Journal of Computational and Applied Mathematics 348 (2019) 120–129

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Learning shape metrics with Monte Carlo optimization

Serdar Cellat, Yu Fan, Washington Mio, Giray Ökten *

Department of Mathematics, Florida State University, Tallahassee, FL 32306, USA

ARTICLE INFO

Article history:
Received 8 March 2018
Received in revised form 16 July 2018

Keywords:
Landmark models
Simulated annealing
Shape metrics

ABSTRACT

Quantifying and modeling shape variation within a population, identifying morphological contrasts across groups, and categorizing individuals or objects according to morphological similarity are central problems in numerous domains of science and applications. In this paper, we present an approach to optimal shape categorization through a new family of metrics for shapes presented as a finite collection of labeled landmark points. We develop a technique to learn metrics that optimally differentiate and categorize shapes using Monte Carlo optimization methods. We discuss the theory and the practice of the methods and apply them to the analysis of synthetic data and the classification of multiple species of fruit flies based on the shape of their wings.

1. Introduction

The problems of modeling, classifying and recognizing shapes permeate many domains of science and applications. To exemplify, the investigation of organismal phenotypic variation and its genetic underpinnings often leads to complex problems in shape analysis (cf. [1–3]). The intricate morphology of pollen grains is of great interest to paleontologists as they form the most abundant and extensive record of plant diversity (cf. [4]). In ecology, there is evidence that variation in shape of spatial vegetation patterns might signal critical changes in ecosystems (cf. [5,6]). These are just a few examples in the broad landscape of problems that involve shape quantification and analysis.

Loosely speaking, the shape of an object is its geometry modulo position, orientation and scale, although depending on the context scale might not be filtered out. The first formal mathematical treatment of shape, due to Kendall, adopts a shape representation based on labeled landmark points [7]. This is illustrated in Fig. 1 that shows a fruit fly wing represented by twelve landmark points placed at vein crossings [8]. Kendall [9] constructed a shape space equipped with a metric that quantifies morphological similarity and contrast, providing a setting for systematic statistical analysis of shape variation (cf. [10]).

In the original model, all landmark points were treated as equally important in the process of constructing a shape space and defining a shape metric. Oftentimes, however, the morphological differences that most sharply contrast two or more shape populations, such as different species of fruit flies, are concentrated in particular regions. This suggests the investigation of inhomogeneous variants that highlight regions of interest. Statistical approaches to "weighted" landmark systems have been investigated in [11,12]. In this paper, we develop a shape space formulation that yields a family of shape spaces and metrics parameterized by $n \times n$ positive definite, symmetric matrices, where n is the number of landmark points. The matrices encode weights assigned to landmarks or linear combinations thereof, thus having the effect of attributing different importance to different parts of a shape.

This family of shape spaces and metrics at hand, in practice, one now has the problem of selecting a model that is "optimally" suited to an application. In this context, an important goal is to have computationally feasible ways of choosing

* Corresponding author.
E-mail address: okten@math.fsu.edu (G. Ökten).

https://doi.org/10.1016/j.cam.2018.08.043
0377-0427/© 2018 Elsevier B.V. All rights reserved.

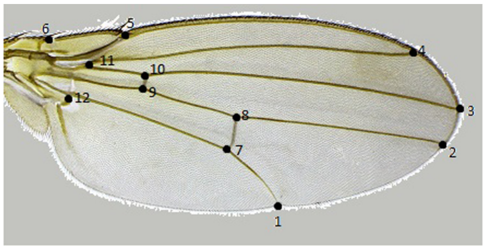


Fig. 1. Fruit fly wing with twelve landmark points.

a shape model that is particularly effective in solving a shape categorization problem such as classifying species of fruit flies. As the search space (of symmetric matrices) is typically high dimensional, this paper also investigates Monte Carlo methods to learn shape models for effective shape categorization. We formulate the question as an optimization problem in the vector space of trace-zero, symmetric matrices that we approach with simulated annealing. We illustrate the approach with synthetic data and validate the method through the classification of multiple species of fruit flies based on rather subtle differences in the shape of their wings.

We note that since the introduction of the landmark model there has been an explosion in developments in the field of shape analysis, particularly over the last two decades. Since the literature is vast, we just provide a few references that may help the interested reader navigate some of the shape literature. In many approaches, various types of morphological signatures or features are associated with objects and shape analysis is performed in feature space. These signatures typically are of a geometric, statistical or topological nature (cf. [13–17]). In other approaches, shape spaces whose elements are curves, surfaces or point clouds representing the shape of objects are employed and a variety of shape metrics have been investigated using techniques from differential and metric geometry (cf. [18–23]). Spectral geometry methods also have been used in multiple studies of shape (cf. [24,25]). These approaches have been applied to problems in a broad variety of areas. Nonetheless, development of techniques for learning a shape model that is well suited to a particular application is still incipient. This paper addresses the problem in the context of supervised learning of landmark models.

The remainder of the paper is organized as follows. In Section 2, we first review a formulation of the standard landmark model and then construct a family of shape spaces with weighted landmarks. In Section 4, we use Monte Carlo methods to solve optimization problems associated with learning shape models that optimize shape classification. A set of experiments with synthetic data are discussed in Section 5 to illustrate the method and the gains using the simulated annealing approach to the learning problem. Applications are presented in Section 6, where we apply our method to taxonomic classification of fruit flies based on the shape of their wings. We close with some additional discussion and remarks.

2. Landmark models of shape

In a landmark model, a shape in $\mathbb{R}^k$ is represented by a labeled collection of n points $p_1, \ldots, p_n \in \mathbb{R}^k$, which we encode as a $k \times n$ matrix $P = \begin{bmatrix} p_1 & \ldots & p_n \end{bmatrix}$. (Throughout the paper, we write vectors in $\mathbb{R}^k$ as column vectors and transposition of matrices is indicated by a superscript T.) The only restriction imposed is that not all landmark points be the same.

