



Landmark-based algorithms for group average and pattern recognition

Snehalata Huzurbazar^a, Dongyang Kuang^b, Long Lee^{c,*}

^a Department of Biostatistics, School of Public Health, West Virginia University, 1 Medical Center Dr., PO Box 9190, Morgantown, WV 26506, USA

^b Department of Mathematics and Statistics, University of Ottawa, Canada

^c Department of Mathematics and Statistics, University of Wyoming, Laramie, WY 82071-3036, USA

ARTICLE INFO

Article history:

Received 21 February 2018

Revised 15 August 2018

Accepted 5 September 2018

Available online 13 September 2018

Keywords:

Group average

Pattern recognition

Features extraction

Landmark

Template matching

Residual momentum

Cluster analysis

Outliers

Structure abnormality

ABSTRACT

We introduce a class of mathematical algorithms with the aim of establishing a framework of finding a group average and extracting prominent features in a group of landmark represented shapes or image templates. A group average is an estimator that is said to best represent the common features of the group being studied. The proposed algorithms, as a tool of feature extraction, extract information about momentum at each landmark through the process of template matching. Once the convergence criterion is satisfied numerically, the algorithms produce a group average and a local coordinate system for each member of the observing group, in terms of the residual momentum. We present several examples to illustrate the use of the proposed algorithms for finding a group average. Using the metrics computed between the group average and each member of the group, we successfully run a cluster analysis for datasets that contain a heavy percentage of outliers. Finally, we apply the collected residual momenta computed in the proposed algorithms in some statistical methods to demonstrate a potential application of the algorithms for detecting structure abnormality.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

In the past 60 years, image analysis and related fields, including pattern recognition, have developed at an enormous pace especially propelled by computational power. Pattern recognition relies on data where prominent features are extracted from image templates. The process of identifying these features generally includes two major steps: (1) transforming the image or shape representations from nonlinear spaces into locally linear coordinates for any necessary calculations, and (2) making inferences from these computations via statistical analyses. Among various representations used, landmark representation, despite its long history and simplicity, has proven to be one of the most effective tools in making statistical inferences on shape or image recognition and matching [14,18,20]. In particular, recent progress on automatic key points detection and unsupervised landmark annotation [39,43] help bring back the spotlight on this classical descriptor by eliminating the labor-intensive manual landmark annotation.

The traditional Procrustes methods extract residual vectors of landmark locations after calculating a group average, a quantity similar to the statistical mean [3,4], and then inferences are based

on classical statistical analyses of the residual vectors. Alternatively, a class of diffeomorphic methods, first introduced in the 1990s [11,17] with more recent theoretical and numerical development [1,2,22,30–32,40,44], model the given image data through certain manifolds and relate two deformable templates through geodesics on the selected manifolds. For detailed information about the development of this class of methods, we refer readers to the book and paper by Younes et al. [33,45]. The geodesic of a Riemannian manifold is governed by a class of partial differential equations (PDE) called the Euler-Poincaré (EP) equations [7,21,29,34,44]. When images or shapes are represented by landmarks, geodesic matching algorithms can be constructed by finding the proper initial conditions of the equivalent finite dimensional particle system of the EP equations. An efficient template matching algorithm that takes advantage of the landmark representation and the particle formulation of the EP equation was introduced by Camassa et al. [5,6]. This method uses a constant matrix to update the search direction of the geodesic shooting for finding the initial conditions that evolve the reference template to the target one, instead of the traditional methods of forward-backward integration for updating the gradient (known as the LDDMM), or Newton's optimization. Another novel feature of this method is that a non-smooth conic kernel for the particle system is used to accelerate the convergence of matching. A brief introduction of this geodesic matching algorithm is given in Section 2. Using the above

* Corresponding author.

E-mail addresses: snehalata.huzurbazar@hsc.wvu.edu (S. Huzurbazar), dkuang@uOttawa.ca (D. Kuang), ilee@uwyo.edu (L. Lee).

geodesic matching algorithm as our underlying warping method, in this paper we propose a landmark-based algorithm that gives an estimator as an average of a group on a general Riemannian manifold. The proposed algorithm utilizes the mappings (Exp and Log maps) under the landmarks representation and communicates between the landmark space and its cotangent plane (the momentum field) through an iterative process of solving the EP equations in its Hamiltonian formulation. The momentum field provides a local template-based coordinate system and is linear in nature [21], which yields advantages for statistical analyses.

The paper is organized as follows. In Sections 2 and 3, an averaging algorithm with different weighting schemes on the momentum field under the setting of deformable template is constructed. We show how the momentum representation for each template is collected during the averaging procedure. Several examples are presented to illustrate the use of the proposed algorithms for finding an estimator that represents prominent features in a group of shapes. In Section 4 we demonstrate a potential application of the group average (as an estimator) for a small group of images. We use the momentum representation based on the calculated average template and a Bayesian predictive statistical approach to detect landmarks whose momentum coordinates encode the most significant differences between groups under comparison.

2. Basic formulation for template matching

For template matching, based on the principle of diffeomorphism, the warping between two images is established through the geodesic defined on the diffeomorphism group $\text{Diff}(\mathbb{R}^n)$ or $\text{Diff}(D)$ for some $D \subseteq \mathbb{R}^n$, which is an open subset of the space $I_d + C_c^\infty(\mathbb{R}^n, \mathbb{R}^n)$, where I_d is an identity map. Suppose that such a space is denoted by $\mathcal{G}$. For $\mathbf{u} \in T_e\mathcal{G}$, a vector in the tangent space of $\mathcal{G}$ with identity $e \in \mathcal{G}$, its norm is defined by:

$$\|\mathbf{u}\|_t^2 := \langle \mathbf{u}, \mathbf{u} \rangle_t := \int_D \mathbf{L}\mathbf{u} \cdot \mathbf{u} \, d\mathbf{x}. \quad (2.1)$$

In this paper, we consider a special operator L that is defined as $L = (I_d - a^2 \Delta)^b$, $b \geq 1$, for which the associated Green's kernel of L is given by

$$G_{b-n/2}(|\mathbf{x}|) = \frac{2^{n/2-b}}{(2\pi a)^{n/2} a^b \Gamma(b)} |\mathbf{x}|^{b-n/2} K_{b-n/2} \left(\frac{|\mathbf{x}|}{a} \right), \quad (2.2)$$

where, $K_{b-n/2}$ is the modified Bessel function of the second kind of order $b-n/2$ and $\Gamma(b)$ is the usual notation for the Gamma function [34]. A notable special parametric choice is the two-dimensional ($n=2$) Green's function with $a=1$ and $b=3/2$, resulting in the simple form³

$$G_{1/2}(|\mathbf{x}|) = \frac{1}{2\pi} e^{-|\mathbf{x}|}. \quad (2.3)$$

This conic-shape kernel while technically does not satisfy the definition of diffeomorphism is advantageous for template matching [6]. This is the kernel we will use later in our numerical experiments.

The length of a curve $\phi : [0, 1] \rightarrow \mathcal{G}$ is defined by using the right translation

$$l(\phi) = \int_0^1 \|\mathbf{u}\|_t \, dt, \quad \text{for } \mathbf{u} = \frac{\partial \phi}{\partial t}(\phi^{-1}(\mathbf{x}, t), t). \quad (2.4)$$

Furthermore, an energy functional can be defined by

$$E(\phi) = \int_0^1 \langle \mathbf{u}, \mathbf{u} \rangle_t \, dt. \quad (2.5)$$

The geodesic is usually obtained by minimizing the energy functional E rather than the length l . This is because minimizing E

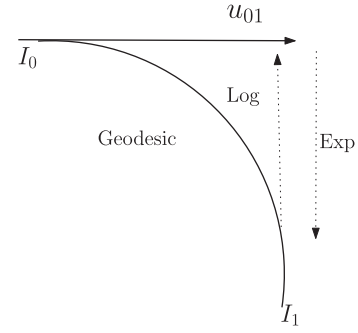


Fig. 1. A schematic for Riemann maps between the manifold and the tangent space. $\mathbf{u}_{01}$ serves as the “shape coordinate” for template I_1 based on I_0 . Through this map, the information on the nonlinear manifold is extracted and mapped to the space that is tangent to the manifold at I_0 . The tangent space is linear, on which statistical analysis can easily be applied.

chooses not only the best path but also the simplest parametrization of time in ϕ that satisfies the constraints $\phi(\mathbf{x}, 0) = I_0$ and $\phi(\mathbf{x}, 1) = I_1$, where I_0 is referred to as the reference, whereas I_1 is the target [34]. In this case, we have $l(\phi) = E(\tilde{\phi}) = l(\tilde{\phi})^2$, where $\tilde{\phi} = \phi(\mathbf{x}, f(t))$ with f being an invertible increasing function (a reparametrization of time) from $[0, 1]$ to itself. Note that we can assume the existence and the uniqueness of such a geodesic in our problems, as long as the manifold we consider is compact and we only work on some neighborhood of $e \in \mathcal{G}$.

The Euler-Lagrange equation, arising from minimizing the energy functional, turns out to have exactly the same form as the EP equations [31,34]:

$$\mathbf{m}_t + (\mathbf{u} \cdot \nabla) \mathbf{m} + (\nabla \cdot \mathbf{u}) \mathbf{m} + (\nabla \mathbf{u}^t) \mathbf{m} = 0, \quad (2.6)$$

with $\mathbf{m} = L\mathbf{u}$, subject to the “time” boundary conditions $\phi(\mathbf{x}, 0) = I_0$ and $\phi(\mathbf{x}, 1) = I_1$. Assuming that the space is geodesic complete, we can convert this boundary-value problem into finding the corresponding initial condition for an initial-value problem. Let the corresponding initial condition $\mathbf{u}_{01}$ be defined through a proper logarithmic map from M to $T_{I_0}M$, the tangent space of M at I_0 , by:

$$\mathbf{u}_{01} = \text{Log}_{I_0}(I_1). \quad (2.7)$$

The corresponding exponential map from $T_{I_0}M$ to M then can be defined accordingly by

$$I_1 = \text{Exp}_{I_0}(\mathbf{u}_{01}). \quad (2.8)$$

The Exp map is an isometry between $T_{I_0}M$ and M to the first order [34], here we take the results from finite-dimensional Riemannian geometry since this paper will only deal with finite landmarks:

$$\begin{aligned} d(I_i, I_j)^2 &= \|\mathbf{u}_{0i} - \mathbf{u}_{0j}\|^2 - \frac{1}{3} \|\mathbf{u}_{0i} \wedge \mathbf{u}_{0j}\|_R^2 \\ &\quad + o(\max(\|\mathbf{u}_{0i}\|, \|\mathbf{u}_{0j}\|)^4), \end{aligned} \quad (2.9)$$

where $\|\cdot\|_R$ is the norm of curvature tensor and $I_i, I_j \in M$. By letting $i=0, j=1$, $\mathbf{u}_{00} = \mathbf{0}$, the term $\frac{1}{3} \|\mathbf{u}_{0i} \wedge \mathbf{u}_{0j}\|_R^2 = 0$. Combining all the previous results, we will have

$$l(\phi) = E(\tilde{\phi}) = l(\tilde{\phi})^2 = d(I_0, I_1)^2 = \|\mathbf{u}_{01}\|^2 + o(\|\mathbf{u}_{01}\|^4). \quad (2.10)$$

Thus, numerically, the length of tangent vector can be used to estimate the length of the geodesic. This first-order isometry is illustrated in Fig. 1.

