^{n} \inf_{O\in SO(2),\gamma\in\Gamma} \left(\cos^{-1}\left(\langle q, O(q_i,\gamma)\rangle\right)\right)^2. \tag{2.2}$$

While this mean is an entire equivalence class, we simply select one element of it for subsequent analysis, that is, we choose some $\bar{q} \in [\bar{q}]$; the specific $\bar{q}$ that is chosen has no impact on the subsequent analysis. The computation of this mean involves an optimization problem which is solved using gradient descent (Kurtek et al. (2013), Dryden and Mardia (2016)).

Given a mean shape, we study variability within and across shape classes using tPCA (Vaillant et al. (2004)). As a first step we project all of the data into the tangent space at the mean using $v_i = \exp_{\bar{q}}^{-1}(q_i^*) \in T_{[\bar{q}]}(\mathcal{S})$, $i = 1, \ldots, n$, where $\exp^{-1}$ denotes the inverse-exponential map (Srivastava et al. (2011)), $q_i^* = O^*(q_i, \gamma^*)$, and O^* and γ^* are the rotation and reparameterization of q_i, respectively, that minimize $d_{\mathcal{S}}([\bar{q}],[q_i])$. In principle, there is a sample covariance for the vectors v_i in the infinite-dimensional tangent space. In practice, the curves are sampled using a finite number of points, say m. Thus, one can simply form the observed tangent data matrix $V \in \mathbb{R}^{n\times 2m}$ (by stacking the x, y coordinates for each v_i into a vector of size $2m$) and then compute the covariance matrix, $Q \in \mathbb{R}^{2m\times 2m}$, using $Q = (1/(n-1))V^T V$. For elastic shape data $n \ll 2m$; LDA and QDA classification methods rely on covariance matrices, which are singular in this setting, necessitating dimension reduction. While one can potentially apply any common statistical technique for dimension reduction to the data matrix V, we use the tPCA approach (Vaillant et al. (2004), Kurtek et al. (2012)). First, we use SVD to compute $Q = U\Sigma U^T$, where U is an orthonormal matrix with columns specifying the principal directions of shape variation, and Σ is a diagonal matrix with nonnegative entries arranged in decreasing order specifying the principal component (PC) variances. Selecting $r < n$, one has a lower-dimensional Euclidean representation of the shapes in the tangent space as $C \in \mathbb{R}^{n\times r}$, with $c_{ij} = v_i U_j$, $i = 1, \ldots, n$, $j = 1, \ldots, r$. These PC data are used for tangent space classification of shapes with LDA or QDA.

3. Classification of shapes on tangent spaces. We describe four procedures for classification of shapes using LDA and QDA in tangent spaces. The first approach is in current use and serves as a baseline. We also discuss practical considerations in the context of data analysis. Throughout this section, we assume that the training data is balanced across classes. The unbalanced case can be handled using reweighting when computing the sample mean shape and by weighting the aggregated likelihoods. All projections are computed using the inverse-exponential map which projects a shape into the tangent space at a projection point along the corresponding geodesic path; in other words, the vector in tangent space from the point of projection to a shape has the same length and direction as the geodesic path from point of projection to shape in shape space.

3.1. *One shot classification on a single tangent space.* In this baseline approach to classification (Pal et al. (2017), Kurtek et al. (2012)), we begin by computing the overall mean shape $\bar{q}$ and a PC coefficient representation of the shapes in the tangent space at $\bar{q}$, using training data pooled over all K classes. Since the linearized shape data often have large dimension compared to the amount of available training data, we use tPCA for dimension reduction arriving at r dimensions. Under the assumption of normality in this r-dimensional space, the log-likelihood of a new observation x under QDA is given by

$$l_{\bar{q}}(x; \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k) = -\frac{1}{2}\log|2\pi\hat{\Sigma}_k| - \frac{1}{2}(x-\hat{\mu}_k)^T\hat{\Sigma}_k^{-1}(x-\hat{\mu}_k), \tag{3.1}$$

where $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are the mean and covariance estimated using training data in the PC coefficient space of class k. For LDA we use the pooled estimate of the $r \times r$ covariance matrix, $\hat{\Sigma}_P = \frac{1}{K}\sum_{k=1}^{K}\hat{\Sigma}_k$, in place of each $\hat{\Sigma}_k$. After computing $l_{\bar{q}}(\mathbf{x}; \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)$ for each class, we choose the class with the largest log-likelihood. In this method we use a single projection point and a single PC space for dimension reduction; for brevity, we refer to this approach as SS. Furthermore, we use a "one-shot" (OS) decision for classification.

3.2. *Aggregated pairwise classification on single tangent space.* In the multiclass case, one can improve upon the SS-OS approach, especially when one class of shapes is very different from the others. The unusual class may be easy to identify, and yet, plays a significant role in determining the lower-dimensional PC space. As a result, the estimated PCs may be ineffective in discriminating between the other classes, leading to a higher than needed misclassification rate. This motivates us to introduce an approach based on PC decompositions of all possible pairwise covariance matrices.

Under this approach, we find $\bar{q}$, the mean shape of the training data pooled over all K classes. The shapes are then projected into the tangent space $T_{[\bar{q}]}(\mathcal{S})$ using the inverse-exponential map. For each pair of classes, $i < j$, $i = 1, \ldots, K$, $j = 1, \ldots, K$, PCs are extracted from the covariance matrix computed using training data in classes i and j only. All training data are then represented in terms of these PCs, and the pairwise log-likelihood of a new observation x for class k (based on the LDA or QDA model), $l_{\bar{q}}^{i,j}(x; \hat{\mu}_k^{i,j}, \hat{\Sigma}_k^{i,j})$, is computed. Here, $\hat{\mu}_k^{i,j}$ and $\hat{\Sigma}_k^{i,j}$ correspond to the mean and covariance of class k, estimated using PC coefficients from the pairwise tPCA model constructed by classes i and j. Thus, there exist $M = \binom{K}{2}$ corresponding log-likelihoods for each class, also indexed by i and j. The M log-likelihoods are aggregated by taking the mean $\bar{l}_{\bar{q}}(x; \hat{\mu}_k^{i,j}, \hat{\Sigma}_k^{i,j}) = M^{-1}\sum_{i<j} l_{\bar{q}}^{i,j}(x; \hat{\boldsymbol{\mu}}_k^{i,j}, \hat{\Sigma}_k^{i,j})$. Use of the geometric mean of the likelihoods (arithmetic mean of the log-likelihoods) has a long history in Bayesian statistics and has been used for, for example, combining expert opinion (Berger (2013)), combining partial Bayes factors (Berger and Pericchi (1996)) and synthesizing different analyses (Yu, MacEachern and Peruggia (2011)). The new observation is then assigned to the class with the maximum mean log-likelihood (OS approach). In this method we use a single tangent space for projection and pairwise PC spaces in this single tangent space for dimension reduction; for brevity we refer to this approach as SP. In later sections we demonstrate that this approach can provide significant improvements in classification over the SS-OS method.