2.1. The classical model

Before introducing the general weighted model, we briefly review a formulation of the classical model that is based on the Euclidean metric on the space $\mathbb{R}^{k\times n}$ of all $k \times n$ matrices, induced by the usual inner product

$$\langle P, Q\rangle = \sum_{j=1}^{n} \langle p_j, q_j\rangle = \sum_{i=1}^{k}\sum_{j=1}^{n} p_{ij}q_{ij}. \tag{1}$$

(We abuse notation by also writing $\langle\, , \rangle$ for the usual dot product on $\mathbb{R}^k$.) The corresponding norm is the Frobenius norm $\|P\| = \langle P, P\rangle^{1/2}$.

The first step in defining shape is to obtain a representation that is insensitive to translations. This is done by restricting P to the $(kn-k)$-dimensional subspace $M(k, n)$ of centered matrices; that is, the subspace of matrices that satisfy $p_1+\cdots+p_n = 0$. We refer to the orthogonal projection $\pi : \mathbb{R}^{k\times n} \to M(k, n)$ as the centering operation. If $c = (p_1 + \cdots p_n)/n$ is the centroid of P, then centering is given by $\pi(P) = \begin{bmatrix} p_1 - c & \ldots & p_n - c \end{bmatrix}$.

For scale invariance, we simply restrict centered matrices to the unit sphere $\mathcal{P}(k, n)$ in $M(k, n)$. The corresponding scaling operation is given by $P \mapsto P/\|P\|$. Note that $P \neq 0$ because we are assuming that not all landmarks coincide. Centered matrices of unit Frobenius norm are known as *pre-shapes*. The pre-shape space $\mathcal{P}(k, n)$ is thus a (unit) sphere of dimension $kn - k - 1$.

To define shape, it only remains to account for orientation. In other words, to identify pre-shapes that differ by an orthogonal transformation as they have the same shape. To this end, define an action of the orthogonal group $O(k)$ by $(U, P) \mapsto UP$, where $U \in O(k)$ and $P \in \mathcal{P}(k, n)$. (In some models only rotations are considered, in which case $O(k)$ is replaced with the special orthogonal group $SO(k)$.) Clearly, $O(k)$ acts by isometries, that is, $\|UP\| = \|P\|$. The shape space $\mathcal{S}(k, n)$ is defined as the orbit space of this action; that is, the quotient space of $\mathcal{P}(k, n)$ under the equivalence relation $P_1 \sim P_2$ if and only if there is $U \in O(k)$ such that $P_1 = UP_2$. We denote the shape represented by a pre-shape P by $[P]$. The distance between shapes is defined as

$$d([P], [Q]) = \min_{U \in O(k)} \|P - UQ\|. \tag{2}$$

It is simple to verify that this definition is well posed. The calculation of shape distance is discussed in the next section in a more general setting. The process of finding U that minimizes (2), optimally aligning Q to P, is known as Procrustes alignment.

2.2. Weighted landmark models

We now construct a family of weighted landmark models, one for each positive definite, symmetric $n \times n$ matrix Σ using an inner product on $\mathbb{R}^{k\times n}$ that has the effect of attributing weights to certain linear combinations of landmarks. Given Σ, define an inner product on $\mathbb{R}^{k\times n}$ by

$$\langle P, Q \rangle_\Sigma = \langle P\Sigma, Q \rangle, \tag{3}$$

whose corresponding norm is denoted $\| \cdot \|_\Sigma$.

Remark. If Σ is diagonal with diagonal entries $\lambda_j > 0$, $1 \le j \le n$, then $\|P\|^2_\Sigma = \sum_{j=1}^n \lambda_j \|p_j\|^2$. Thus, we may interpret the eigenvalues of Σ as a weight system for the landmarks. More generally, diagonalize Σ writing $\Sigma = U\Lambda U^T$, with U orthogonal and Λ diagonal. Under the change of coordinates $P \mapsto \bar{P} = PU$, we have that $\|P\|_\Sigma = \|\bar{P}\|_\Lambda$. Hence, the Σ-norm may be viewed as attributing weights to linear combinations of the original landmarks via the eigenvalues and eigenvectors of Σ.

We now describe pre-shapes in the Σ-metric, starting with the centering operation. To motivate the construction, we first note that the usual centroid $c = (p_1 + \cdots + p_n)/n$ of $P \in \mathbb{R}^{k\times n}$ may be viewed as the unique minimizer of $V(x) = \sum_{j=1}^n \|p_j - x\|^2$, $x \in \mathbb{R}^k$. In matrix notation, this may be rewritten as follows. For $x \in \mathbb{R}^k$, let $\mathbb{I}(x) = \begin{bmatrix} x & \dots & x \end{bmatrix}$ be the $k \times n$ matrix whose columns are all equal to x. Then, $P - \mathbb{I}(x) = \begin{bmatrix} p_1 - x & \dots & p_n - x \end{bmatrix}$ so that the centroid may be viewed as the unique minimizer of $\|P - \mathbb{I}(x)\|^2$. Thus, we define the Σ-centroid of P as

$$c_\Sigma = \operatorname*{argmin}_{x \in \mathbb{R}^k} \|P - \mathbb{I}(x)\|^2_\Sigma. \tag{4}$$

Lemma 1. *The Σ-centroid of $P \in \mathbb{R}^{k\times n}$ is given by $c_\Sigma = P\Sigma e/|\Sigma|$, where $e \in \mathbb{R}^n$ is the vector whose components are all equal to 1 and $|\Sigma| = \langle e, \Sigma e\rangle$. ($|\Sigma| > 0$ since Σ is positive definite.)*

Proof. The Σ-centroid minimizes $u(x) = \frac{1}{2}\|P - \mathbb{I}(x)\|^2_\Sigma$. Noting that $\mathbb{I}(x) = xe^T$, we may write $u(x) = \left\langle (P - xe^T)\Sigma, P - xe^T \right\rangle/2$. A calculation shows that $\nabla u(x) = \langle e, \Sigma e\rangle\, x - P\Sigma e$, which only vanishes at $c_\Sigma = P\Sigma e/|\Sigma|$. □

The subspace $M_\Sigma(k, n)$ of Σ-centered $k \times n$ matrices is thus defined by the linear equation $P\Sigma e = 0$. The next proposition shows that the Σ-centering operation is given by $P \mapsto \begin{bmatrix} p_1 - c_\Sigma & \dots & p_n - c_\Sigma \end{bmatrix}$. Clearly, if Σ is the identity matrix this coincides with the usual centering operation.