Under the setting of landmark representation, a diffeomorphism ϕ can be represented by its actions on the landmark space L^N , i.e.

$$\phi : L^N \rightarrow L^N, \quad I := \{\mathbf{x}_i\}_{i=1}^N \rightarrow \phi(I) := \{\phi(\mathbf{x}_i)\}_{i=1}^N. \quad (2.11)$$

³ In this case the space is $I_d + C_0^\infty(\mathbb{R}^n, \mathbb{R}^n)$

This gives a finite dimensional approximation of the original infinite functional space⁴. In the landmark space L^N , it is convenient to define a pair of variables $(\mathbf{q}_i, \mathbf{p}_i)$ for each landmark $\mathbf{x}_i = \mathbf{q}_i$ and the corresponding momentum $\mathbf{p}_i$. For given initial data $\{(\mathbf{q}_i, \mathbf{p}_i)\}_{i=1}^N$, the time evolution of the Euler-Lagrangian (or the EP) equations is obtained by evolving the particle system in time for $(\mathbf{q}_i, \mathbf{p}_i)$ [5]:

$$\begin{aligned}\frac{d\mathbf{q}_i}{dt} &= \sum_{j=1}^N G_{b-1}(|\mathbf{q}_i - \mathbf{q}_j|) \mathbf{p}_j, \\ \frac{d\mathbf{p}_i}{dt} &= - \sum_{j=1, j \neq i}^N (\mathbf{p}_i \cdot \mathbf{p}_j) G'_{b-1}(|\mathbf{q}_i - \mathbf{q}_j|) \frac{\mathbf{q}_i - \mathbf{q}_j}{|\mathbf{q}_i - \mathbf{q}_j|}.\end{aligned}\quad (2.12)$$

The above particle system gives an easy way of calculating the tangent vector, which is the velocity field $\mathbf{u}(\mathbf{x}, t)$ of the EP equations,

$$\mathbf{u}(\mathbf{x}, t) = \sum_{j=1}^N G_{b-1}(|\mathbf{x} - \mathbf{q}_j|) \mathbf{p}_j, \quad (2.13)$$

and the discrete Hamiltonian is defined by

$$H = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\mathbf{p}_i \cdot \mathbf{p}_j) G_{b-1}(\mathbf{q}_i - \mathbf{q}_j). \quad (2.14)$$

The Hamiltonian is a conserved quantity and is a metric. It is worth noting that numerically solving the particle system (2.12) is equivalent to approximating an Exp map. Moreover, with this particle setting, on the tangent space, numerically it is advantageous for us to directly work on the momentum field $\mathbf{p}$, other than the velocity $\mathbf{u}$ for the Log map. Detailed discussion about the advantage of using the momentum field is in Section 3. Note that the conversion between the velocity and the momentum fields is straightforward by using Eq. (2.13).

For our applications, landmark positions $\{\mathbf{x}_i = \mathbf{q}_i\}_{i=1}^N$ at $t = 0$ and $t = 1$ are known, while the initial momenta $\{\mathbf{p}_i\}_{i=1}^N$ at $t = 0$ are not. A shooting algorithm that searches for the initial momenta is introduced in [6]. In that algorithm, a prediction-correction step:

$$\mathbf{u}_0^{k+1} = \mathbf{u}_0^k + h^k (\mathbf{I}^1 - \mathbf{I}_1^k), \quad (2.15)$$

$$\mathbf{p}_0^{k+1} = \text{Convert}(\mathbf{u}_0^{k+1}; \mathbf{I}_0), \quad (2.16)$$

where

$$\mathbf{u}_0^k : \text{the } k\text{th iteration of the velocity field at } t = 0 \text{ for } \mathbf{I}_0, \quad (2.17)$$

$$\mathbf{I}_1^k : \text{the } k\text{th iteration of } \mathbf{I}_1 \text{ obtained by solving the particle system,} \quad (2.18)$$

$$h : \text{the correction parameter at } k\text{th iteration,} \quad (2.19)$$

is coupled with the particle system (2.12) to search for the appropriate initial conditions $\mathbf{p}_i|_{t=0}$, $i = 1, \dots, N$, so that $\text{Exp}(\mathbf{I}_0; \mathbf{p}_i|_{t=0}) = \mathbf{I}_1$. This shooting algorithm numerically generates the Riemann Log map. The algorithm is described in Algorithm 1, with a schematic illustration in Fig. 2, and more details and discussion in [6].

3. Group average via a geodesic shooting algorithm

In the literature, there are many classical approaches for obtaining an average of a group [3,4,25,34,37]. The basic idea of these approaches is to calculate the Karcher mean or Fréchet mean with

Algorithm 1 Geodesic shooting algorithm for template matching.

Select corresponding landmarks (particles) between two shapes I_0 and I_1 ;
Initialize the vector field $\mathbf{u}_0^{(0)}$ associated with the landmark points of I_0 with an initial guess of momenta $\mathbf{p}^{(0)}$;
for $k = 0, 1, 2, \dots$ **until** convergence **do**

1. Evolve the N -particle system of EP Eq. (2.2) to some fixed time ($t = 1$), using the momenta associated with $\mathbf{u}_0^{(k)}$ to obtain the approximate landmarks $\mathbf{q}_i^{(k)}$ for $\mathbf{I}_1^{(k)}$, where $\mathbf{q}_i$ is the i^{th} landmark, $i = 1, \dots, N$.
2. If the difference $\|\mathbf{I}_1 - \mathbf{I}_1^{(k)}\| < \epsilon$ for some norm, then stop.
3. The vector field is dynamically updated by $\mathbf{u}_0^{(k+1)} = \mathbf{u}_0^{(k)} + M^{(k)} \cdot (\mathbf{I}_1 - \mathbf{I}_1^{(k)})$, where $M^{(k)}$ is an unknown matrix that can be chosen so that the algorithm converges at least locally.
4. Update the momenta of the particle system $\mathbf{p}_0^{(k+1)}$ from $\mathbf{u}_0^{(k+1)}$.

end for

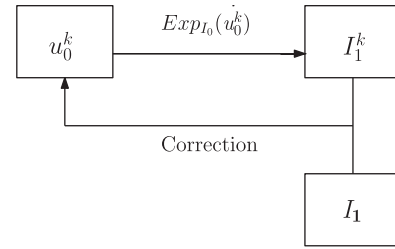


Fig. 2. A schematic illustration for the shooting algorithm. In the k^{th} iteration before convergence, the particle system is used to calculate the other end of the geodesic by adopting $(I_0, \mathbf{u}_0^k)$ as the initial data. A miss-fit function between $\mathbf{I}_1^k$ and the target $\mathbf{I}_1$ is calculated to adjust $\mathbf{u}_0^k$ for the next approximation $\mathbf{u}_0^{k+1}$.

different choice of metrics [12,23]. Recently, more sophisticated estimators related to tangent spaces of Riemannian manifolds have been studied for their potential to be used as averages of groups [13,36,40,41]. Among all these approaches, the classical morphometric average, like the Procrustes average, in a sense, ignores the nonlinear and non-flat geometric features. On the other hand, to consider such features in finding an average under the setting of diffeomorphic deformation, it is necessary to find a nice linear space (at least locally), on which performing linear combinations is both mathematically sound and physically meaningful. To this end, we choose the momentum field as such a space for our proposed algorithm (described below), because the momentum provides a local template-based coordinate system which is also linear in nature [21,36,40]. As a result, in our algorithm, when we find an average of a group in the momentum field and map the result through the Exp map back to the manifold, in principle, certain information of the non-flat geometry will still be preserved. At convergence, the initial momentum that drives the average to each individual observation will serve as the momentum representation of that landmark template to be used for later statistical analysis.

Let $\bar{I}$ denote the average of the group $T := \{I_k\}_{k=1}^m$. We use Fig. 3 to illustrate the proposed averaging procedure, with details given in Algorithm 2.

Note that in Algorithm 2, w_j in Eq. (3.3) is called the j th weight. Two special weighting schemes are worth noting. In the first scheme, the weight is chosen to be

$$w_i = 1/m \quad \text{for all } i\text{'s.} \quad (3.6)$$

This weighting scheme mimics the direct average method for the Karcher mean in flat Eulerian spaces, which gives the solution of

² Readers are referred to [15,16] for methods dealing with non-labeled landmarks.

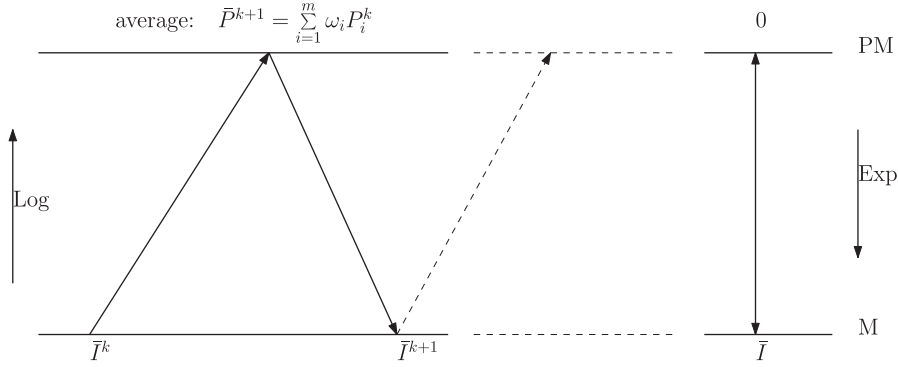


Fig. 3. A schematic of the averaging scheme where M stands for the manifold, PM denotes the momentum bundle. As mentioned in the previous section, in the landmark setting, the Log mapping is calculated using a prediction-correction loop, while the Exp mapping can be done by solving the particle system [6]. Once convergence is reached, P_i will be used as residual momentum representation for each individual observation in the group.

Algorithm 2 Algorithm for finding an average of a group.

Choose an initial guess $\bar{I}^0 \notin T$ for the average;

for $k = 0, 1, 2, \dots$ **until convergence do**

for $i = 0, 1, 2, \dots, m$ **do** a) Extract tangent vector through Log mapping:

$$\mathbf{u}_i^k = \text{Log}_{\bar{I}^k}(I_i) \quad (3.1)$$

 b) Convert each tangent vector to momentum:

$$\mathbf{p}_i^k = \text{Convert}(\mathbf{u}_i^k) \quad (3.2)$$

end for

 Take the weighted average on the momentum field via:

$$\bar{\mathbf{p}}^{k+1} = \sum_{i=1}^m \omega_i \mathbf{p}_i^k \quad (3.3)$$

if $\|\bar{\mathbf{p}}^{k+1} - \bar{\mathbf{p}}^k\| > \epsilon$ **then**
 Compute the next guess:

$$\bar{\mathbf{u}}^{k+1} = \text{Convert}(\bar{\mathbf{p}}^{k+1}) \quad (3.4)$$

$$\bar{I}^{k+1} = \text{Exp}_{\bar{I}^k}(\bar{\mathbf{u}}^{k+1}) \quad (3.5)$$

else

Break;

Return $\bar{I}^k$ as average $\bar{I}$.

end if

end for

the least sum of distances squared, i.e.

$$\sum_{i=1}^m d(\bar{I}, I_i)^2. \quad (3.7)$$

Note that for this weight, each member has an equal opportunity for contributing to the group average, which makes this scheme equivalent to that introduced in [36]; the latter reference includes a thorough discussion of the so-called “center of mass” in Riemannian manifolds under a statistical setting.