3.3. *Aggregated classification on pairwise tangent spaces.* The projection from the nonlinear shape space to the linear tangent space distorts intershape distances and requires registration that involves the search for optimal rotations and reparameterizations. The amount of distortion depends on multiple factors, including the point of projection and the dispersion of the data. Pursuing the heuristics of pairwise comparisons by projection to the tangent space

at the sample Karcher mean, we consider projections to all pairwise tangent spaces, followed by aggregation.

Under this approach and for each pair of groups in the training data, $i < j$, we compute the pairwise mean shape, $\bar{q}_{i,j}$, and determine the tangent space, $T_{[\bar{q}_{i,j}]}(\mathcal{S})$. We estimate PCs in $T_{[\bar{q}_{i,j}]}(\mathcal{S})$ using the training data in classes i and j only. The data from all classes are projected into each pairwise tangent space, indexed by i and j, and then projected into the PC subspace built using the same pair of classes. Thus, as in the previous section, there are M tPCA-based means and covariances for each class k, again indexed by the pair i and j to denote the two classes used to build the tPCA model. This leads to the log-likelihood of a new observation x (based on the LDA or QDA model), $l_{\bar{q}_{i,j}}(x; \hat{\boldsymbol{\mu}}_k^{i,j}, \hat{\Sigma}_k^{i,j})$, defined in each pairwise tangent space. The multiple log-likelihoods are then combined as before using $\bar{l}_{\bar{q}_{\cdot,\cdot}}(x; \hat{\mu}_k^{i,j}, \hat{\Sigma}_k^{i,j}) = M^{-1} \sum_{i<j} l_{\bar{q}_{i,j}}(x; \hat{\boldsymbol{\mu}}_k^{i,j}, \hat{\Sigma}_k^{i,j})$. Here, we omit the superscript i, j on the log-likelihood l since the projection point $\bar{q}_{i,j}$ already includes this notation. The new observation is then assigned to the class with the maximum average log-likelihood (OS approach). In this case we use pairwise tangent spaces and pairwise PC spaces for dimension reduction; for brevity, we refer to this approach as PP. In many cases this procedure further improves upon the SP-OS method.

An alternative approach would be to build a model for each class on the tangent space at its class mean and then compare the resulting log-likelihoods (Srivastava et al. (2006)). However, each tangent space has its own coordinate system, and so the likelihoods from these different spaces cannot be directly compared. In contrast, our approach builds a model for each class in each tangent space, creating directly comparable log-likelihoods. These well-defined log-likelihoods are then averaged for comparison. Another possibility is to parallel transport representations from each tangent space to a common tangent space, defined at the overall mean (see Section 9.8.2 in Srivastava and Klassen (2016)), such that all models are defined with respect to a common coordinate system. The details of these two procedures, along with the resulting classification performance for all datasets considered in Section 4, are provided in Supplementary Material A (Cho, Kurtek and MacEachern ((2021a), Section 4)).

3.4. *Aggregated pairwise classification with recursion.* The distortion induced by a projection point far from a pair of classes can lead to a very small and numerically unstable contribution to the aggregated log-likelihood. If severe enough, the classification rule can be destabilized. The impact of these poor projection points (and PC spaces) can be limited through use of a recursive procedure. We begin with the calculation of $\bar{l}_{\bar{q}_{\cdot,\cdot}}(x; \hat{\mu}_k^{i,j}, \hat{\Sigma}_k^{i,j})$ for each class k and new observation x. For the recursion the class with the smallest mean log-likelihood is identified and dropped, leading to a similar problem with $K - 1$ classes. The recursion continues with a succession of problems with fewer classes until a single class remains.

As an example, suppose there are K classes. In the first stage there are $M_1 = \binom{K}{2}$ tangent spaces defined at projection points $\{\bar{q}_{i,j}, i < j\}$. The mean log-likelihood for class k is $\bar{l}_{\bar{q}_{\cdot,\cdot}}(x; \hat{\mu}_k^{i,j}, \hat{\Sigma}_k^{i,j}) = M_1^{-1} \sum_{i<j} l_{\bar{q}_{i,j}}(x; \hat{\boldsymbol{\mu}}_k^{i,j}, \hat{\Sigma}_k^{i,j})$. If $\bar{l}_{\bar{q}_{\cdot,\cdot}}(x; \hat{\mu}_{k_1}^{i,j}, \hat{\Sigma}_{k_1}^{i,j})$ (where $k_1 \in \{1, \ldots, K\}$) is the smallest among all mean log-likelihoods, class k_1 is dropped from the comparison. Then, we reaggregate all of the log-likelihoods without projections associated with class k_1. In the second stage there are $M_2 = \binom{K-1}{2}$ projection points that do not include class k_1: $\{\bar{q}_{i,j}, i < j, i, j \neq k_1\}$. The new mean log-likelihood for class k and new observation x is given by $\bar{l}_{\bar{q}_{\cdot,\cdot}}(x; \hat{\mu}_k^{i,j}, \hat{\Sigma}_k^{i,j}) = M_2^{-1} \sum_{i<j, i,j \neq k_1} l_{\bar{q}_{i,j}}(x; \hat{\boldsymbol{\mu}}_k^{i,j}, \hat{\Sigma}_k^{i,j})$. If $\bar{l}_{\bar{q}_{\cdot,\cdot}}(x; \hat{\mu}_{k_2}^{i,j}, \hat{\Sigma}_{k_2}^{i,j})$ is the smallest reaggregated log-likelihood among those of all k, except k_1, class k_2 is dropped from the comparison. Again, we reaggregate the log-likelihoods without projections involving classes k_1 and k_2 and repeat this procedure. After repeating it $K - 1$ times, we obtain a

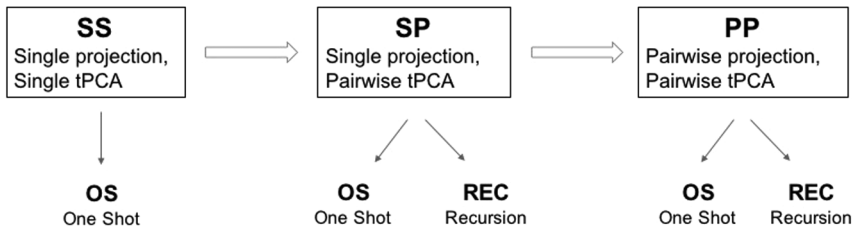

FIG. 2. *A diagram of the proposed shape classification methods.*

unique class for the final classification decision. This recursive approach (REC) can be used instead of the OS method described earlier for SP and PP. In general, REC does not provide the same classification decision as OS.