Proposition 1. *The orthogonal projection $\pi_\Sigma : \mathbb{R}^{k\times n} \to M_\Sigma(k, n)$ is given by $\pi_\Sigma(P) = \begin{bmatrix} p_1 - c_\Sigma & \dots & p_n - c_\Sigma \end{bmatrix}$.*

Proof. Let $e_1, \dots, e_k$ be the canonical basis of $\mathbb{R}^k$ and set $V_i = e_i e^T/|\Sigma|^{1/2} \in \mathbb{R}^{k\times n}$. The matrices V_i, $1 \le i \le k$, form a Σ-orthonormal set. Indeed,

$$\left\langle V_i, V_j \right\rangle_\Sigma = \frac{1}{|\Sigma|} \left\langle e_i e^T, e_j e^T \Sigma \right\rangle = \frac{1}{|\Sigma|} \left\langle e_i, e_j \right\rangle \langle e, \Sigma e\rangle = \left\langle e_i, e_j \right\rangle = \delta_{ij}.$$

As a matter of fact, these matrices form an orthonormal basis of the orthogonal complement of $M_\Sigma(k, n)$. This follows from the fact that, for any Σ-centered P and $1 \le i \le k$,

$$\langle V_i, P \rangle_\Sigma = \frac{1}{|\Sigma|} \left\langle e_i e^T, P\Sigma \right\rangle = \frac{1}{|\Sigma|} \langle e_i, P\Sigma e\rangle = \frac{1}{|\Sigma|} \langle e_i, 0\rangle = 0.$$

Hence, the orthogonal projection π_Σ may be calculated as

$$\begin{aligned}\pi_\Sigma(P) &= P - \sum_{i=1}^{k} \langle V_i, P\rangle_\Sigma V_i = P - \frac{1}{|\Sigma|}\left\langle e_i e^T, P\Sigma\right\rangle e_i e^T \\ &= P - \frac{1}{|\Sigma|} e_i \langle e_i, P\Sigma e\rangle\, e^T = P - \frac{1}{|\Sigma|} e_i e_i^T P\Sigma e e^T \\ &= P - e_i e_i^T \left(\frac{P\Sigma e}{|\Sigma|}\right) e^T = P - e_i e_i^T c_\Sigma e^T \\ &= P - \begin{bmatrix} p_1 - c_\Sigma & \ldots & p_n - c_\Sigma \end{bmatrix}.\end{aligned} \tag{5}$$

This concludes the proof. □

The pre-shape space $\mathcal{P}_\Sigma(k, n)$ is defined as

$$\mathcal{P}_\Sigma(k, n) = \{P \in M_\Sigma(k, n) : \|P\|_\Sigma = 1\}, \tag{6}$$

the unit sphere in $M_\Sigma(k, n)$. It is straightforward to verify that $\mathcal{P}_\Sigma(k, n)$ is invariant under the $O(k)$-action $(U, P) \mapsto UP$ and that $O(k)$ acts by Σ-isometries. As before, the shape space $\mathcal{S}_\Sigma(k, n)$ is defined as the orbit space of this action and equipped with the metric

$$d_\Sigma([P], [Q]) = \min_{U \in O(k)} \|P - UQ\|_\Sigma. \tag{7}$$

The minimization problem in (7) has a well-known closed form solution (cf. [7,11]), described next.

Proposition 2. *Let $P, Q \in \mathcal{P}_\Sigma(k, n)$, $X = P\Sigma Q^T$, and $X = U_1 \Lambda V_1^T$ be a singular value decomposition of X, where $U_1, V_1 \in O(k)$ and Λ is the diagonal matrix of singular values. Then, $\hat{U} = U_1 V_1^T$ minimizes $f(U) = \|P - UQ\|_\Sigma$, $U \in O(k)$.*

Proof. We write

$$\begin{aligned}\|P - UQ\|_\Sigma^2 &= \langle P - UQ, P - UQ\rangle_\Sigma = \|P\|_\Sigma^2 - 2\langle P, UQ\rangle_\Sigma + \|UQ\|_\Sigma^2 \\ &= 2 - 2\langle P, UQ\rangle_\Sigma = 2 - 2\,\langle P\Sigma, UQ\rangle = 2 - 2\left\langle P\Sigma Q^T, U\right\rangle.\end{aligned}$$

Thus, minimizing f is the same as maximizing $\left\langle P\Sigma Q^T, U\right\rangle$. Using the SVD of X, we write $\left\langle P\Sigma Q^T, U\right\rangle = \left\langle U_1 \Lambda V_1^T, U\right\rangle = \left\langle \Lambda, U_1^T U V_1\right\rangle$. It follows that $\left\langle P\Sigma Q^T, U\right\rangle \leq \operatorname{trace} \Lambda$ because Λ is diagonal with non-negative entries and all entries of $U_1^T U V_1$ are less than or equal to 1 since this matrix is orthogonal. The upper bound given by the trace is realized if $U_1^T U V_1 = I_k$, where I_k is the identity matrix. Thus, $\hat{U} = U_1 V_1^T$ minimizes f as claimed. □

2.3. Mean shapes

Having constructed the shape space $(\mathcal{S}_\Sigma(k, n), d_\Sigma)$, we now have an environment to carry out statistical shape analysis with weighted landmark systems. In the classical setting, $\Sigma = I_n$, shape statistics has been investigated extensively (cf. [10,9]). A basic statistic is a mean shape that we discuss next.