The second weighting scheme is motivated by the idea of finding the geometric median [13], a robust estimator when the dataset has corrupted members. Using the idea from the Weiszfeld’s algorithm for Eulerian spaces [24], the assigned weight for this scheme is calculated by

$$w_i = \frac{1}{d_i} / \sum_{i=1}^m \frac{1}{d_i}. \quad (3.8)$$

For this choice of weight, the estimator is the value that minimizes the sum of distances, denoted by

$$D(\bar{I}, T) := \sum_{i=1}^m d(\phi_i), \quad (3.9)$$

where ϕ_i is the geodesic connecting the unknown group average $\bar{I}$ and each group member I_i , i.e. the restriction is now $\phi_i(\mathbf{x}, 0) = \bar{I}$ and $\phi_i(\mathbf{x}, 1) = I_i$ for each i .

The existence and uniqueness of both “mean” ($\omega_i = \frac{1}{m}$) and “median” ($w_i = \frac{1}{d_i} / \sum_{i=1}^m \frac{1}{d_i}$) were investigated in [13,36], respectively with assumptions on the manifold that either one of the following holds:

- (a) the sectional curvature $\kappa < 0$;
- (b) $0 < \kappa < K$, and $\text{diam}(U) < \frac{\pi}{2\sqrt{K}}$ with U a subset of the manifold that can cover all observations.

Following the idea of Kuhn and others [13,24], we now give a proof that shows the convergence of Algorithm 2 for both choices of weights, (3.6) and (3.8) under condition (b).

Theorem 1 (Local Convergence of Algorithm 2).

1. Algorithm 2 converges to the minimizer of (3.9) for $\omega_i = \frac{1}{m}$.
2. Algorithm 2 converges to the minimizer of (3.7) for $\omega_i = \frac{1}{m}$.

Proof. If the condition (b) is satisfied, there exists a unique minimizer on the manifold [13,36]. We only need to show the introduced algorithm actually converges under this condition.

We first show claim (1). Let $f(\mathbf{x}) = \sum_{i=1}^m d(\mathbf{x}, \mathbf{x}_i)$, where $\mathbf{x}$ and $\{\mathbf{x}_i\}_{i=1}^m \in M$ and M is a manifold. Let $\mathbf{x}, \mathbf{y} \in M$. The distance $d(\mathbf{x}, \mathbf{y})$ is defined by $\|\text{Log}_{\mathbf{x}}(\mathbf{y})\|_{L, \mathbf{x}} = \|\text{Log}_{\mathbf{y}}(\mathbf{x})\|_{L, \mathbf{y}}$, with $\|\cdot\|_{L, \mathbf{x}}$ the norm induced by operator L .

Since f is a geodesic convex function, there will be no local optimum on U . In order to show that the algorithm converges, we only need to prove $f(\hat{\mathbf{x}}^{k+1}) < f(\hat{\mathbf{x}}^k)$ if $f(\hat{\mathbf{x}}^k)$ is not the minimum.

Let $\hat{\mathbf{x}}^{k+1}$ be the $k+1$ iteration of Algorithm 2, with the following defined as:

$$\mathbf{u}_i^{k+1} = \text{Log}_{\hat{\mathbf{x}}^{k+1}}(\mathbf{x}_i), \quad \hat{\mathbf{x}}^{k+1} = \text{Exp}_{\hat{\mathbf{x}}^k}(\mathbf{u}^{k+1}) \text{ and } \mathbf{x}_i = \text{Exp}_{\hat{\mathbf{x}}^k}(\mathbf{u}_i^k). \quad (3.10)$$

Hence,

$$f(\hat{\mathbf{x}}^{k+1}) = \sum_{i=1}^m d(\hat{\mathbf{x}}^{k+1}, \mathbf{x}_i) \quad (3.11)$$

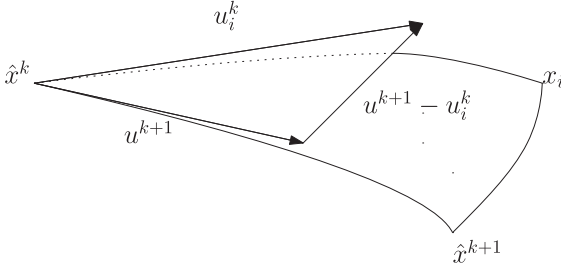


Fig. 4. Triangles on M and $T_x^k M$.

$$= \sum_{i=1}^m d(\text{Exp}_{\hat{x}^k}(\mathbf{u}^{k+1}), \text{Exp}_{\hat{x}^k}(\mathbf{u}_i^k)) \quad (3.12)$$

$$\leq \sum_{i=1}^m \|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_{L, \hat{x}^k}. \quad (3.13)$$

It is worth noting that the last inequality holds, only if all observations $\mathbf{x}_i \in U$ and the condition (b) is satisfied. In this case, we can use Toponogov Comparison Theorem to obtain the inequality.

Let $\Delta_{\hat{x}^k, \hat{x}^{k+1}, \mathbf{x}_i}$ be the geodesic triangle on the manifold, see Fig. 4. The Toponogov Comparison Theorem states that if observations $\mathbf{x}_i$ satisfy the condition (b), for each corresponding side of the two triangles, the distance on the M is less than the corresponding distance on $T_{\hat{x}^k} M$, i.e. $d(\hat{x}^{k+1}, \mathbf{x}_i) \leq \|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_{L, \hat{x}^k}$.

Hereafter, for a fixed $\hat{x}^k$, we drop $\hat{x}^k$ in $\|\cdot\|_{L, \hat{x}^k}$ and denote the norm by $\|\cdot\|_L$ for simplicity. Suppose that the algorithm uses the update $\mathbf{u}^{k+1} = \sum_{i=1}^m \omega_i \mathbf{u}_i^k$, where $\omega_i = \frac{1}{\|\mathbf{u}_i^k\|_L} / \sum_{i=1}^m \frac{1}{\|\mathbf{u}_i^k\|_L}$, then $\mathbf{u}^{k+1}$ is the center of gravity of weights $\frac{1}{\|\mathbf{u}_i^k\|_L}$ placed at $\mathbf{u}_i^k$, and hence $\mathbf{u}^{k+1}$ minimizes the function⁵:

$$g(\mathbf{u}) = \sum_{i=1}^m \frac{1}{\|\mathbf{u}_i^k\|_L} \|\mathbf{u} - \mathbf{u}_i^k\|_L^2. \quad (3.14)$$

$$\text{Thus } g(\mathbf{u}^{k+1}) \leq g(\mathbf{0}) = \sum_{i=1}^m \|\mathbf{u}_i^k\|_L = f(\hat{x}^k).$$

On the other hand,

$$g(\mathbf{u}^{k+1}) = \sum_{i=1}^m \frac{1}{\|\mathbf{u}_i^k\|_L} [(\|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L - \|\mathbf{u}_i^k\|_L) + \|\mathbf{u}_i^k\|_L]^2 \quad (3.15)$$

$$= \sum_{i=1}^m \frac{1}{\|\mathbf{u}_i^k\|_L} (\|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L - \|\mathbf{u}_i^k\|_L)^2 + 2 \sum_{i=1}^m \|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L - f(\hat{x}^k). \quad (3.16)$$

Substituting this into $g(\mathbf{u}^{k+1}) \leq f(\hat{x}^k)$, we get

$$\sum_{i=1}^m \frac{1}{\|\mathbf{u}_i^k\|_L} (\|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L - \|\mathbf{u}_i^k\|_L)^2 + 2 \sum_{i=1}^m \|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L \leq 2f(\hat{x}^k). \quad (3.17)$$

Hence, we have

$$\sum_{i=1}^m \|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L < f(\hat{x}^k). \quad (3.18)$$

Now combining Eqs. (3.13) and (3.18) yields

$$f(\hat{x}^{k+1}) \leq \sum_{i=1}^m \|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L < f(\hat{x}^k). \quad (3.19)$$

Note that the equalities in (3.19) hold if and only if $\mathbf{u}^{k+1} = \mathbf{0}$, which is equivalent to $\hat{x}^{k+1} = \hat{x}^k = \bar{x}$, i.e. the iteration stops and the minimum is reached. This proves that Algorithm 2 converges when the weight (3.8) is used. Moreover, at convergence, the result of Algorithm 2 minimizes the sum of the distances.

Following the above structure, and using the same notation, we can prove claim (2). Below, we repeat the proof but skip some steps to avoid redundancy. Redefining $f(\hat{x}) = \sum_{i=1}^m d(\mathbf{x}, \mathbf{x}_i)^2$, we have

$$f(\hat{x}^{k+1}) \leq \sum_{i=1}^m \|\mathbf{u}^{k+1} - \mathbf{u}_i^k\|_L^2. \text{ Since } \omega_i = \frac{1}{m}, \text{ the function } g \text{ is now defined by}$$

$$g(\mathbf{u}) = \sum_{i=1}^m \|\mathbf{u} - \mathbf{u}_i^k\|_L^2 = f(\mathbf{u}). \text{ So by the same argument, we have}$$

$$f(\hat{x}^{k+1}) \leq g(\hat{x}^{k+1}) \leq g(\mathbf{0}) = \sum_{i=1}^m \|\mathbf{u}_i^k\|_L^2 = f(\hat{x}^k).$$

The equality holds only if $\hat{x}^{k+1} = \hat{x}^k$. This proves the convergence of Algorithm 2 with equal weights, and the result gives the least sum of distances squared. $\square$

We remark that similar to Weizfeld's algorithm for flat Eulerian spaces, the initial guess for Algorithm 2 cannot be chosen as any of the group members l_i , otherwise the algorithm will not converge. However, there is a way to get around with this restriction in flat Eulerian spaces [41], so that the algorithm converges for any initial guess. The extension of it on general Riemannian manifolds is still under investigation. We also remark that Pennec [36] investigated other geometric measurements in a statistical setting for the equally weighted schemes, see [36] for further details.

The two weighting schemes give rise to different average templates for a group, based on which, the resulting residual momentum representation for each group member will be different. We remark that the result of Theorem 1 relies heavily on condition (b) (see page 10) which may not hold for every dataset, as discussed in [28]. Nevertheless, in this paper, we intend to use the algorithm as a feature extraction tool, a pre-processor to extract momentum features for the purpose of detecting local changes that are statistically significant. The data we use are often recorded on the same or similar objects, such as the annotated brain landmarks of different people or of the same people at different times. In these cases, observations are assumed "close", and hence convergence can be expected. In other cases, the algorithms may not converge⁶, or the minimizer is not unique.⁷ In such cases, we will use a smaller scale parameter a in Green's kernel⁸ $\frac{1}{2\pi} e^{-\frac{\|\mathbf{x}\|}{a^2}}$, so that the matrix $\{G_{ij}\}$

with entry $G_{ij} = \frac{1}{2\pi} e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{a^2}}$ will become more diagonally dominated when used for calculating the geodesic. Landmarks in this case will become less correlated with each other and the deformation will look more linear. By doing so, the geometry of the landmark deformation may become closer to the flat case, and we can hope that the diameter inequality in condition (b) could then be satisfied and the proposed algorithms can at least converge numerically.

⁴ For example, curvatures are negative.

⁵ For example, curvatures are positive, but diameter inequality is not satisfied for observations.

⁶ All the following examples are conducted using $a^2 = 1$ unless specifically specified otherwise.

³ Since $\|\mathbf{u}_i^k\|_L$ is the denominator, observations $\mathbf{x}_i$ cannot be selected as an initial guess for starting the algorithm.