3.5. *Practical considerations.* In Sections 3.2 and 3.3 we described two methods for multiclass shape classification that use pairwise procedures for dimension reduction to different degrees. Furthermore, in Section 3.4 we proposed a recursive approach for aggregating pairwise classification results. Figure 2 provides a flowchart of all of the approaches, starting with the baseline method SS and progressing to increasingly pairwise procedures (SP and PP). For SP and PP, the user has an additional choice of using the OS or REC decision for classification.

In general, in terms of classification accuracy, PP outperforms SP which outperforms the baseline SS. However, the three methods have different computational costs which must be taken into consideration when choosing the best approach for a given multiclass shape classification problem. The main computational bottleneck is the search for the sample Karcher mean which is performed using an iterative, gradient-based algorithm. The SS and SP approaches require only one such computation, albeit with a potentially large sample size, since they rely on a single tangent space approximation. In contrast to SS, which uses a single PC space for classification, SP estimates all pairwise PC spaces, making it computationally more expensive. However, this increase in computational complexity is minimal. On the other hand, PP requires $\binom{K}{2}$ computations of the sample Karcher mean, making it much more computationally expensive, especially when K is large. Thus, there is a tradeoff between computational cost and classification accuracy. In addition, for SP and PP the final classification decision based on aggregated likelihoods can be made via the OS or REC approaches. The OS method is simple and easy to interpret. But, if the training data contains a mix of very similar and very diverse classes, then the REC procedure helps classification performance by eliminating the worst classes in a stepwise manner. Importantly, when pairwise log-likelihoods are very similar, REC may remove the correct class from the relevant set, which can result in worse performance than the OS method.

Finally, there are two additional choices in these classification procedures: (1) LDA vs. QDA and (2) the dimensionality of the PC space. The first choice has been widely explored in the past for various problems, and we do not discuss it here further. The second choice is nontrivial, and there is no single prescription that applies across problems and datasets. In general, we aim to achieve a low-dimensional, yet faithful, Euclidean representation of shape data via PCs. We have found that the proposed approaches are robust to these two choices. This is confirmed in Section 4 for real-data examples.

4. Empirical studies. We apply the proposed approaches to multiclass classification of two real-shape datasets: plant leaves and animals. For the recursive approach we focus on the first and last stages of the recursion. More detailed classification results that include intermediate stages are provided in Supplementary Material A (Cho, Kurtek and MacEachern

((2021a), Section 5)). Source code for replication is available in Supplementary Material B (Cho, Kurtek and MacEachern (2021b)). First, we consider a problem with a relatively small number of classes by selecting only a few species from the leaf data. Then, we consider the entire datasets of leaves and animals. For these two entire datasets we additionally compare classification performance of the proposed procedures to two nonparametric distance-based methods: (1) k-nearest neighbors and (2) nearest mean. Nearest neighbors classification is based on geodesic shape distances, computed using equation (2.1), between a test case and all training cases. To break ties for k-nearest neighbors when $k > 1$, we reduce the neighborhood size stepwise from k to $k-1, k-2, \ldots, 1$ (if necessary) until there are no ties (Weinberger, Blitzer and Saul (2006)). Additional results of k-nearest neighbors with other ways to break ties are in Supplementary Material A (Cho, Kurtek and MacEachern ((2021a), Section 6)). The nearest mean classification was performed as follows. First, using training data, we estimate a mean shape for each class using equation (2.2). Second, we compute the shape distance from each test case to the shape mean for each class using equation (2.1). Finally, the test case is assigned to the class giving the minimum distance. Supplementary Material A (Cho, Kurtek and MacEachern ((2021a), Section 1)), includes an additional one-dimensional toy simulation that motivates the use of pairwise procedures for classification. We begin with a brief description of the two datasets.

4.1. *Data description*. We first work with the Flavia Plant Leaf dataset[1] (Wu et al. (2007)). The closed outlines used in our work were extracted from images of plants captured using a digital camera. The entire dataset consists of 1907 observations of plant leaves split into 32 classes corresponding to the species of the plants. We note that this dataset does not contain any landmark information in the form of the starting point on each leaf, thus requiring the full search over rotations and reparameterizations (including an exhaustive search over cut points) to compute optimal registrations. In case such landmark information is provided, it can be readily incorporated into the proposed framework via landmark-constrained elastic shape analysis (Strait et al. (2017)). Le and Kume (2000) showed that growth of biological structures (rat skulls in particular) tends to follow geodesic paths in shape spaces. This finding was corroborated by Hotz et al. (2010) for leaves. While we consider a different problem of classification, this suggests that the shape of a leaf provides useful information for differentiating between different leaf species. In Section 4.2 we use a small, carefully selected subset of four leaf species in a semisynthetic simulation experiment to highlight the benefits of the proposed classification approaches. Then, in Section 4.3 we report classification performance on the entire dataset.

Bai, Liu and Tu (2009) provide shape data for animals whose outlines were segmented from natural images. The entire dataset consists of 100 observations for 20 types of animals: bird, butterfly, cat, cow, crocodile, deer, dog, dolphin, duck, elephant, fish, flying bird, chicken, horse, leopard, monkey, rabbit, mouse, spider, tortoise. These 20 animal types will be used in Section 4.4 for classification. In Figure 3(a) we show a single example for each animal class (in the same order as the above list). In Figure 3(b) we show a few examples of cats, monkeys and spiders. We note that there is a lot of variation in pose within each class, making the classification problem very difficult.