Let α be a Borel probability measure on the shape space $(\mathcal{S}_\Sigma(k, n), d_\Sigma)$. A *mean shape* of α is a minimizer of the Fréchet function $V : \mathcal{S}_\Sigma(k, n) \to \mathbb{R}$ defined as

$$V([P]) = \frac{1}{2}\,\mathbb{E}\,[d_\Sigma^2([P], \cdot)] = \frac{1}{2}\int_{\mathcal{S}_\Sigma(k,n)} d_\Sigma^2([P], [Q])d\alpha([Q]), \tag{8}$$

the expected value of the distance-square function (up to the factor 1/2). For a dataset of shapes $[P_1], \ldots, [P_\ell] \in \mathcal{S}_\Sigma(k, n)$, a *mean shape* is defined as a mean of the associated empirical distribution $\alpha_\ell = \left(\delta_{[P_1]} + \ldots + \delta_{[P_\ell]}\right)/\ell$, the uniform mixture of Dirac measures supported on the data points. In other words, a minimizer of

$$V_\ell([P]) = \frac{1}{2\ell}\sum_{i=1}^{\ell} d_\Sigma^2([P], [P_i]). \tag{9}$$

Unlike the Euclidean case, in general, the mean shape is not unique. However, for distributions whose support have "sufficiently small" diameter, empirical evidence suggests that uniqueness may hold. Under this assumption, we describe an efficient algorithm for estimating the mean shape that will be used extensively in our applications to metric learning. The algorithm is based on an attracting fixed point principle, reminiscent of the method proposed in [26] for the estimation of the mean in the classical landmark model in $\mathbb{R}^2$.

For a dataset of shapes, as above, the method searches for the (empirical) mean shape at the pre-shape level; that is, on the subspace $M_\Sigma(k,n) \subset \mathbb{R}^{k\times n}$ of Σ-centered matrices subject to the constraint $\|P\|_\Sigma = 1$. In this setting, the Fréchet function becomes

$$V_\ell(P) = \frac{1}{2\ell}\sum_{i=1}^{\ell} \langle P - U_iP_i, P - U_iP_i\rangle_\Sigma, \tag{10}$$

where $U_i \in O(k)$ optimally aligns P_i to P. We write the Lagrangian as

$$L(P,\lambda) = V_\ell(P) - \frac{\lambda}{2}\langle P, P\rangle_\Sigma, \tag{11}$$

whose P-gradient with respect to the inner product $\langle\cdot\,,\cdot\rangle_\Sigma$ is

$$\nabla_P L(P,\lambda) = \nabla V_\ell(P) - \lambda P, \tag{12}$$

where $\nabla V_\ell(P) = P - \frac{1}{\ell}\sum_{i=1}^{\ell} U_iP_i$ (the term U_i depends on P, but its derivative vanishes because of the optimal alignment property). Thus, at a point where $\nabla_P L = 0$, we must have $\lambda P = \nabla V_\ell(P)$, implying that P and $\nabla V_\ell(P)$ must be parallel. Since $\|P\|_\Sigma = 1$, at a constrained minimum we have

$$P = \operatorname{sign}(a_P)\frac{\nabla V_\ell(P)}{\|\nabla V_\ell(P)\|_\Sigma}, \tag{13}$$

where $a_P = \langle \nabla V_\ell(P), P\rangle_\Sigma = 1 - \frac{1}{\ell}\sum_{i=1}^{\ell}\langle U_iP_i, P\rangle_\Sigma$. Thus, at a minimum, P should be a fixed point of the mapping $T : \mathcal{P}_\Sigma(k,n) \to \mathcal{P}_\Sigma(k,n)$ defined by

$$T(P) = \operatorname{sign}(a_P)\frac{\nabla V_\ell(P)}{\|\nabla V_\ell(P)\|_\Sigma}. \tag{14}$$

Experiments suggest that mean pre-shapes are attracting fixed points of T, leading to the following algorithm.

Mean Shape Algorithm (pseudo code)

```
1. Choose a threshold $\epsilon > 0$ and initialize the search with a random pre-shape $P$.
2. Calculate $T(P)$.
3. Calculate $\varepsilon(P) = \|T(P) - P\|_\Sigma$.
4. If $\varepsilon(P) < \epsilon$, output $P$. Else, update $P \leftarrow T(P)$ and go to Step 2.
```

3. Learning metrics for shape classification

In this section, we present a method for learning a shape model that is "optimal" for a classification problem involving multiple shape classes, with the aid of a training set. We follow the general principle that a good metric should make the various classes as compact and well separated as possible. This is evocative of the principle used in subspace learning via linear discriminant analysis. However, here the principle is used in a very different manner since we are learning shape metrics, not performing linear dimension reduction. We use the notions of *within-class scatter* as an overall measure of compactness of the shape classes and *between-class scatter* to quantify their separation. These are assembled into a single cost function that we seek to minimize to optimize the choice of shape model.

If $r > 0$, then the mapping $P \mapsto P/\sqrt{r}$ induces an isometry between the shape spaces defined by Σ and $r\Sigma$. Thus, we normalize Σ to satisfy $\det\Sigma = 1$. Write $\Sigma = \exp A$, where A is a symmetric matrix. The logarithmic change of coordinates $\Sigma \mapsto A$ maps positive definite, symmetric matrices to arbitrary symmetric matrices and the normalization $\det\Sigma = 1$ translates to $\operatorname{tr} A = 0$; the zero trace constraint. From a computational perspective, this is very helpful as the search for optimal models may be performed on the full linear subspace T_n of $n\times n$ trace-zero, symmetric matrices with no additional constraints. For this reason, we formulate metric learning in A-coordinates.