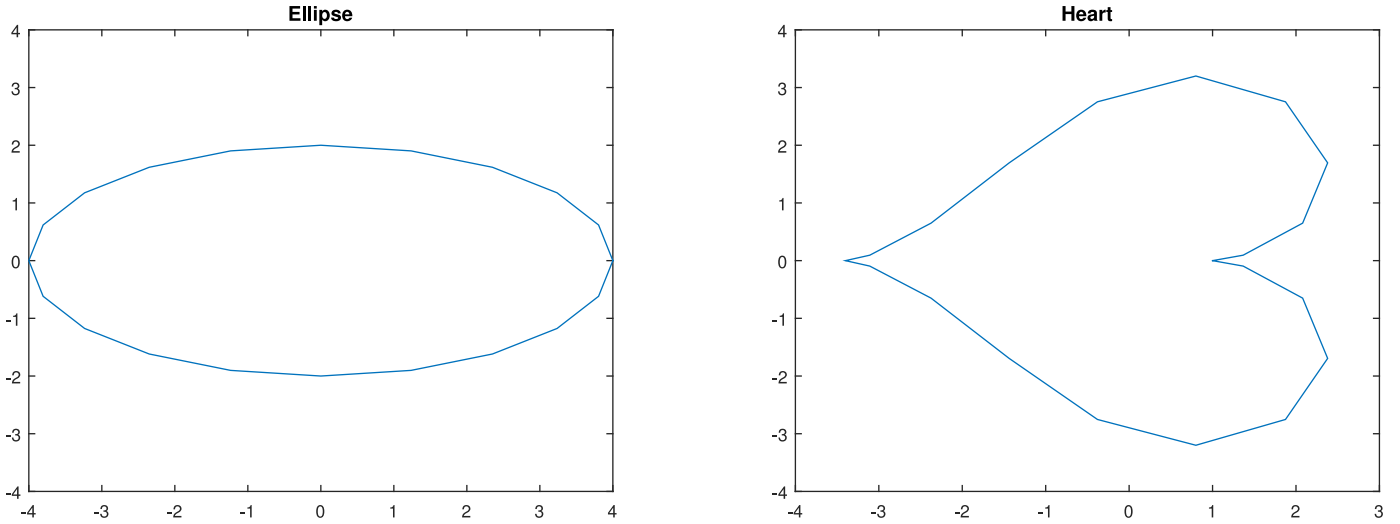


Fig. 5. A typical ellipse and a typical outlier in the group we consider.

3.1. Example 1

In this example, we calculate the average for a group of 20 shapes by using the two different weighting schemes, respectively. 20 landmarks are used to represent each shape. In the group we consider, a portion of the group ($1 - \alpha$, where $0 < \alpha < 0.5$) are ellipses $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$, where (a, b) have a bivariate normal distribution: $\mathcal{N}((4, 2), 0.2I)$, and I is the identity matrix. The rest of this group are treated as outliers produced from copies of the heart-shape curve:

$$\begin{cases} x = \frac{1}{5}(13 \cos(\theta) - 5 \cos(2\theta) - 2 \cos(3\theta) - \cos(4\theta)) \\ y = \frac{1}{5}(16 \sin(\theta)^3). \end{cases} \quad (3.20)$$

Fig. 5 shows a typical ellipse and a typical outlier in this group. We show the comparison between the two weighting schemes in Fig. 6, where α is chosen to be 0.1, 0.2, 0.3, and 0.4 for different levels of outliers. It is clear from the figure that the equally weighted scheme absorbs the traits equally from each member of the group; namely when the level of outliers is higher (larger α), the shape of the resulting average deviates more from the shapes of the majority, which are a distribution of ellipses. On the contrary, the other weighting scheme gives a robust estimator, in the sense that the obtained average stays close to the majority of the group, in particular when α is large.

We remark that the use of the weighting schemes depends on the goals of the application. For example, the equally weighted scheme is sensitive to outliers, which can be advantageous if the goal is to detect outliers in streaming data. On the other hand, if a most representative average is the aim of the application, then the robust algorithm should be used to obtain a robust estimator.

3.2. Example 2

In the second example, we apply our algorithms to outlines of four different fighters annotated by landmarks. The data were analyzed in [38] and can be downloaded at [http://visionlab.uta.edu/shape_data.htm]. The four fighters are Mirage (Group 1), F14 with wings extracted (Group 2), F14 with wings extended (Group 3), and F15 (Group 4). Each fighter category has 30 outline templates. The original number of landmarks for each shape template roughly ranges from 300 to 1600. In order to conduct our algorithms effectively, we sample only 30 points for each template. Fig. 7 shows the averaged outlines of these 4 fighters with sampled

points, calculated by using Algorithm 2 with two different weighting schemes.

In Fig. 8, we compare the group averages computed by the two weighting schemes with the presence of outliers. The solid line is the group average of the templates selected entirely from Group 1. The dash line is the group average of templates selected from both Group 1 (75%) and Group 3 (25%). We refer to such a composition of templates as a dataset with 25% outliers. We plot the group average with outliers against that without outliers. Similar to the previous example, we found that the group average computed with the robust weighting scheme is less affected by the presence of outliers.

In real datasets, outliers may come from different sources. In order to examine how our algorithms perform in identifying outliers in these cases, we compile four different datasets as described below and perform a cluster analysis on them.

1. Dataset 1: 50 templates, 30 from Group 1 and 20 from Group 3.
2. Dataset 2: 50 templates, 30 from Group 1 and 20 from Group 2.
3. Dataset 3: 50 templates, 30 from Group 1, 10 from Group 3, and 10 from Group 4.
4. Dataset 4: 60 templates, 30 from Group 1, 10 from Group 2, 10 from Group 3, and 10 from Group 4.

In all the datasets, the bulk of the data are from Group 1 with outliers from different sources. The experiment is conducted with the following steps assuming the number of clusters are known:

- Step 1: Compute the group average by three methods: (a) The traditional Procrustes method [3,4], (b) Algorithm 2 with equal weights, and (c) Algorithm 2 with robust weights.
- Step 2: After the group average is obtained, we compute the distance (metric) between each template and the estimated group average. In particular, for the traditional Procrustes method we compute the L^2 metric, while for Algorithm 2, we compute the H metric by Eq. (2.14).
- Step 3: For each selected metric D , we normalize the distance from each data value to the average by:

$$D_i^{\text{normalized}} = \frac{D_i - D_{\min}}{D_{\max} - D_{\min}}, \quad (3.21)$$

where $D_{\min}$ is the minimum of the metrics found within each dataset, while $D_{\max}$ is the maximum. We plot the

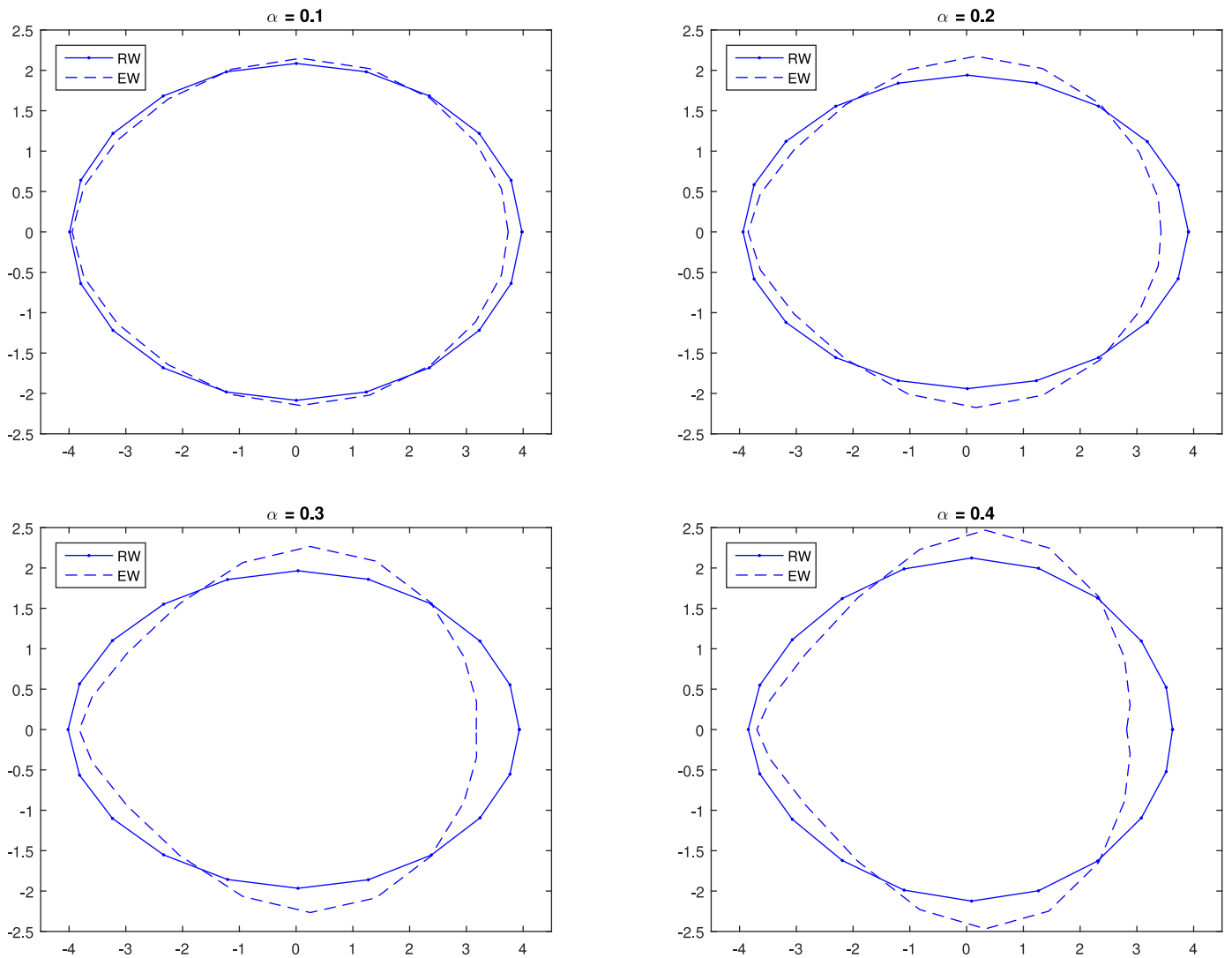


Fig. 6. Comparison of results from the two weighting schemes. “RW” stands for the robust weighting scheme, while “EW” is short for the scheme with equal weights. The EW scheme are more sensitive to the fluctuations of the group components.

magnitudes of the normalized metrics in the order of templates.

Fig. 9 shows the results of our experiments. In the Figure, the rows, from top to bottom, are the results of dataset 1, 2, 3, and 4, respectively, while the columns, from left to right, are L^2 metric computed between each member of the dataset and the group average computed by using the Procrustes approach, H metric computed between each member of the dataset and the group average obtained by using the equal weights, and H metric computed between each member of the dataset and the group average by using the robust weights. A visualization of these distances accompanied by silhouette plots of the cluster with Dataset 1 is shown in Fig. 10. Note that since a silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters, the cluster formed by outliers are more separated (the values in the silhouette plot are mostly one) from the major data if the robust average is used. We observe from the Figures:

1. The results indicate that Group 3 is more distinct from Group 1 than Group 2 and Group 4, while Group 2 is the closest to Group 1. This is consistent with the visualization in Fig. 7.
2. All three methods are able to classify the different groups of shapes. However, Algorithm 2 with the robust weights seems

Table 1

Comparison of the accuracy with k -means clustering for the three methods. The proposed algorithm with the robust weights has the highest successful rate to classify the groups (95%–100%), while the Procrustes method is in the middle (84%–98%), and the proposed algorithm with the equal weights is the worst (64%–96%).

	Procrustes	Equal weight	Robust weight
Dataset 1	49/50	48/50	50/50
Dataset 2	44/50	32/50	48/50
Dataset 3	42/50	39/50	49/50
Dataset 4	49/60	49/60	57/60

to better classify the “outliers”, compared with the other two methods, in terms of the visualization.