4.2. *Semisynthetic simulation*: *Classification of a small subset of the leaf dataset*. We first consider classification of a small subset of leaves from the Flavia Plant Leaf dataset. To highlight the benefits of the proposed methods, we have carefully selected three classes that are difficult to distinguish (classes 1, 3 and 4 drawn in Figure 4(a)) in black, blue and green,

[1] http://flavia.sourceforge.net/

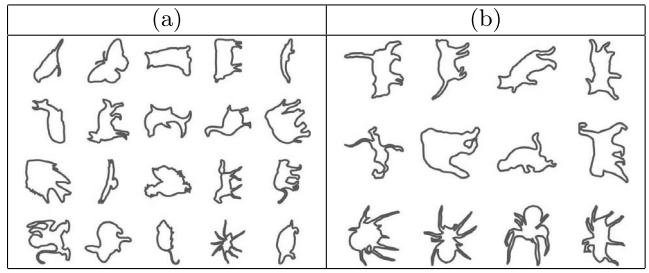

FIG. 3. (*a*) *One sample from each animal class*. (*b*) *Examples of cats, monkeys and spiders*.

respectively) and one outgroup that is easily distinguishable from the other three classes (class 2 drawn in Figure 4(a) in red). To assess classification performance, the dataset is randomly split into a training set of 40 leaves from each class and a test set consisting of the remaining leaves (23 in class 1, 16 in class 2, 20 in class 3 and 16 in class 4). The first two PCs for one training dataset for all classes are plotted in the left panel of Figure 4(b). It is evident that class 2 (red points) is very far from the other classes and is easily distinguishable using any reasonable classification method, including LDA and QDA. However, classification among the other three classes is much more challenging, as there is considerable overlap between the classes. These observations can be easily confirmed using the plots in the right panel of Figure 4(b).

Table 1 illustrates the impact of the projection point on the misclassification rate. We provide results for pairwise classification for classes 1, 3 and 4 only and note that classification involving class 2 is always very good. The table reports total misclassification rates, averaged over 25 random splits of the data, for 12 points of projection. We additionally average the misclassification rates over the numbers of PCs (two through 10) used to define the lower-dimensional space to provide a single performance summary. The candidate projection points are $\bar{q}$ (mean shape of all training samples across the four classes), $\bar{q}_1$, $\bar{q}_2$, $\bar{q}_3$, $\bar{q}_4$ (mean shape of each individual class), $\bar{q}_{1,2}$, $\bar{q}_{1,3}$, $\bar{q}_{1,4}$, $\bar{q}_{2,3}$, $\bar{q}_{2,4}$, $\bar{q}_{3,4}$ (all pairwise mean shapes) and $\bar{q}_{1,3,4}$ (mean shape across classes 1, 3 and 4). The pairwise mean shape projection point results are highlighted in bold, while projection points involving class 2 are underlined. The results indicate that the pairwise mean shape projection points lead to better classification performance. Furthermore, projection points, which include the outgroup in the computation of the mean shape, perform poorly. The magnitude of the impact of the projection point on the misclassification rate is striking and is consistent across the three pairwise problems.

Next, we apply the various procedures described in Section 3 to 25 random splits of the data; the splits were constructed in the same way as described earlier. We consider both LDA

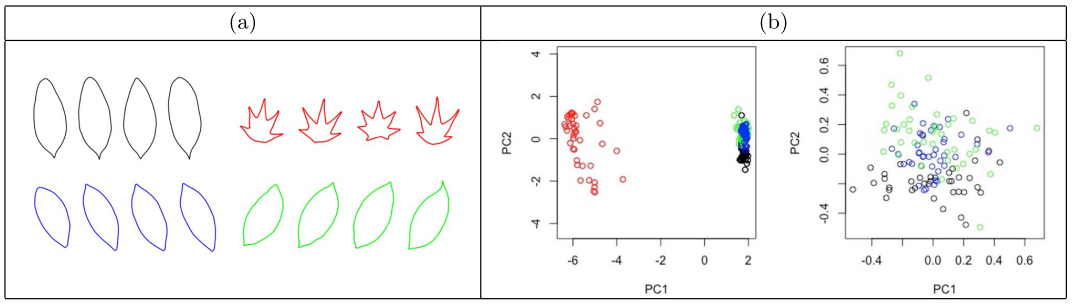

FIG. 4. (*a*) *Sample leaf shapes from four plant species. Classes* 1 (*black*), 3 (*blue*) *and* 4 (*green*) *are similar. Class* 2 (*red*) *is easily discernible*. (*b*) *Plots of the first two PCs in the tangent space at the mean of all four classes* (*left*) *and at the mean of the three similar classes* (*right*). *Each point represents one leaf.*

TABLE 1

Total misclassification rates (%) of LDA and QDA averaged over 25 random splits of the data (four classes of leaves shown in Figure 4). We consider various points of projection for the three pairwise problems for classes 1, 3 and 4. The pairwise mean shape projection point results are highlighted in bold. Projections involving the outgroup are underlined

Class 1 vs. class 3			Class 1 vs. class 4			Class 3 vs. class 4		
Projection	QDA	LDA	Projection	QDA	LDA	Projection	QDA	LDA
$\bar{q}_{1,4}$	1.73	1.70	$\boldsymbol{\bar{q}_{1,4}}$	**5.08**	**4.91**	$\boldsymbol{\bar{q}_{3,4}}$	**4.93**	**4.47**
$\bar{q}_1$	2.26	2.50	$\bar{q}_{1,3,4}$	6.64	6.22	$\bar{q}_4$	5.30	5.05
$\boldsymbol{\bar{q}_{1,3}}$	**2.75**	**3.12**	$\bar{q}_3$	6.66	15.04	$\bar{q}_{1,4}$	5.47	6.63
$\underline{\bar{q}_{1,2}}$	2.97	3.64	$\bar{q}_1$	7.11	6.95	$\bar{q}_1$	6.53	6.90
$\bar{q}_4$	2.99	10.51	$\bar{q}_{1,3}$	7.18	7.72	$\bar{q}_3$	6.88	6.75
$\bar{q}_3$	3.31	21.44	$\bar{q}_{3,4}$	7.23	11.70	$\bar{q}_{1,3,4}$	8.26	8.59
$\bar{q}_{3,4}$	3.73	15.62	$\bar{q}_4$	7.64	13.75	$\bar{q}_{1,3}$	8.83	10.05
$\bar{q}_{1,3,4}$	4.14	4.01	$\underline{\bar{q}}$	7.93	8.15	$\underline{\bar{q}}$	10.00	11.33
$\underline{\bar{q}_{2,4}}$	4.54	15.31	$\underline{\bar{q}_{1,2}}$	8.35	8.68	$\underline{\bar{q}_{2,3}}$	12.09	11.22
$\underline{\bar{q}}$	4.95	6.06	$\underline{\bar{q}_{2,4}}$	10.14	22.46	$\underline{\bar{q}_{2,4}}$	12.41	12.70
$\underline{\bar{q}_{2,3}}$	7.05	17.57	$\underline{\bar{q}_{2,3}}$	10.71	17.12	$\underline{\bar{q}_{1,2}}$	13.11	16.52
$\underline{\bar{q}_2}$	12.42	24.93	$\underline{\bar{q}_2}$	18.27	26.66	$\underline{\bar{q}_2}$	20.16	19.83