Let m be the number of shape classes in a metric learning problem. We denote by C_i the collection of training samples for the ith class, $1 \le i \le m$, and let $n_i = |C_i|$. The elements of C_i are denoted $P(i,j)$, $1 \le j \le n_i$, where each $P(i,j)$ is a $k\times n$ matrix representing a shape. Given $A \in T_n$, we let $P(i,j;A) \in \mathcal{P}_\Sigma(k,n)$ be the pre-shape obtained by centering and normalizing $P(i,j)$ with respect to $\Sigma = \exp A$. We let $M_i(A) \in \mathcal{P}_\Sigma(k,n)$ be a pre-shape that represents the Σ-mean of the class C_i, and $\hat{P}(i,j;A)$ the pre-shape obtained by aligning $P(i,j;A)$ to the group mean $M_i(A)$ so that $d_\Sigma([P(i,j;A)],[M_i(A)]) = \|\hat{P}(i,j;A) - M_i(A)\|_\Sigma$. With this notation, the within-class scatter function is defined as

$$S_W(A) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \|\hat{P}(i,j;A) - M_i(A)\|^2_{e^A} \tag{15}$$

and the between-class scatter function as

$$S_B(A) = \sum_{i=1}^{m} n_i \|\hat{M}_i(A) - M\|^2_{e^A}, \tag{16}$$

where the pre-shape $M \in \mathcal{P}_\Sigma(k, n)$ represents the mean of the entire training set $\cup_{i=1}^m C_i$, and $\hat{M}_i(A)$ is obtained by Σ-aligning M_i to M.

We are interested primarily in the ratio $S_W(A)/S_B(A)$, as it gives a single scalar that quantifies the suitability of a proposed Σ-model. A small value indicates that the metric defined by $\Sigma = \exp A$ shapes up the training data into well delineated clusters. Thus, we formulate the metric learning problem as the minimization of $F : T_n \to \mathbb{R}$ defined as

$$F(A) = \frac{S_W(A)}{S_B(A)} + c\|A\|^2, \tag{17}$$

where $c > 0$ is a constant. The second summand is a regularization term that ensures that the optimization problem is well posed.

Deterministic gradient methods for the minimization of F have a few drawbacks. Besides the usual problem of escaping local minima, computations may get costly since each numerical estimation of ∇F requires $O(n^2)$ evaluations of the cost function because $\dim T_n = \frac{1}{2}(n^2 + n) - 1$. For these reasons, we turn to Monte Carlo methods that are discussed in the next section.

4. Monte Carlo optimization

We approach the optimization of the cost function defined in (17) with the simulated annealing method, a technique developed by Kirkpatrick et. al [27] that is an adaptation of the Metropolis–Hastings algorithm [28,29]. For a survey of simulated annealing, we refer the reader to [30]. We give a brief description of the method and how it is used in our application.

Let f be a real-valued cost function defined on a *state space* (also referred to as solution space) Ω. The aim is to find a global minimum or maximum of f on Ω. We consider the minimization of f with the assumption that the problem is well posed, but maximization may be treated similarly. The method assumes a neighborhood structure that prescribes a neighborhood $N(\omega)$ of any $\omega \in \Omega$. For example, if Ω is Euclidean space, $N(\omega)$ may be a hypercube about ω with a fixed size. The simulated annealing algorithm constructs a random walk that, at each step, makes a transition from a state (candidate solution) ω to another state ω' in its neighborhood $N(\omega)$.

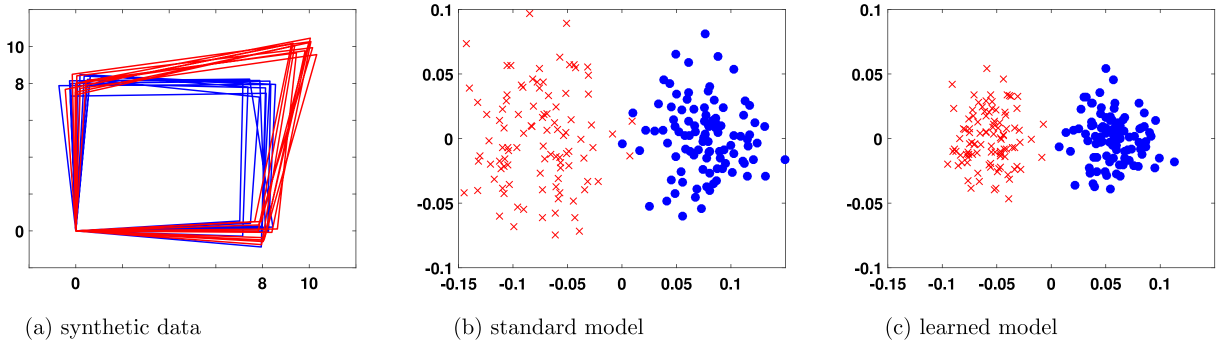
The initial state $\omega \in \Omega$ usually is chosen at random. At each iteration, a candidate neighboring state $\omega' \in N(\omega)$ also is randomly generated. Let $E_{\omega,\omega'}$ be the event that ω' is accepted as the next state. The probability of $E_{\omega,\omega'}$ is called the *acceptance probability* and is given by

$$P(E_{\omega,\omega'}, T) = \begin{cases} \exp\left(-\dfrac{f(\omega') - f(\omega)}{T}\right), & \text{if } f(\omega') > f(\omega); \\ 1, & \text{if } f(\omega') \le f(\omega). \end{cases} \tag{18}$$

The acceptance probability depends on a *temperature parameter* $T > 0$, which follows a cooling schedule $T = t_k$ satisfying $t_k \downarrow 0$, as $k \to \infty$. This cooling schedule is a crucial element of the search mechanism. Under certain hypotheses, if the convergence $t_k \downarrow 0$ is sufficiently slow, then the algorithm converges to an optimal solution with probability 1. A discussion of convergence results may be found in [31]. A common choice is a cooling schedule that satisfies $t_{k+1} = ct_k$, where $0 < c < 1$. The next algorithm uses such a cooling schedule, as well as a repetition schedule $\mathcal{M} = \{M_k\}_{k=0}^\infty$, which gives the number of iterations M_k at temperature t_k.

Simulated Annealing Algorithm (pseudo code)

```
1. Select an initial state $\omega \in \Omega$, an initial temperature $T = t_0$, a repetition schedule $\mathcal{M}$, and $0 < c < 1$.
2. Set a counter at $k = 0$ and a stop criterion.
3. For $i = 1 : M_k$
      (a) Randomly generate a state $\omega' \in N(\omega)$.
      (b) Compute the acceptance probability $P(E_{\omega,\omega'}, T)$.
      (c) Generate a random $u \in (0, 1)$ from the uniform distribution.
      (d) If $u \le P(E_{\omega,\omega'}(k))$, then update $\omega \leftarrow \omega'$.
4. If the stop criterion is satisfied, output the solution $\omega$.
   Else, update $k \leftarrow k + 1$, $T \leftarrow cT$, and go to Step 3.
```



(a) synthetic data (b) standard model (c) learned model

Fig. 2. (a) 4-landmark data; MDS visualizations of the organization of training data under the (b) standard and (c) learned shape metrics, resp.