To further confirm that Algorithm 2 with the robust weights indeed outperforms the other two methods in our experiments, we use the extracted metric (3.21) as feature for a k -means clustering analysis. The clustering analysis is performed 5 times and the best overall accuracy is given in Table 1. From Table 1, we observe that using distances with respect to a robust average yields the best clustering, since the robust average best represents the majority of the data, helping better separate the outlier group. The average

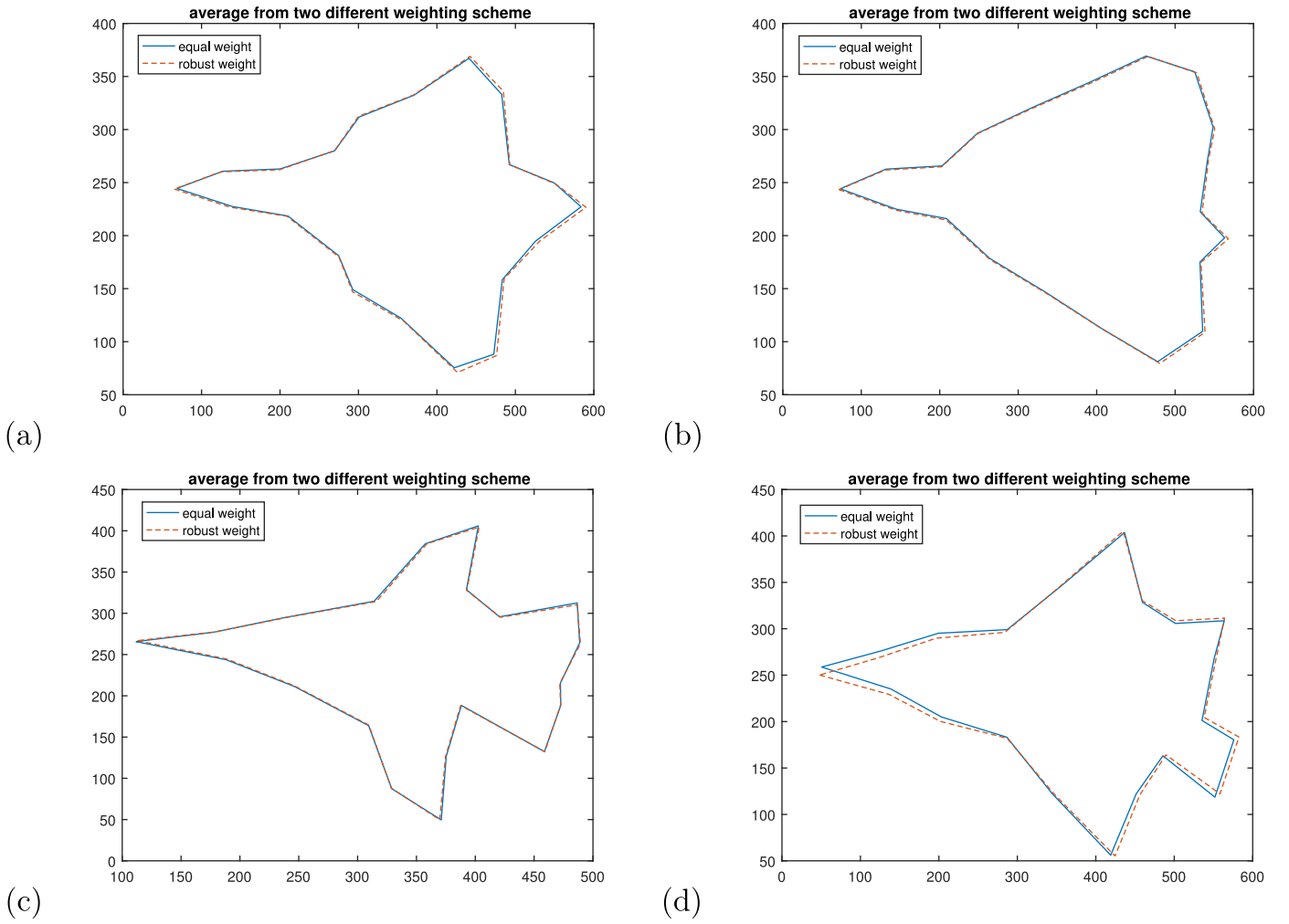


Fig. 7. The group averages for the outlines of four different fighter planes: (a) Mirage (Group 1), (b) F14 with wings extracted (Group 2), (c) F14 with wings extended (Group 3), and (d) F15 (Group 4).

with the equal weights performs the worst due to its nature that it is calculated assuming no obvious outliers are in the group.

Note 1. In the paper by Thakoor et al. [38], similar to our experiment in Table 1, the authors compared several state-of-the-art classification methods. The top three methods in their experiments are (1) circular Hidden Markov Model (HMM) + Maximum Likelihood, (2) Support Vector Machine (SVM) + Fourier descriptor, and (3) HMM + weighted likelihood classification (their proposed method). The classification precisions of the three methods for the 4 fighters selected in Table 1 are 96.7%–100%, 96.7%–100%, and 93.3%–100%, respectively. While their proposed method does not have the best classification precision, their method is less time-consuming. These classification methods are fundamentally different from the one used in our manuscript. Namely, they do not find the average template and the associated metric. Nonetheless, for our method with robust weights, the classification precision is 95%–100% for using a much smaller set of landmark points on shape perimeter (30 landmarks for the proposed method versus 256 landmarks for the methods in [38]).

Note 2. Procrustes method remains a popular method for finding an average shape in statistical shape analysis in the landmark space, due to its simplicity and efficiency. Many recent methods for finding an average template, such as those in [22,26], are not developed for the landmark space. It will not be a fair comparison

for comparing these methods with the proposed algorithm directly, unless we implement these methods in the landmark space, which may not be possible numerically and theoretically.

Note 3. Example 2 suggests that the proposed algorithms with the H metric do a moderately better job than the traditional Procrustes method, in terms of classification. In the next section, however, we will present a potential application of the group average. We will argue that the proposed algorithms provide extra information via the residual momenta that are useful for other classification purposes, whereas the Procrustes methods can only provide the information about the landmark locations.

4. An application of group average

Once an average for a group is defined and properly calculated, the next two important questions are

- Whether or not new data can be classified into a member of this group by comparing with the average?
- What features from the new data can be used to justify the decision made in (a)?

These two questions are part of an even bigger topic of feature extraction in pattern recognition and classification, namely finding features with which data from different categories can be effectively discriminated. To answer these two questions, we need to

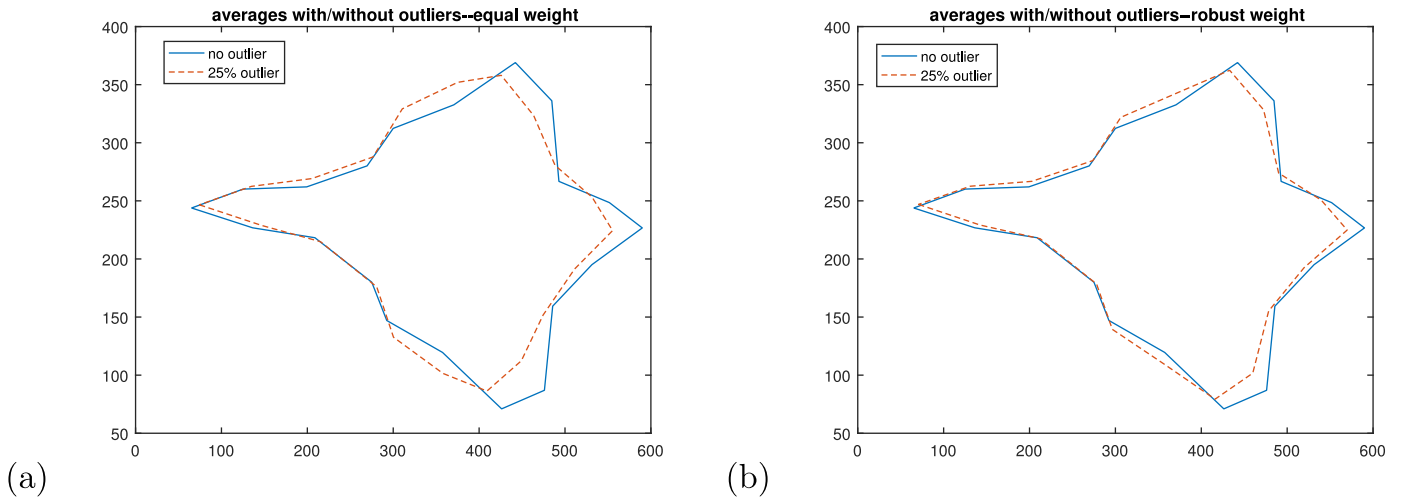


Fig. 8. Group averages in the presence of outliers. Solid lines: group averages of templates selected entirely from Group 1 in Fig. 7. Dash lines: group averages of templates selected from Group 1 (75%) and Group 3 (25%). (a) Group averages computed with equal weights. (b) Group averages computed with robust weights.

seek a set of meaningful variables that can carry necessary information about the nature of the data. The two variables we choose for the task are (i) a global variable H , which is the Hamiltonian (see Eq. (2.14)) that indicates the energy required for the deformation between two sets of landmarks [6], and (ii) a local vector variable $\mathbf{P}$, which is the collection of momentum at each landmark calculated by Algorithm 2.

The global variable H could be used to distinguish templates of different kinds, for example brains between humans and chimpanzees. In [6], we have utilized H for clustering and classifying different simple contour curves. However, it may not be sensitive enough to compare subcategories of the same kind, such as brain structures between people with and without schizophrenia where only minor local changes could cause a difference. To distinguish such a difference, we need to resort to the local variable. In this section, we will focus on applying the residual momentum $\mathbf{P}$ from Algorithm 2 for abnormality detection. A Bayesian predictive approach will be used to assess the information hidden behind the momentum representation $\mathbf{P}$. At the same time, we will discuss advantages of using the momentum representation over the traditional location representation.

In recent years, the local shape-based descriptions of brain regions that describe brain morphometry in regions involved with brain function disorders attracts much attention [42]. In particular, if the shape of brain structure is annotated by landmarks, then feature extraction methods that can detect the abnormality of anatomical landmarks has the potential for fast disease diagnosis [46]. With this in mind, to justify numerically the advantage of using momentum representation, we use a dataset of 28 parasagittal brain images for illustration. Each image is annotated by 13 landmarks. The dataset can be sorted into two groups. One contains 14 images of non-schizophrenic people, while the other has 14 images from confirmed schizophrenic cases. For more details about this dataset, we refer the reader to the link [19]. The annotated landmarks of a typical brain image in this dataset are shown in Fig. 11. In applications like this, observations on landmarks of different individuals are essentially close, so we can expect the convergence of Algorithm 2.

4.1. Detection

For each image in the dataset, the annotated landmarks represent locations of human brain structures. It is well known that landmark locations could be different between a schizophrenic and

a non-schizophrenic template. Such differences are sometimes beyond visualization and can only be identified through mathematical and statistical analysis. These landmarks carry certain features and information that can help distinguish schizophrenic images from a group of non-schizophrenic templates. We refer to the collection of these landmarks as a predictor. A predictor is viewed as a signature for describing the difference between templates. For the remainder of this section, we focus on finding the landmarks that comprise a predictor for our image data. The upshot of this exercise could help medical researchers locate those specific landmarks (structures of brain), and shed light on the development of the disease.