and QDA in PC spaces of dimension ranging from two to 20. Table 2 presents the total misclassification rates, averaged over 25 random splits of the data into training and test sets. As a general trend the misclassification rate, when using LDA, decreases as the number of PCs increases. When using QDA, the misclassification rate decreases as the number of PCs increases from two to 10 and then increases slightly. These patterns show the interplay between dimension reduction and the complexity of the model being fit. Overall, QDA performs better than LDA, although LDA performs well for the PP approach with an OS decision for classification (Pairwise Projections, Pairwise PCs and Stage 1). The standard approach SS-

TABLE 2

Total misclassification rates (%) of LDA and QDA for the four leaf species datatset, averaged over 25 random splits of the data into training and test sets. The stages refer to the number of recursion steps as outlined in Section 3.4

	LDA					QDA				
	Overall projection			Pairwise projections		Overall projection			Pairwise projections	
	Overall PC	Pairwise PCs		Pairwise PCs		Overall PC	Pairwise PCs		Pairwise PCs	
PCs	OS	St.1	St.3	St.1	St.3	OS	St.1	St.3	St.1	St.3
2	25.97	18.51	18.88	10.99	11.15	24.16	19.31	18.45	9.92	10.77
4	22.03	13.33	11.73	8.85	6.19	18.88	12.69	10.29	6.24	4.75
6	17.97	9.92	9.49	6.72	4.80	14.67	9.01	8.05	5.28	4.00
8	16.21	8.85	8.16	5.65	4.27	11.25	7.31	5.92	4.43	4.00
10	13.33	8.11	7.31	5.23	4.21	10.51	7.04	5.71	3.89	3.09
12	11.15	7.25	6.88	4.85	4.48	8.69	6.93	5.07	3.68	3.63
14	9.17	7.31	6.56	4.69	4.43	9.49	7.57	5.55	4.53	3.57
16	8.00	7.52	6.40	4.53	4.21	9.76	7.57	6.13	4.32	3.84
18	6.99	6.67	6.13	4.43	4.37	10.08	8.48	6.88	4.64	4.75
20	6.67	6.72	6.51	4.43	4.48	10.24	9.23	7.84	5.44	6.03

OS (Overall Projection, Overall PC, One Shot) provides the poorest performance among the methods.

Table 2 allows us to assess the value of various components of the procedures we have described: choice of projection point (single or pairwise), choice of PC space (single or pairwise) and choice of decision for classification (one shot or recursive). The table shows that the use of pairwise projection points is beneficial in classification based on both LDA and QDA. The comparison of the One Shot columns to the Stage 1 columns shows the value of selecting pairwise PC spaces rather than performing classification in a single PC space based on all training data. Finally, the value of recursion over the one-shot approach is demonstrated by the reduction in misclassification rate from Stage 1 to Stage 3. Supplementary Material A (Cho, Kurtek and MacEachern ((2021a), Section 3)) reports the classification performance of the proposed methods when a nonelastic shape analysis approach (SRVF representation under arc-length parameterization) is used to define the LDA and QDA models. The results reported there are directly comparable to the ones shown in Table 2, and it is clear that elastic shape analysis provides superior classification performance.

4.3. *Real data example*: *Classification of the entire leaf dataset*. We now apply all of the proposed procedures to the entire leaf dataset. Since the 32 species of leaves have various shapes, that is, some classes are very similar while others are very different, we are again interested in investigating the differences in misclassification rate across the various modeling and final decision choices provided by the proposed methods. As in Section 4.2, we use 25 random splits of the data into training and test sets. We use 40 training samples from each class and the remaining samples for testing. This results in a balanced training set but an unbalanced test set. The total number of test cases across all classes is 627.

Table 3 shows the total misclassification rates for LDA and QDA, again averaged over the 25 random splits of the data into training and test sets. In general, the misclassification rates are larger than those in Table 2, as expected. However, the proposed methods perform quite well, with PP-OS in an eight-dimensional PC space providing the lowest misclassification

TABLE 3

Total misclassification rates (%) of LDA and QDA, for the entire leaf dataset of 32 species, averaged over 25 random splits of the data into training and test sets. The stages refer to the number of recursion steps as outlined in Section 3.4

	LDA					QDA				
	Overall projection			Pairwise projections		Overall projection			Pairwise projections	
	Overall PC	Pairwise PCs		Pairwise PCs		Overall PC	Pairwise PCs		Pairwise PCs	
PCs	OS	St.1	St.31	St.1	St.31	OS	St.1	St.31	St.1	St.31
2	54.50	33.60	33.72	24.04	23.30	45.69	24.73	30.33	18.09	19.50
4	41.72	31.20	27.60	21.79	19.42	33.35	18.01	22.02	13.45	14.36
6	32.54	29.14	25.93	20.26	17.91	24.38	15.87	19.27	10.51	12.57
8	30.39	27.73	25.06	18.62	17.23	22.41	14.87	17.42	9.73	11.88
10	29.22	26.67	24.58	17.71	16.96	21.38	15.05	17.58	9.95	12.01
12	27.30	26.02	24.52	17.23	16.87	20.34	15.62	17.86	10.53	12.53
14	26.18	25.42	24.54	16.89	16.82	20.85	16.37	18.71	11.32	13.29
16	25.37	24.91	24.43	16.56	16.82	22.21	17.70	19.37	12.27	14.35
18	24.78	24.71	24.41	16.33	16.81	23.77	19.18	20.21	13.22	15.41
20	24.57	24.38	24.36	16.08	16.82	26.08	20.89	21.75	14.53	16.81