Remark. A simple, commonly used criterion is to stop the algorithm as the temperature falls below a set threshold. Another useful criterion is to stop if improvements in the cost function consistently fall below a set threshold after a prescribed number of steps. The latter is used in our experiments and applications. In our implementation of the algorithm, we keep track of the best overall state ω_{best} in the memory: when a random state ω' is generated, $f(\omega')$ is compared with $f(\omega_{\text{best}})$, and the state with smaller function value is set as the new ω_{best}. When the algorithm stops with output w, we compare ω with ω_{best}, and the final output is the one that has the lower function value. With this implementation, we guarantee the algorithm picks the best state among all the states it has visited.

In our applications, Ω is the space T_n of trace-zero $n \times n$ symmetric matrices. Given a state $A \in T_n$, the neighborhood structure used is of the form

$$N(A) = \{B \in T_n \mid \sup_{1\le i,j\le n} |a_{ij} - b_{ij}| < \delta\}, \tag{19}$$

for some fixed $\delta > 0$. Random states A' in $N(A)$ are generated as follows. For $1 \le i \le j \le n$, we generate independent random numbers ζ_{ij} from the uniform distribution on $(-\delta, \delta)$ and add them to the entries a_{ij} and a_{ji} of A. Then, we orthogonally project the resulting symmetric matrix to T_n, obtaining a trace-zero symmetric matrix A'. Following [27], in all experiments, we set $c = 0.9$[1] .

5. Preliminary experiments

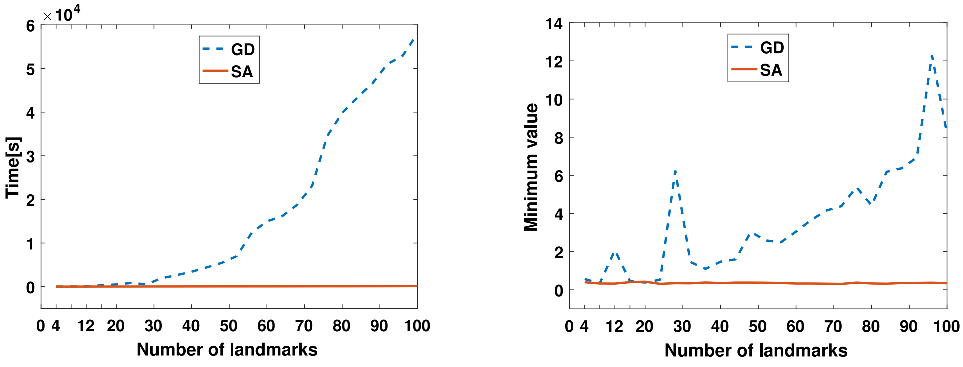
Our first example uses synthetic data to illustrate the improvement in shape discrimination that can be achieved through the metric learning method. We generated two shape groups, G_1 and G_2, as follows. Each shape in G_1 is represented by four landmarks, $p_1, p_2, p_3, p_4 \in \mathbb{R}^2$, sampled independently from Gaussian random variables with standard deviation $\sigma = 0.5$ and centered at $(0, 0)$, $(0, 8)$, $(8, 8)$ and $(8, 0)$, respectively. G_2 is generated in a similar manner, except that the center of the Gaussian for the third landmark is shifted to $(10, 10)$. For each group, we generated 100 training shapes[2] and 500 test shapes. Fig. 2a shows a few samples in each group, with some landmarks joined by line segments to facilitate visualization.

Fig. 2b and c show multidimensional scaling (MDS) visualizations of the organization of the two groups in shape space under the standard and learned shape models, respectively. In this simple case, the standard shape metric already yields a good group separation, but the separation is sharpened by the learned shape model that, by design, also makes the clusters more compact. For a more objective comparison, we employed the test set of 1000 shapes to estimate the classification accuracy using the nearest neighbor classifier for both shape metrics. The classification accuracy improved from 94.2% for the classical shape model to 98.3% for the learned model.

To illustrate the gains resulting from the simulated annealing (SA) approach over gradient descent (GD) and examine the sensitivity of the learned model to the training set, we consider variants of the previous example in which we increase the number n of landmarks by sampling Gaussians centered at more points along the "edges" of the rectangles, randomizing the choice of training set. We randomly generated 100 shapes for two shape groups for several values of n in the range $4 \le n \le 100$. Fig. 3a shows the computing time for GD and SA against the number of landmarks. As expected, the improvement in computational efficiency is significant. At each iteration, GD requires $O(n^2)$ evaluations of the cost function, whereas a single evaluation is needed in SA. The quality of the learned shape metric, as measured by the cost function, also improves substantially with SA. Fig. 3b shows the minimum values of the cost function obtained by each method for several values of n. The improvement with SA is particularly pronounced for larger values of n, suggesting the existence of many local minima in higher dimensions.

[1] We have empirically tested the effect of different values for c on the performance of the algorithm using some standard test functions from global optimization, specifically, Ackley, Rosenbrock and Six-Hump Camel functions, and concluded that $c = 0.9$ results in the best performance.