It is worth noting that in order to obtain convincing results, more than one method should be applied for validation. One of the traditional Procrustes methods uses superposition of datasets to compare the variations for each landmark. Adopting this approach, Bookstein [4] showed that landmarks 1 and 13 are the “most different” landmarks between the two groups in the dataset. For comparison, below we apply Algorithm 2 and use a Bayesian approach on momentum coordinates of each template to detect the abnormal landmarks.

Our algorithm begins with finding the momentum coordinates for each template in the database. Momentum coordinates $\{P_i^N\}_{i=1}^{14}$ for templates $\{I_i^N\}_{i=1}^{14}$ of the non-schizophrenic group will be calculated through Algorithm 2. For each i , P_i^N is a 13 by 2 matrix, in which the j th row ($j = 1, 2, \dots, 13$) contains the momentum coordinate calculated at the j th landmark. We denote the resulting average of this group by $\bar{I}_N$. After $\bar{I}_N$ is found, momentum representations based on the average calculated from the previous step for each individual in the schizophrenic group, denoted $\{P_i^S\}_{i=1}^{14}$, will be computed by the Riemannian logarithm map, i.e. $P_i^S = \text{Log}_{\bar{I}_N}(I_i^S)$. To apply statistical methods for the momentum coordinates, we construct the observation data matrix for each landmark. For each landmark $j = 1, 2, \dots, 13$, we collect the j th row of the momentum coordinates $\{P_i^N\}_{i=1}^{14}$ of the non-schizophrenic group to form a 14×2 matrix N_j . Similarly, we construct the data matrix S_j for the schizophrenic group.

Note 4. We note that for the k th image in each group, e.g. the non-schizophrenic group, the momentum coordinators $N_1^k, \dots, N_{13}^k$ are correlated, whereas N_1^k is independent of N_1^l , the momentum coordinate of first landmark of the l th image in this group. Our statistics in the next section are based on the independent momentum coor-

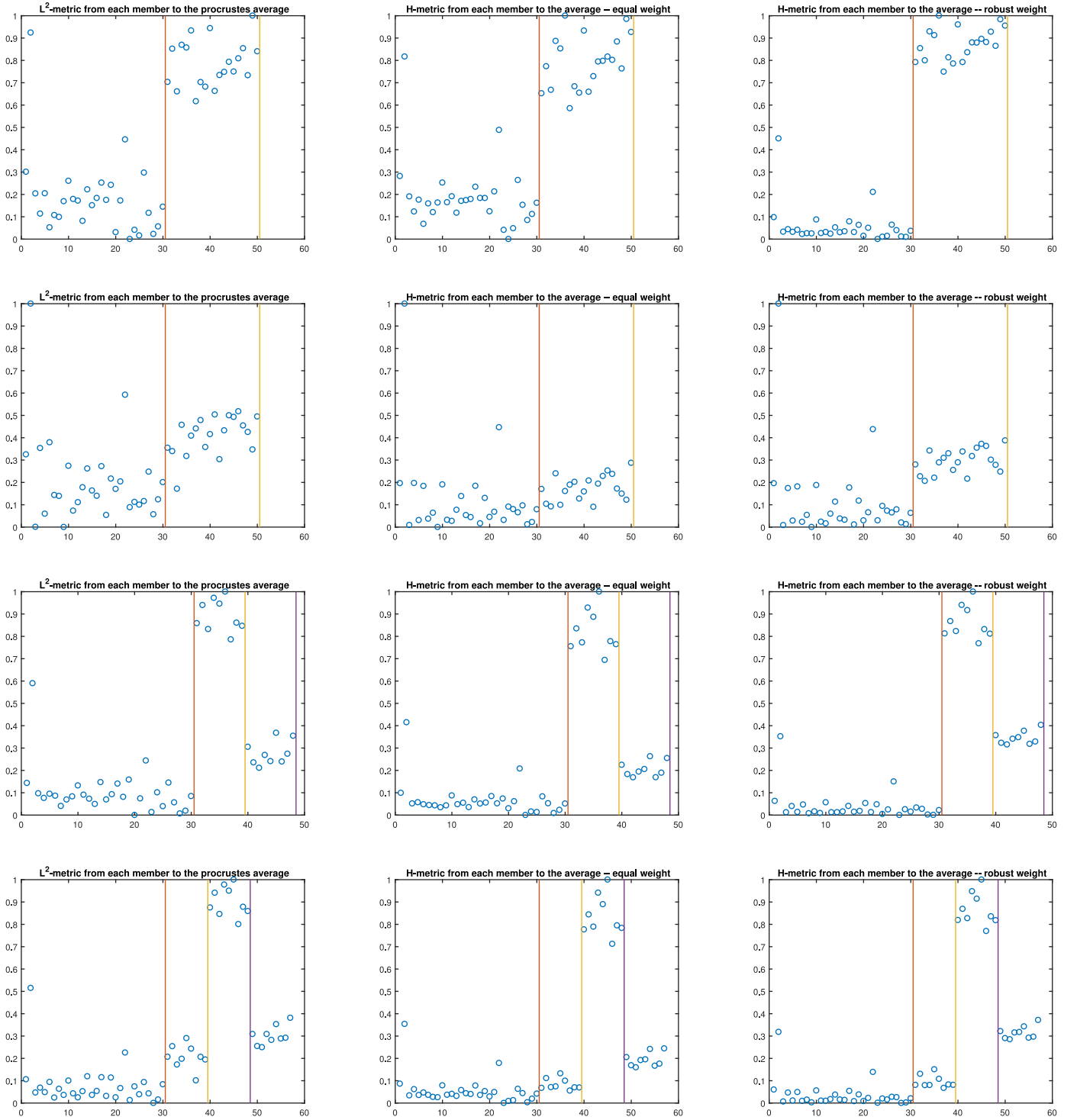


Fig. 9. Rows (from top to bottom): dataset 1, 2, 3, and 4. Columns (from left to right): L^2 metric computed between each member of the dataset and the group average computed by using the Procrustes approach, H metric computed between each member of the dataset and the group average obtained by using the equal weights, and H metric computed between each member of the dataset and the group average by using the robust weights.

dinates computed, respectively, from the 14 different images with respect to the average within the two groups.

4.2. Statistical methods for momentum data

In this section, we use two statistical methods to process the momentum data $\{N_j, S_j\}$ for landmarks $j = 1, 2, \dots, 13$. Our aim is to identify those landmarks associated with changes between

non-schizophrenia group and the compared schizophrenia group through analysis on their residual momentum representation. If both methods give consistent conclusions that are also consistent with the classical Procrustes method, the result by our method will be trustworthy in this task. We do not use classical hypothesis testing since our sample sizes are relatively small and most of the classical results rely on large sample asymptotics. Thus, we use a simple distance-based approach and also construct a Bayesian pre-

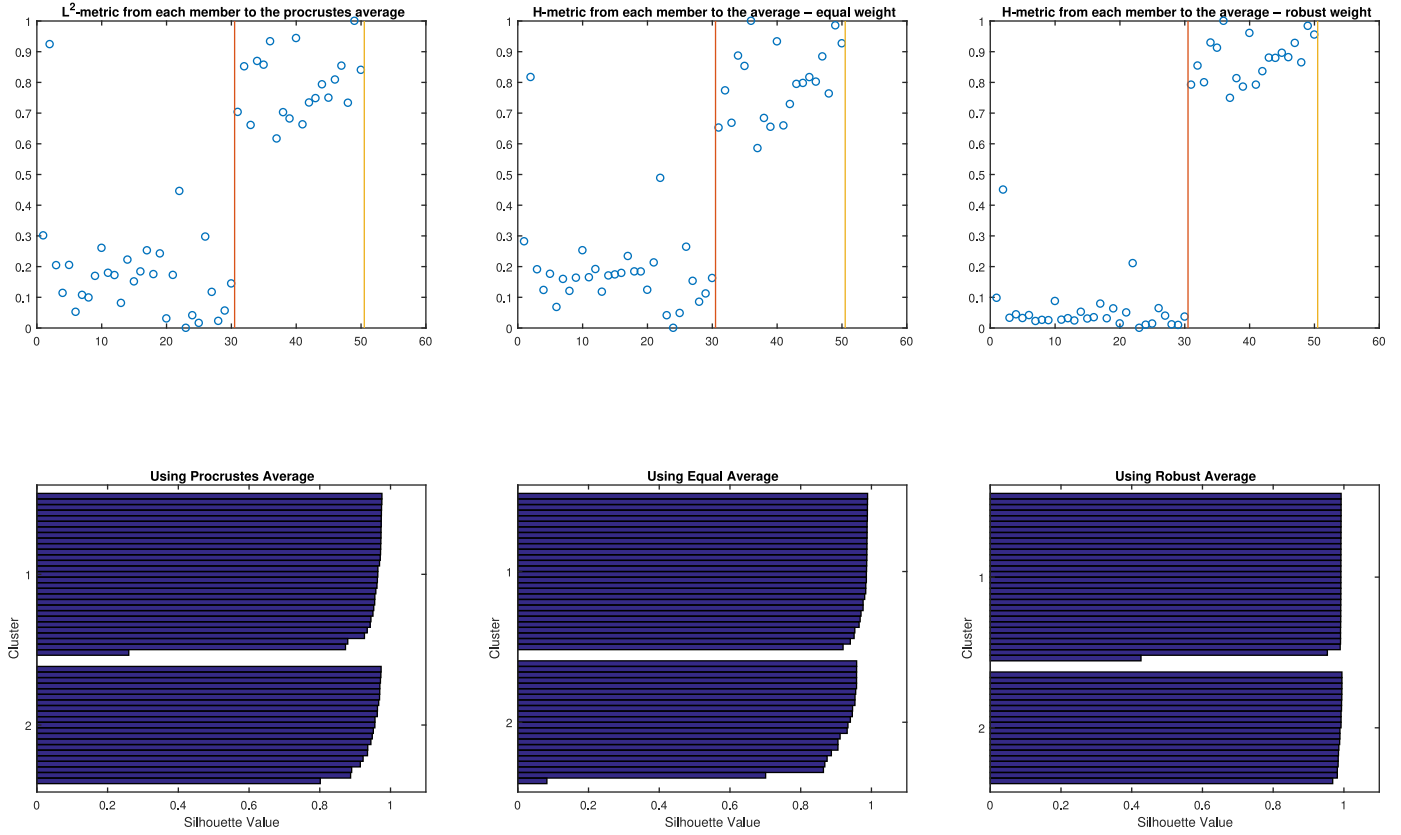


Fig. 10. Top row: Row 1 in Fig. 9. Bottom row: Silhouette plots of the clusters detected by the three methods in the top row. The cluster formed by outliers are more separated (i.e. the values in the silhouette plot are mostly one) from the major data if the robust average is used.

Table 2

l_2 norm of the mean momentum at each landmark in schizophrenic group. Landmark 1, 6 and 13 have unusually large values.

Norm of Mean Momentum for N Group							
Landmark	1	2	3	4	5	6	7
$ \bar{E}(N_j) $	0.0001	9.39e-06	1.17e-05	9.07e-06	1.11e-05	0.0001	3.74e-05
Landmark	8	9	10	11	12	13	
$ \bar{E}(N_j) $	2.6610e-05	1.3453e-05	1.9547e-05	2.4342e-05	4.8575e-06	5.9788e-05	

dictive distribution which conditions on the data and does not rely on asymptotics.