TABLE 4

Total misclassification rate (%) of SS, SP-OS, PP-OS for LDA and QDA (number of PCs chosen to minimize error), and $k = 1, 3, 5, 7$-nearest neighbors and nearest mean classifiers for the entire leaf dataset of 32 species, averaged over the identical 25 splits of the data into training and test sets in Table 3

SS		SP-OS		PP-OS		Nearest neighbor				Nearest mean
LDA	QDA	LDA	QDA	LDA	QDA	$k=1$	$k=3$	$k=5$	$k=7$	
24.57	22.41	24.38	14.87	16.08	**9.73**	11.78	11.29	11.87	12.18	20.29

rate of only 9.73% based on the QDA model. We note that methods that use pairwise projection points and pairwise PC spaces perform better than their single projection counterparts. The LDA misclassification rates decrease as the dimensionality of the PC spaces increases. For QDA the misclassification rates decrease up to a point and then increase slightly. These trends are the same as those observed in the previous section. Overall, QDA performs better than LDA across all methods. We note that the full recursion does not always perform better than the one-shot method in this case. For LDA it reduces the misclassification rate up to around a 14-dimensional PC space. For QDA the one-shot method always performs better. A careful examination of intermediate stages (see Supplementary Material A (Cho, Kurtek and MacEachern ((2021a), Section 5)) suggests that the recursion may help up to a stage where all outgroups are removed from the data. After that, the log-likelihoods exhibit greater numerical stability and averaging across linearizations provides a modest benefit.

Table 4 compares the misclassification rates of PP-OS for LDA and QDA, and multiple classifiers based on shape distance: the $k = 1, 3, 5, 7$-nearest neighbors and nearest mean classifiers. Even though PP-OS for LDA with 20 PCs shows larger misclassification rate than all nearest neighbor methods, PP-OS for QDA with eight PCs has better performance. Both PP-OS methods significantly outperform the nearest mean approach and the single projection methods SS and SP-OS.

Upon closer examination there are three sets of classes of leaves that are most difficult to distinguish via the proposed model-based shape classification approaches. They correspond to plant species: (1) Anhui Barberry, Oleander and Ford Woodlotus, (2) Wintersweet and Camphortree and (3) Japan Arrowwood and Sweet Osmanthus. A sample image of a leaf from each class is presented in Supplementary Material A (Cho, Kurtek and MacEachern (2021a)) (Figures 5, 6 and 7 correspond to (1), (2) and (3), respectively). It is evident that, within each set, the different classes of leaves tend to have very similar outer shapes. However, they clearly differ in color and the internal branching structure of the veins. In fact, when biologists manually classify plant leaves, they use all of the visual information available in the image and are thus better able to distinguish between the leaf species. This suggests that a model-based approach for classification that is able to integrate additional features of the imaged leaves (beyond the shape of their outlines) will lead to further improvements in performance and would more closely imitate the process used by biologists in this task.

4.4. *Real-data example*: *Classification of the animal dataset*. This last set of results considers a dataset of 20 animals observed under very different poses, which presents some additional challenges for shape classification. While the shapes of leaves within the same class were very similar, within class variability for the animal shapes is much larger, due to the poses in which the animals were imaged. We use 25 random splits of the data into training and test sets of sizes 60 and 40, respectively. Thus, the total number of test cases across all classes is 800.

Table 5 shows the total misclassification rates for this example, averaged over the 25 splits of the data into training and test sets. The overall patterns are very similar to the leaf dataset

TABLE 5

Total misclassification rates (%) of LDA and QDA, for the animal dataset, averaged over 25 random splits of the data into training and test sets. The stages refer to the number of recursion steps as outlined in Section 3.4

	LDA					QDA				
	Overall projection			Pairwise projections		Overall projection			Pairwise projections	
	Overall PC	Pairwise PCs		Pairwise PCs		Overall PC	Pairwise PCs		Pairwise PCs	
PCs	OS	St.1	St.19	St.1	St.19	OS	St.1	St.19	St.1	St.19
12	62.99	61.52	59.54	45.67	53.49	54.58	48.77	49.05	34.40	41.60
14	61.17	60.59	58.90	44.50	52.75	52.95	47.68	48.07	33.28	41.14
16	60.51	59.85	58.20	43.41	51.87	51.67	47.10	47.45	32.68	40.48
18	59.85	59.22	57.47	42.24	51.14	50.64	46.53	46.38	32.39	40.21
20	59.08	58.40	56.58	41.37	50.08	50.49	46.21	46.13	32.31	40.49
22	58.17	57.69	56.29	40.55	49.35	50.58	46.06	46.27	32.40	40.84
24	58.14	57.07	55.95	39.82	48.97	50.63	46.47	46.41	32.84	41.58
26	57.49	56.54	55.69	39.16	48.63	50.85	46.87	46.82	33.53	42.32
28	56.96	56.15	55.43	38.61	48.37	51.40	47.34	47.38	34.48	43.74
30	56.47	55.83	55.28	37.95	47.98	51.85	48.12	47.88	35.57	44.82

example. However, the misclassification rates in this case are much worse. Overall, QDA performs better than LDA, and the PP procedure leads to smaller misclassification rates than the SP method. The baseline SS method is the worst, as before. The recursion helps only for the SP method and hurts for the PP methods. We also note that many more PCs are required in this example to achieve good classification performance, a result of larger variability within and across classes.

Table 6 compares the misclassification rates of PP-OS for LDA and QDA, and the same distance-based classifiers that were used for the entire leaf dataset. Here, both PP-OS methods for LDA with 30 PCs and QDA with 20 PCs show larger misclassification rates than the nearest neighbors methods. However, the performance of the nearest neighbors methods deteriorates as the number of nearest neighbors considered increases. When seven nearest neighbors are used, the performance of our QDA-based classifier is comparable to the nearest neighbors approach. This perhaps suggests that a local classifier, akin to a few nearest neighbors, is better suited for this dataset due to large heterogeneity within the shape classes. Both PP-OS procedures significantly outperform the global nearest mean method and the single projection methods SS and SP-OS.

This example is very different from the leaf example, which presents a much more controlled environment: the leaf images were all acquired in a fixed position. Thus, the animal dataset poses a more difficult classification problem. The two examples present complementary assessments of the performance of the proposed classification methods.