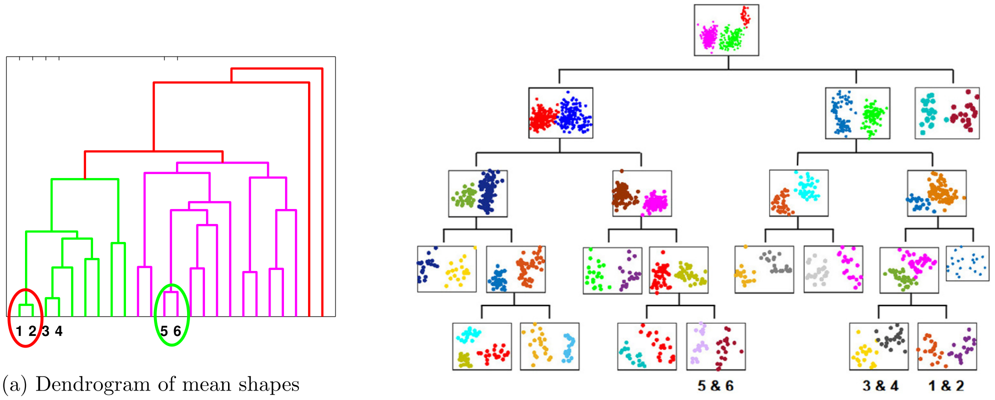
[2] We picked the number of training shapes to be 100 because that is about the number of samples we can afford to use for training in the fly wing dataset that will be discussed in the next section. This helps to assess the power of our numerical approach when the training size is as small as 100.



(a) Computation time

(b) Cost function

Fig. 3. Gradient descent versus simulated annealing: (a) computation time against the number of landmarks; (b) computed minimum values of the cost function against the number of landmarks.



(a) Dendrogram of mean shapes

(b) Classification tree

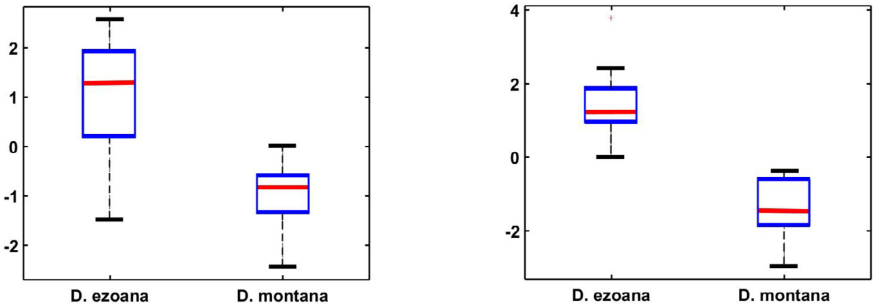
Fig. 4. Wing of 24 fly species: (a) dendrogram of mean shapes indicating morphological similarity across species; (b) taxonomic classification tree. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 3 suggests SA is fairly insensitive to the number of landmarks. The performance of SA as a function of the dimension of the function was further investigated numerically by Cellat [32]. When SA was applied to the test functions by Ackley and Rosenbrock, its performance did not deteriorate as the dimension of the function was increased from 2 to 100 in the former, and 2 to 25 for the latter function. However, for the Rastrigin and Sphere functions, there was a slight deterioration in the performance as the dimension increased. The results also serve as empirical evidence of stability of the output of the algorithm with respect to the training set as characterized by the minimum of the cost function.

6. Classification of fly wings

In this section, we present the results of a morphological analysis of fruit fly wings that used the simulated annealing method discussed in Section 4. The dataset used, courtesy of Houle Lab at Florida State University, comprised wing shapes for 24 *Drosophila* species. Each fly wing was represented by 12 landmark points as indicated in Fig. 1. The objective was to perform taxonomic classification based on wing shapes at the highest possible accuracy. The number of wing shapes for each fly species ranged from 181 to 268, with 215 shapes on average. We split the data into training and test sets using 100 randomly selected samples for each species as training shapes and the remaining samples as test shapes.

As the number of species was rather large and key morphological contrasts could be located in different parts of the wing, a hierarchical approach to shape classification proved to yield better results than categorization based on a single shape metric. A classification tree that used different shape models at its various nodes was constructed as follows. Starting with the classical shape model, $\Sigma = I_n$, we computed the mean training shape for each species and performed a hierarchical clustering of the 24 mean shapes using the single linkage method. Fig. 4a shows the resulting dendrogram that guided our initial hierarchical grouping of the various species, as indicated in Fig. 4b. At the top level, the species were grouped into three clusters colored green, magenta and red in the dendrogram. Similar groupings, guided by the dendrogram, were



(a) standard metric

(b) learned metric

Fig. 5. Discrimination of *D. ezoana* and *D. montana* on the training set.

constructed at other levels of the hierarchy. At each node of the resulting tree, we tested how well the standard shape metric solved the corresponding classification problem. For example, at the top node, the standard metric and the nearest neighbor classifier yielded a classification accuracy of 99.7% for the test set indicating that there was no need to learn a new shape metric at that node. For nodes at which the accuracy was lower, we applied our method to learn a shape metric and improve classification rates. For example, the distance between the mean shapes for the species *D. ezoana* and *D. montana* (labeled 1 and 2) is the smallest indicating that their wings exhibit significant shape similarity and may be difficult to distinguish. The classification accuracy at that node improved from 62.3% to 74.6% for the test set. Fig. 5 offers a boxplot visualization of the improved species discrimination with the learned metric, constructed as follows. We first used tangent-space PCA to obtain a Euclidean representation of the data using the first 8 principal components, which account for 91% of the variation. The data was then projected onto an axis via linear discriminant analysis. Altogether, the method was applied to 3 nodes of the tree. More specifically, metric learning was used for the discrimination of the three pairs of species whose mean shapes were most similar, namely $(1, 2)$, $(3, 4)$ and $(5, 6)$, as indicated in Fig. 4a. The classification accuracy at each of these 3 nodes increased by at least 10% with the learned metric. The overall classification accuracy using the full classification tree improved from 89.9% (with the standard shape metric at all nodes) to 97.8% using the tree modified by the 3 learned metrics.

7. Summary and concluding remarks

We introduced a shape space formulation of statistical models that represent shape by a finite collection of labeled landmarks and attribute different weights to landmarks or linear combinations thereof. The models are thus able to highlight particular regions of morphological interest. We formulated the problem of shape categorization as a learning problem that seeks to identify particular metrics that optimize shape classification by making clusters as compact and well separated as possible. As the learning problem typically involves optimization in high dimensions, we employed Monte Carlo methods for computational feasibility and efficiency. We illustrated the method with applications to synthetic data and applied it to the categorization of multiple species of fruit flies based on often subtle differences on wing shape.