4.2.1. Method 1(norm of mean momentum)

The first statistical method we use to analyze the momentum data is to directly compare the norm of mean momentum at each landmark. This method follows an intuitive idea that landmarks with large mean momentum calculated from the schizophrenic group may indicate abnormalities. In this method, for each N_j and S_j , $j = 1, \dots, 13$, we compute the l_2 norm of mean momentum by

$$||\bar{E}(N_j)|| = \left\| \frac{1}{14} \sum_{k=1}^{14} N_j^k \right\|_2. \quad (4.1)$$

Note that for each N_j^k , there are two components. The l_2 norm evaluates the square root of the sum of component squares. Tables 2 and 3 summarize the l_2 norm of the mean momenta calculated at each landmark for the two groups, respectively. For the schizophrenic group (see Table 3), we note that Landmarks 1, 6 and 13 have unusually large values compared to other landmarks. We group these landmarks together as a predictor for this method.

4.2.2. Method 2(Bayesian posterior predictive distributions)

While the method by observing the sample mean can help provide certain information on abnormal location of landmarks, using Bayesian (posterior) predictive distributions we can obtain distributions for the landmarks in each of the two groups. This offers the advantage that the inference is not on the means of landmarks, but on the distributions of these landmarks themselves. Landmarks are ‘observables’ while their means are parameters and ‘latent’; inference on observables can be validated with future data as more images are obtained. Predictive distributions of the momentum data for each landmark are obtained via a Gibbs sampler for sampling from the posterior predictive distributions; details of the statistical modeling and its implementation are in Appendix A.

For each landmark, the respective predictive distributions of the two groups are compared to decide whether there is a significant difference. The predictive distribution $z = p(x, y | \{x_i, y_i\}_{i=1}^{14})$ is a surface, which is hard for us to analyze the difference. Instead, we use the contour lines of $p(x, y | \{x_i, y_i\}_{i=1}^{14}) = 0.95$ for their comparison. Fig. 12 shows the plots of the 95% contour lines of the two groups for each of the 13 landmarks.

To quantify the difference between the two contour lines for each landmark, we compute the ratio of the overlapping area of the two contour regions over the union of the two regions. Let Ω_i^N

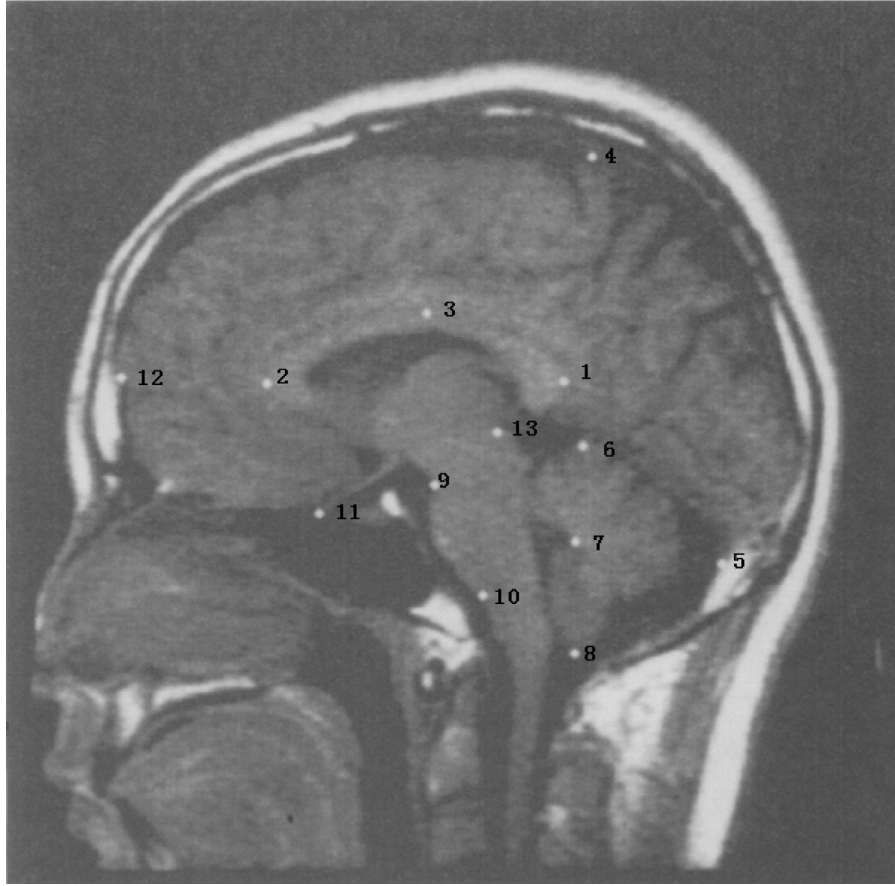


Fig. 11. Landmark Positions. (1) splenium, posteriormost point on corpus callosum; (2) genu, anteriormost point on corpus callosum; (3) top of corpus callosum, uppermost point on arch of callosum (all three to an approximate registration on the diameter of the callosum); (4) top of head, a point relaxed from a standard landmark along the apparent margin of the dura; (5) tentorium of cerebellum at dura; (6) top of cerebellum; (7) tip of fourth ventricle; (8) bottom of cerebellum; (9) top of pons, anterior margin; (10) bottom of pons, anterior margin; (11) optic chiasm; (12) frontal pole, extension of a line from 1 through 2 until it intersects the dura; (13) superior colliculus.

Table 3
 l_2 norm of the mean momentum at each landmark in non-schizophrenic group.

Norm of Mean Momentum for S Group							
Landmark	1	2	3	4	5	6	7
$\ \bar{E}(S_j)\ $	0.2231	0.0338	0.0533	0.0995	0.0932	0.1391	0.0432
Landmark	8	9	10	11	12	13	
$\ \bar{E}(S_j)\ $	0.0722	0.0635	0.0270	0.0826	0.0873	0.2182	

and Ω_i^S represent the contour region for each group, respectively, for the i^{th} landmark. We define the ratio as $r_i := \frac{|\Omega_i^N \cap \Omega_i^S|}{|\Omega_i^N \cup \Omega_i^S|}$. We expect that the smaller the value r_i is, the more different the two distributions will be. In practice, we can approximate r_i by $\hat{r}_i$ to avoid the extra computation of 2-D integration, where $\hat{r}_i := \frac{B_i^N \cap B_i^S}{B_i^N \cup B_i^S}$. Here $B_i^N := X_i^N \times Y_i^N$ with X_i^N and Y_i^N are the central 95% intervals of the marginal predictive distributions. B_i^S is defined exactly the same way.

Table 4 shows the approximated ratios for the 13 landmarks. Landmarks with $r < 0.5$ (i.e. approximately less than a half of the total region is overlapped) is considered as having significant difference between the two groups, and they will be grouped together as a predictor. In this case, the predictor for this method includes landmarks 1, 2, 6 and 13. We remark that the momentum field used in our calculations is advantageous. For comparison, we repeat exactly the same calculations, but replace the momentum data by the original position data. Table 5 is the result of using position data, the counter part to Table 4. As we can see that no

landmarks have ratios that are less than 0.5 in Table 5. This indicates that the momentum representation has better separability than the location one.

The above two methods provide different viewpoints for our investigation. Method 1 finds a predictor by directly examining the peak values of the parameters (the 2-norm of mean momentum in our case). Method 2 uses Bayesian inference to search for a predictor. By examining the mean momentum, we found an additional abnormal landmark 6 when compared to Procrustes Method. The Bayesian approach, which also models the data, helps to dig out another suspected abnormal landmark 2. The three methods here seem to discover abnormal landmarks at different scales. In Table 6, we summarize the predictors found in the literature and our methods. The results are consistent.

Finally, we comment that from Table 6, the structure differences between the groups of schizophrenia and non-schizophrenia in our database most likely occur at the locations annotated by landmarks 1, 2, 6 and 13. Indeed, the region enclosed by these landmarks (see Fig. 13) were investigated in clinical trials and found to be relevant to schizophrenia from some of the literature [27,35].

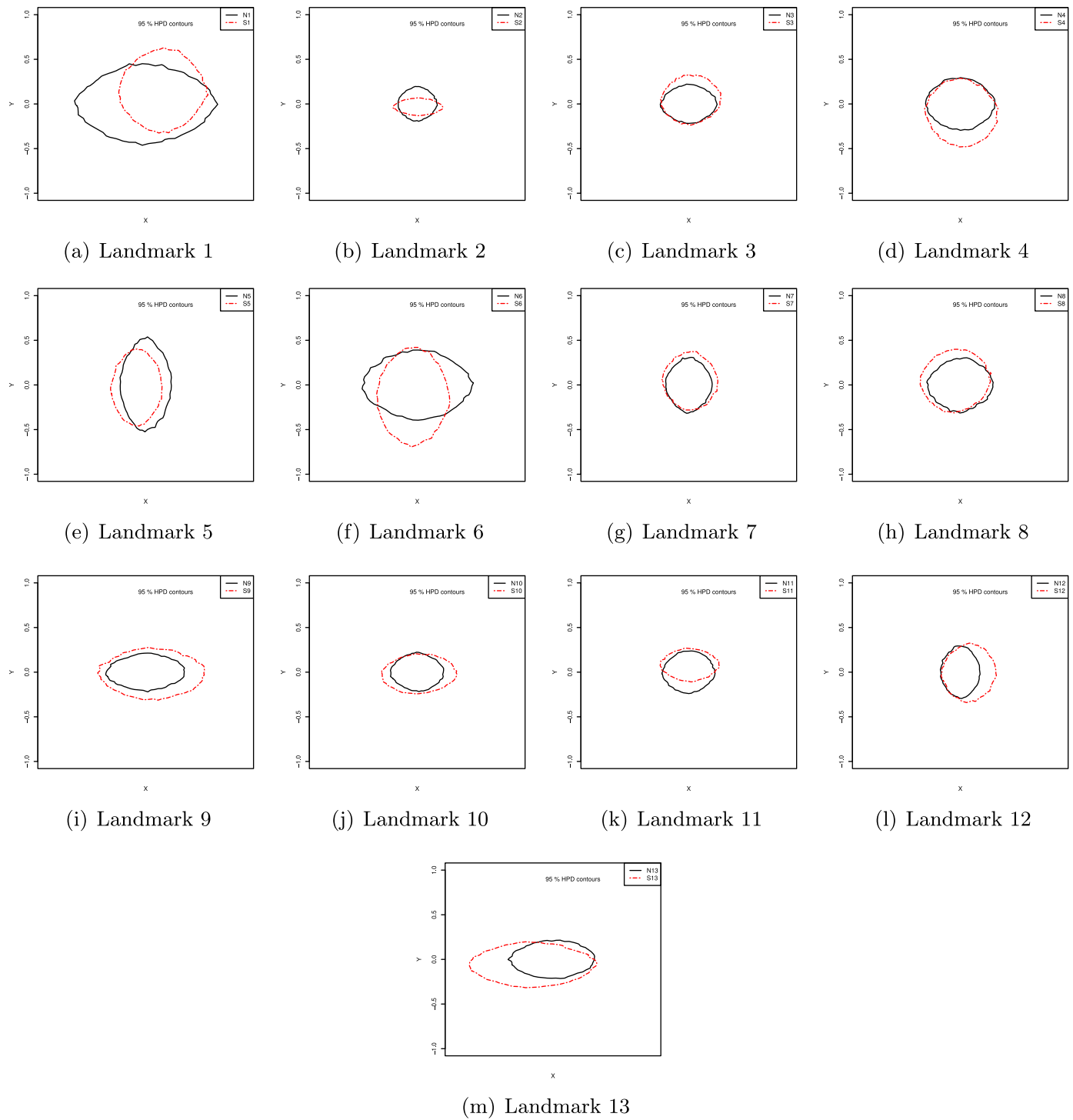


Fig. 12. 95% contour plots of the predictive distribution for each of the 13 landmarks. Black solid line is computed from N group, Red dashed line is computed from S group. Landmarks with their N and S contour plots differing much will be picked out as a predictor.