TABLE 6

Total misclassification rate (%) of SS, SP-OS, PP-OS for LDA and QDA (number of PCs chosen to minimize error), and $k = 1, 3, 5, 7$-nearest neighbors and nearest mean classifiers for the animal dataset of 20 species, averaged over the identical 25 splits of the data into training and test sets in Table 5

SS		SP-OS		PP-OS		Nearest neighbor				Nearest mean
LDA	QDA	LDA	QDA	LDA	QDA	$k = 1$	$k = 3$	$k = 5$	$k = 7$	
56.47	50.49	55.83	46.21	37.95	32.31	25.35	**25.01**	25.80	26.44	52.92

5. Discussion. An important step in elastic shape analysis is the move from the infinite dimensional, curved space, where shapes naturally abide, to a finite dimensional linear space that allows the use of a suite of standard statistical tools. In this article we have shown that this linearization is not trivial and that the details of the linearization can have a major impact on subsequent statistical inference. The linearization consists of two main components: a projection point to determine a tangent space and dimension reduction by choice of PCs. A key step in the linearization process is the registration of all shapes to the projection point. This makes the choice of projection a key step for the success of subsequent analyses.

We propose aggregation as a mechanism to make use of multiple linearizations driven by different projection points and PCs. Aggregation allows us to focus on the pairwise classification problem where the existing literature provides a sound heuristic for the linearization. By itself, the use of pairwise PCs followed by aggregation of likelihoods from statistical models provides a substantial benefit with more flexibility, even compared to alternative nonparametric classification approaches.

Additionally, we note that the presence of an outgroup can harm the aggregation by contributing linearizations that have little relevance for the classes to which an observation might plausibly be assigned. While one could use alternative procedures to account for outgroups, for example, based on voting principles, we propose a recursion that can be quickly computed from the results of the aggregation. The recursion has proven successful for a problem with a modest number of classes and a clear outgroup. For problems with a profusion of classes and great within-class variation, the recursion harms performance.

The nearest neighbor approach performs much better than the single projection, pairwise tPCA methods as well as the pairwise projection, pairwise tPCA LDA approach, especially for the animal dataset. The nearest neighbors method is very local, particularly when the classification decision relies on only a few neighbors, vs. the two proposed model-based methods, which are more global. However, the performance of the nearest neighbor approach deteriorates as more neighbors are considered in the classification decision. The locality of the classification procedure seems to be much more important in the case of the animal dataset than the leaf dataset, with the one nearest neighbor approach providing the best performance. This matches intuition based on what we know about this dataset: the animal shapes were extracted from natural images under many different poses. More global methods that either consider more neighbors or define models that consider all of the poses would naturally be expected to perform worse in this case.

In terms of computational cost, the most time consuming portion of our approach is registration, that is, the search for cut-point, optimal rotation and reparameterization for each shape and target pair. Assume there are N training samples for each of K classes. For nearest neighbors classifiers a test sample is registered to all NK training samples; for the pairwise projection methods a test sample is registered to $\binom{K}{2}$ class means; for the single projection or nearest mean procedures, a test sample is registered to K means. The proposed PP approach rivals the nearest neighbors methods in terms of accuracy and is, in many settings, much quicker. It is slower than the single projection and nearest mean methods, but it has much better classification accuracy. We recommend the PP approach when good classification performance is desired and N is large relative to K.

The proposed approaches are applicable in classification problems on general Riemannian manifolds. However, we have observed empirically that the largest gains in classification performance over competing methods are observed when linearization of the space involves nonlinear registration, such as in elastic shape analysis. Supplementary Material A (Cho, Kurtek and MacEachern ((2021a), Section 2)) reports results of additional classification experiments on high-dimensional spheres and the landmark shape space.

This work opens up several methodological and applied problems, several of which we are in the process of examining. One is the choice of projection points for pairwise comparisons.

A sensible alternative to the pairwise mean is the midpoint of the geodesic between the two classwise mean shapes. Although not exactly the same, these two types of projection points are typically close to one another. This will allow us to greatly reduce computational cost: we only have to compute K mean shapes instead of $\binom{K}{2}$. A second is whether there are more effective ways to select the PCs for a given projection and pairwise classification problem. A third is to delineate the circumstances under which the recursion is beneficial. A fourth is whether the recursion can be modified to retain aggregation over a subset of linearizations. Finally, we plan to explore related applied problems in biology. In particular, we will explore relationships between shapes of plants and/or animals and various covariates of interest, such as genomic signatures, environmental variables, etc. In future work we expect to consider these and other problems.

Acknowledgments. We thank Dr. Hamid Laga from Murdoch University and Dr. Xiang Bai from Huazhong University of Science and Technology for providing the outlines from the Flavia Plant Leaf and the animal datasets, respectively. We also thank the Editor, Associate Editor and two reviewers for their thoughtful comments, which significantly improved this manuscript.

Funding. This research was partially supported by NSF DMS 1613110 (to SM), and NSF DMS 1613054, NSF CCF 1740761, NSF CCF 1839252 and NIH R37 CA214955 (to SK).

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to "Aggregated pairwise classification of elastic planar shapes" (DOI: 10.1214/21-AOAS1452SUPPA; .pdf). This supplement includes (1) a toy one-dimensional simulation study that motivates the use of pairwise procedures in classification, (2) results of classification of data on spheres and the landmark shape space, (3) results of classification using a nonelastic curve framework, (4) results of classification using classwise LDA/QDA models with and without parallel transport, (5) detailed classification results for the recursive method, (6) additional results based on k-nearest neighbors classification, and (7) additional figures for the leaf classification example in Section 4.3.

Supplement B: Source code for "Aggregated pairwise classification of elastic planar shapes" (DOI: 10.1214/21-AOAS1452SUPPB; .zip). Matlab and R source code for the models described in this paper and data files.

REFERENCES

BAI, X., LIU, W. and TU, Z. (2009). Integrating contour and skeleton for shape classification. In *IEEE International Conference on Computer Vision Workshops* 360–367.

BERGER, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*, *Springer Series in Statistics*. Springer, New York. MR0804611 https://doi.org/10.1007/978-1-4757-4286-2

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109–122. MR1394065 https://doi.org/10.2307/2291387

CHO, M. H., KURTEK, S. and MACEACHERN, S. N. (2021a). Supplement to "Aggregated pairwise classification of elastic planar shapes." https://doi.org/10.1214/21-AOAS1452SUPPA

CHO, M. H., KURTEK, S. and MACEACHERN, S. N. (2021b). Source code for "Aggregated pairwise classification of elastic planar shapes." https://doi.org/10.1214/21-AOAS1452SUPPB

COOTES, T. F., TAYLOR, C. J., COOPER, D. H. and GRAHAM, J. (1995). Active shape models—Their training and application. *Comput. Vis. Image Underst.* **61** 38–59.