Although we focused on shape classification, the family of weighted shape metrics developed in the paper should be useful in other contexts, particularly in situations where key morphological variation tends to be localized to particular regions. Many of the methods of statistical shape analysis developed for the standard shape metric should have weighted counterparts such as weighted mean shapes and tangent-space principal component analysis used in this paper.

Acknowledgments

We thank David Houle (Biological Sciences, FSU) for sharing the fly wing data. This work was supported in part by NSF grant DBI-1052942.

References

[1] D.W. Thompson, On Growth and Form, Cambridge University Press, 1917.

[2] D. Houle, G. Bolstad, K.V. der Linde, T. Hansen, Mutation predicts 40 million years of fly wing evolution, Nature 548 (2017).

[3] Q. Xu, H. Jamniczky, D. Hu, R.M. Green, R.S. Marcucio, B. Hallgrimsson, W. Mio, Correlations between the morphology of sonic hedgehog expression domains and embryonic craniofacial shape, Evol. Biol. 42 (2015) 379–386.

[4] L. Mander, M. Li, W. Mio, C. Fowlkes, S. Punyasena, Identification of grass pollen through the quantitative analysis of surface ornamentation and texture, Proc. Royal Soc. B 280 (2013) 20132905, http://dx.doi.org/10.1098/rspb.2013.1905.

[5] F. Borgogno, P. D'Odorico, F. Laio, L. Ridolfi, Mathematical models of vegetation pattern formation in ecohydrology, Rev. Geophys. 47 (2009) RG1005, http://dx.doi.org/10.1029/2007RG000256.

[6] L. Mander, S.C. Dekker, M. Li, W. Mio, S.W. Punyasena, T.M. Lenton, A Morphometric analysis of vegetation patterns in dryland ecosystems, Royal Soc. Open Sci. 4 (2017) 160443, http://dx.doi.org/10.1098/rsos.160443.

[7] D.G. Kendall, Shape manifolds, Procrustean metrics, and complex projective spaces, Bull. Lond. Math. Soc. 16 (2) (1984) 81–121.

[8] D. Houle, J. Mezey, P. Galpern, A. Carter, Automated measurement of Drosophila wings, BMC Evol. Biol. 3 (1) (2003) 25, http://dx.doi.org/10.1186/1471-2148-3-25.

[9] H.L. Le, D.G. Kendall, The Riemannian structure of Euclidean shape spaces: a novel environment for statistics, Ann. Statist. 21 (1993) 1225–1271.

[10] I.L. Dryden, K.V. Mardia, Statistical Shape Analysis, John Wiley & Sons, 1998.

[11] C. Goodall, Procrustes Methods in the Statistical Analysis of Shape, J. R. Stat. Soc. Series B Stat. Methodol. 53 (1991) 285–339.

[12] D. Theobald, D.S. Wuttke, Empirical Bayes hierarchical models for regularizing maximum likelihood estimation in the matrix Gaussian Procrustes problem, Proc. Natl. Acad. Sci. USA 103 (49) (2006) 18521–18527, http://dx.doi.org/10.1073/pnas.0508445103.

[13] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 509–521.

[14] T. Sebastian, P. Klein, B. Kimia, Recognition of shapes by editing their shock graphs, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 550–571.

[15] H. Ling, D. Jacobs, Shape classification using the inner distance, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 286–299.

[16] G. Carlsson, A. Zomorodian, A. Collins, L. Guibas, Persistence barcodes for shapes, Int. J. Shape Model. 11 (2005) 149–187.

[17] P. Frosini, G. Jabłoński, Persistence barcodes for shapes, Int. J. Shape Model. 11 (2005) 149–187.

[18] P.W. Michor, D. Mumford, An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach, Appl. Comput. Harmon. Anal. 23 (2007) 74–113.

[19] X. Liu, Y. Shi, I. Dinov, W. Mio, A Computational Model of Multidimensional Shape, Int. J. Comput. Vis. 89 (2010) 69–83.

[20] W. Mio, J.C. Bowers, X. Liu, Shape of elastic strings in Euclidean space, Int. J. Comput. Vis. 82 (2009) 96–112.

[21] X. Liu, W. Mio, Y. Shi, I. Dinov, X. Liu, N. Leporé, F. Leporé, M. Fortin, P. Voss, M. Lassonde, P.M. Thompson, Models of normal variation and local contrasts in hippocampal anatomy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2010, pp. 407–415.

[22] M. Bauer, P. Harms, P.W. Michor, Sobolev metrics on the manifold of all Riemannian metrics, J. Differential Geom. 94 (2013) 187–208.

[23] F. Mémoli, Gromov-Wasserstein distances and the metric approach to object matching, Found. Comput. Math. 11 (2011) 1–71.

[24] M. Reuter, F.-E. Wolter, N. Peineke, Laplace-Beltrami spectra as ‘‘Shape-DNA’’ of surfaces and solids, Comput. Aided Des. 38 (2006) 342–366.

[25] F. Mémoli, A spectral notion of Gromov-Wasserstein distances and related methods, Appl. Comput. Harmon. Anal. 30 (2011) 363–401, http://dx.doi.org/10.1016/j.acha.2010.09.005.

[26] S. Huckemann, H. Ziezold, Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces, Adv. Appl. Probab. (2006) 299–319.

[27] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671–680, http://dx.doi.org/10.1126/science.220.4598.671.

[28] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087–1092.

[29] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1) (1970) 97–109.

[30] K.A. Dowsland, J. Thompson, Simulated annealing, in: Handbook of Natural Computing, Springer-Verlag, 2012, pp. 1623–1655.

[31] D. Mitra, F. Romeo, A. Sangiovanni-Vincentelli, Convergence and finite-time behavior of simulated annealing, Adv. Appl. Probab. 18 (3) (1986) 747–771.

[32] S. Cellat, Metric Learning for Shape Classification: A Fast and Efficient Approach with Monte Carlo Methods (Ph.D. Dissertation), Florida State University, 2018.