DRYDEN, I. L. and MARDIA, K. V. (1992). Size and shape analysis of landmark data. *Biometrika* **79** 57–68. MR1158517 https://doi.org/10.1093/biomet/79.1.57

DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis with Applications in R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3559734 https://doi.org/10.1002/9781119072492

GLAUNÈS, J., QIU, A., MILLER, M. I. and YOUNES, L. (2008). Large deformation diffeomorphic metric curve mapping. *Int. J. Comput. Vis.* **80** 317–336.

GRENANDER, U. and MILLER, M. I. (1998). Computational anatomy: An emerging discipline. *Quart. Appl. Math.* **56** 617–694. MR1668732 https://doi.org/10.1090/qam/1668732

GROVE, K. and KARCHER, H. (1973). How to conjugate C^1-close group actions. *Math. Z.* **132** 11–20. MR0356104 https://doi.org/10.1007/BF01214029

HOTZ, T., HUCKEMANN, S., MUNK, A., GAFFREY, D. and SLOBODA, B. (2010). Shape spaces for prealigned star-shaped objects—Studying the growth of plants by principal components analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 127–143. MR2750135 https://doi.org/10.1111/j.1467-9876.2009.00683.x

JOSHI, S. H., KLASSEN, E., SRIVASTAVA, A. and JERMYN, I. H. (2007). A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^n$. In *IEEE Conference on Computer Vision and Pattern Recognition* 1–7.

KENDALL, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* **16** 81–121. MR0737237 https://doi.org/10.1112/blms/16.2.81

KLASSEN, E., SRIVASTAVA, A., MIO, W. and JOSHI, S. H. (2004). Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 372–383.

KURTEK, S., SRIVASTAVA, A., KLASSEN, E. and DING, Z. (2012). Statistical modeling of curves using shapes and related features. *J. Amer. Statist. Assoc.* **107** 1152–1165. MR3010902 https://doi.org/10.1080/01621459.2012.699770

KURTEK, S., SU, J., GRIMM, C., VAUGHAN, M., SOWELL, R. and SRIVASTAVA, A. (2013). Statistical analysis of manual segmentations of structures in medical images. *Comput. Vis. Image Underst.* **117** 1036–1050.

LAGA, H., KURTEK, S., SRIVASTAVA, A., GOLZARIAN, M. and MIKLAVCIC, S. J. (2012). A Riemannian elastic metric for shape-based plant leaf classification. In *International Conference on Digital Image Computing Techniques and Applications* 1–7.

LE, H. and KUME, A. (2000). Detection of shape changes in biological features. *J. Microsc.* **200** 140–147.

MALLADI, R., SETHIAN, J. A. and VEMURI, B. C. (1996). A fast level set based algorithm for topology-independent shape modeling. *J. Math. Imaging Vision* **6** 269–289. MR1390215 https://doi.org/10.1007/BF00119843

MIO, W., SRIVASTAVA, A. and JOSHI, S. H. (2007). On shape of plane elastic curves. *Int. J. Comput. Vis.* **73** 307–324.

PAL, S., WOODS, R. P., PANJIYAR, S., SOWELL, E. R., NARR, K. L. and JOSHI, S. H. (2017). A Riemannian framework for linear and quadratic discriminant analysis on the tangent space of shapes. In *Workshop on Differential Geometry in Computer Vision and Machine Learning* 726–734.

PIZER, S. M., JUNG, S., GOSWAMI, D., VICORY, J., ZHAO, X., CHAUDHURI, R., DAMON, J. N., HUCKEMANN, S. and MARRON, J. S. (2013). Nested sphere statistics of skeletal models. In *Innovations for Shape Analysis*: *Models and Algorithms*. *Math. Vis.* 93–115. Springer, Heidelberg. MR3075829 https://doi.org/10.1007/978-3-642-34141-0_5

ROBINSON, D. T. (2012). Functional data analysis and partial shape matching in the square root velocity framework. Ph.D. thesis, The Florida State Univ. MR3152393

SIDDIQI, K. and PIZER, S. M., eds. (2008). *Medial Representations*: *Mathematics*, *Algorithms and Applications*. *Computational Imaging and Vision* **37**. Springer, New York. MR2547467 https://doi.org/10.1007/978-1-4020-8658-8

SRIVASTAVA, A. and KLASSEN, E. P. (2016). *Functional and Shape Data Analysis*. *Springer Series in Statistics*. Springer, New York. MR3821566

SRIVASTAVA, A., JAIN, A., JOSHI, S. and KAZISKA, D. (2006). Statistical shape models using elastic-string representations. In *Asian Conference on Computer Vision* 612–621. Springer, Berlin.

SRIVASTAVA, A., KLASSEN, E., JOSHI, S. H. and JERMYN, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1415–1428.

STRAIT, J., KURTEK, S., BARTHA, E. and MACEACHERN, S. N. (2017). Landmark-constrained elastic shape analysis of planar curves. *J. Amer. Statist. Assoc.* **112** 521–533. MR3671749 https://doi.org/10.1080/01621459.2016.1236726

VAILLANT, M., MILLER, M. I., YOUNES, L. and TROUVÉ, A. (2004). Statistics on diffeomorphisms via tangent space representations. *NeuroImage* **23** S161–S169.

WEINBERGER, K. Q., BLITZER, J. and SAUL, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems* 1473–1480.

WU, S. G., BAO, F. S., XU, E. Y., WANG, Y.-X., CHANG, Y.-F. and XIANG, Q.-L. (2007). A leaf recognition algorithm for plant classification using probabilistic neural network. In *IEEE International Symposium on Signal Processing and Information Technology* 11–16.

YOUNES, L. (1998). Computable elastic distances between shapes. *SIAM J. Appl. Math.* **58** 565–586. MR1617630 https://doi.org/10.1137/S0036139995287685

YU, Q., MACEACHERN, S. N. and PERUGGIA, M. (2011). Bayesian synthesis: Combining subjective analyses, with an application to ozone data. *Ann. Appl. Stat.* **5** 1678–1698. MR2849791 https://doi.org/10.1214/10-AOAS444

ZAHN, C. T. and ROSKIES, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Trans. Comput.* **21** 269–281. MR0321383 https://doi.org/10.1109/tc.1972.5008949