Table 4

Overlapped region rate for each landmark using the generated marginal distribution from the extracted momentum data. Landmark 1, 2, 6 and 13 got separated from others for their overlapped rate $r < 0.5$. They could be used as a predictor with their momentum representation.

Results on Momentum Data							
Landmark	1	2	3	4	5	6	7
r	0.4535	0.4286	0.6906	0.6542	0.5561	0.4972	0.7187
Landmark	8	9	10	11	12	13	
r	0.6825	0.5348	0.6155	0.5470	0.5554	0.4538	

Table 5

Overlapped region rate for each landmark using the generated marginal distribution from the original location data. No landmarks have $r < 0.5$, meaning the distribution are quite similar between the two group of data. The potential difference is not well reflected in the location representation.

Results on Position Data							
Landmark	1	2	3	4	5	6	7
r	0.7900	0.6636	0.5480	0.6249	0.5559	0.6045	0.5528
Landmark	8	9	10	11	12	13	
r	0.5887	0.5022	0.6171	0.5560	0.6066	0.6063	

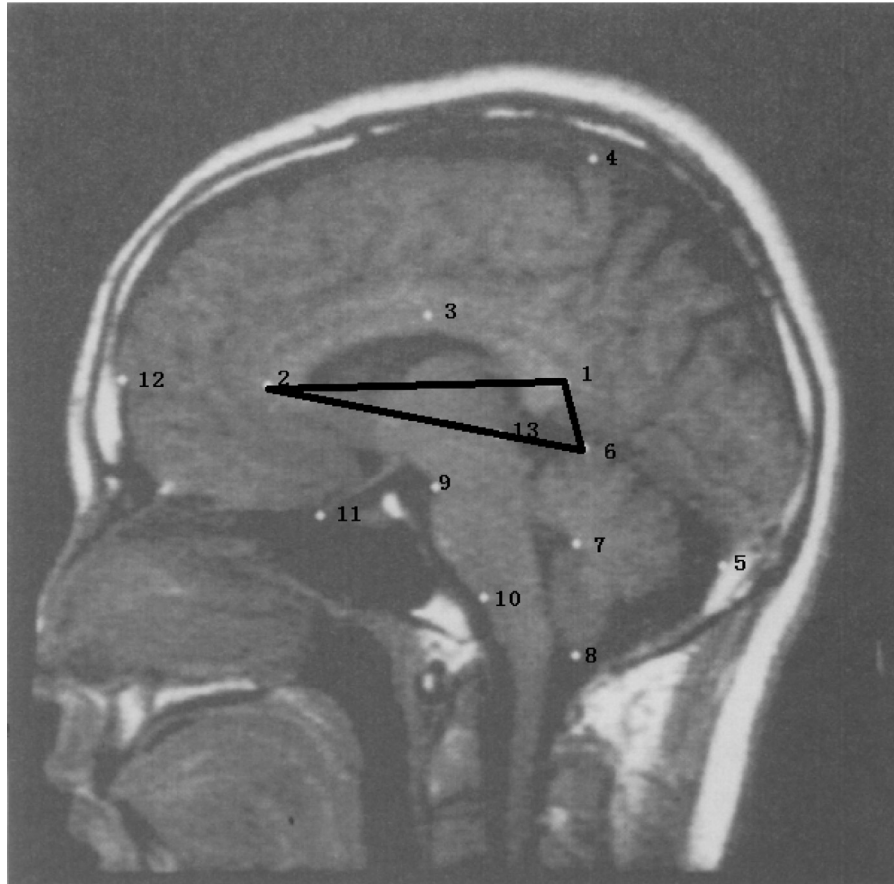


Fig. 13. Region enclosed by detectors 1, 2, 6, 13. Our calculation based on sample data suggests physical changes at mid-brain accompany schizophrenia. Certain important structures within this region such as: Thalamus, Hippocampus and Third Ventricle may be worth investigated. Some medical results [27,35] seem to support this observation.

Table 6

Summary of predictors from different methods.

Method	Predictor
Procrustes method in [4]	1, 13
Method 1: Mean Momentum (MM)	1, 6, 13
Method 2: Predictive Distribution Contour (PDC)	1, 2, 6, 13

Note 5. We would like to emphasize that the proposed algorithm alone is not a method for detecting outliers. The cluster analysis shown in Fig. 9 and Fig. 10 in Section 3 is to illustrate that the proposed algorithm with the robust weight has a better ability to extract the features than the conventional methods (such as Procrustes method), and with the aid of other tools, such as the Hamiltonian metric and the k -mean, we are able to separate the outlier. In Section 4, we first find the average template of the non-schizophrenic group by the proposed algorithm. Then we compute the residual momenta between the average template and each member in the schizophrenic group through template matching.

Finally we apply statistical methods to identify the abnormal momenta and hence the abnormal landmarks. Both our cluster analysis and the strategy of abnormal landmark detection lie entirely in the landmark space with the aid of applying statistical tools on the variables defined through template matching. Our methods are philosophically different from the pixel-based methods developed specifically for detecting outliers for images, for which statistical methods are applied directly to pixels. [8–10]. It is difficult to make direct comparison.

5. Conclusions

We present a class of algorithms for finding a group average. We prove a convergence theorem for the averaging algorithm with different weights under certain conditions. The Hamiltonian metric computed through the averaging algorithm allows us to classify the outliers, while the residual momentum variable resulting from the averaging process provides a local template-based linear coordinate system. The residual momentum variable also serves

as a representation which could enhance potential differences between groups under comparison when using the original location representation cannot, and thus is advantageous for related statistical analysis like abnormality detection. We apply the proposed algorithms, combining statistical approaches, on a medical image database. The database contains 14 non-schizophrenic and 14 schizophrenic templates annotated with 13 landmarks. The region of brain structure changes in schizophrenia identified by our approach is highly coherent with that in the literature.

Though the methods presented here are a first step, the applications of the proposed algorithms could have an impact on clinical medicine, forensics, and surveillance where images are used. For example, medical image abnormality detection that will help physicians for early diagnoses on schizophrenia, Alzheimers, and heart diseases, fast automatic face recognition algorithms that will provide forensic identification, and instantaneous image matching in surveillance that will help governments for intelligence gathering, the prevention of crime, the protection of a process, person, group or object, or for the investigation of crime.

Acknowledgments

LL was partially supported by NSF DMS 1413273. The paper was partially completed while SH was in the Statistics Department at the University of Wyoming and was partially funded by NSF DMS 1100615.

Appendix A. Statistical model specifics and implementation of method 2 in §4.2.2

The data are computed from the analysis of images from two groups of 14 non-schizophrenic and 14 schizophrenic brains, so that within each group at each of 13 landmarks, we have 14 observations (number of brains in group) on bivariate data (x, y components). Notationally, the vectors in the x and y directions result in data $\{(x, y)'_{i,j,k}; i = 1, \dots, 14, j = 1, \dots, 13, k = 1, 2\}$ which we denote as $(\mathbf{X}, \mathbf{Y})'$. In what follows, we use the notation $\{(x, y)'_{(l)}\}$ to denote a subset of $(\mathbf{X}, \mathbf{Y})'$ formed by collecting vectors over the index l .

For fixed (j, k) , we model the vectors $\{(x, y)'_{i,j,k}\}_{(i)}$ as conditionally independent bivariate normal random variables given the mean, $\mu_{j,k} = (\mu_{x,j,k}, \mu_{y,j,k})'$, and covariance matrix $\Sigma_{j,k}$, yielding the joint probability distribution,

$$[\{(x, y)'_{i,j,k}\}_{(i)} | \mu_{j,k}, \Sigma_{j,k}] = \prod_{i=1}^{14} N((x, y)'_{i,j,k} | \mu_{j,k}, \Sigma_{j,k}) \quad (\text{A.1})$$

where $[\cdot]$ denotes a probability distribution with $[\cdot | \cdot]$ for the conditional version. For purposes of data modeling, the bivariate distribution can be rewritten as the product of $[x_{i,j,k}]$, the marginal distribution of the first component of the $(x, y)'_{i,j,k}$ vector, with $[y_{i,j,k} | x_{i,j,k}]$, the conditional distribution of the second component given the first, where both are univariate normal distributions. This reformulation yields a likelihood where the covariance structure is a function of the two standard deviations $(\sigma_{x,j,k}, \sigma_{y,j,k})$ and one correlation $(\rho_{j,k})$. There is more flexibility in assigning priors to scalar parameters (Lunn et al. 2013, p.226), whereas the usual bivariate formulation restricts priors for the covariance matrix, Σ , to the Wishart distribution with little flexibility in the prior's shape. Given data from a (j, k) combination, we denote the contribution to the likelihood for $\theta_{j,k} = (\mu_{x,j,k}, \mu_{y,j,k}, \sigma_{x,j,k}, \sigma_{y,j,k}, \rho_{j,k})$ as $L(\theta_{j,k} | \{(x, y)'_{i,j,k}\}_{(i)})$, which is proportional to the quantity in Eq. (A.1). Then, the overall likelihood is $L(\theta | (\mathbf{X}, \mathbf{Y})') = \prod_{j,k} L(\theta_{j,k} | \{(x, y)'_{i,j,k}\}_{(i)})$, where $\theta = \{\theta_{j,k}\}_{(j,k)}$.

We used independence priors for the mean and covariance parameters,

$$[\{(\mu_{x,j,k}, \mu_{y,j,k}, \sigma_{x,j,k}, \sigma_{y,j,k}, \rho_{j,k})\}_{(j,k)}] = [\{(\mu_{x,j,k}, \mu_{y,j,k})\}_{(j,k)}] \times [\{(\sigma_{x,j,k}, \sigma_{y,j,k}, \rho_{j,k})\}_{(j,k)}]$$

where the components of each were modeled hierarchically. Specifically, for the mean parameters, we took $[\{(\mu_{x,j,k}, \mu_{y,j,k})\}_{(j,k)}] = [\{(\mu_{x,j,k})\}_{(j,k)}][\{(\mu_{y,j,k})\}_{(j,k)}]$, which implies that the means for the x and y components of the vectors had independent prior distributions. Across the (j, k) combinations, the means for the components were assumed to be conditionally independent with $[\{(\mu_{x,j,k})\}_{(j,k)}] | \mu_x, \sigma_x \sim N(\mu_x, \sigma_x^2)$ and $[\{(\mu_{y,j,k})\}_{(j,k)}] | \mu_y, \sigma_y \sim N(\mu_y, \sigma_y^2)$. We used fairly diffuse hyper-prior distributions $[\mu_x] = [\mu_y] \sim N(0, 10)$ and $[\sigma_x^{-2}] = [\sigma_y^{-2}] \sim U(0, 1)$. Fixing the hyper-priors at specific constants resulted in no difference in the final results.

The components of the covariance matrices were also modeled to have independent prior distributions,

$$[\{(\sigma_{x,j,k}, \sigma_{y,j,k}, \rho_{j,k})\}_{(j,k)}] = [\{(\sigma_{x,j,k})\}_{(j,k)}][\{(\sigma_{y,j,k})\}_{(j,k)}][\{(\rho_{j,k})\}_{(j,k)}]$$

with prior conditional independence for (j, k) combinations and distributions $[\sigma_{x,j,k} | a_x, b_x] \sim \Gamma(a_x, b_x)$, $[\sigma_{y,j,k} | a_y, b_y] \sim \Gamma(a_y, b_y)$ and $[a_x] = [a_y] \sim U(0, .5)$, $[b_x] = [b_y] \sim U(0, 1)$, which give diffuse gamma distributions. Finally, the priors for the correlation parameters were also conditionally independent, $[\rho_{j,k} | \mu_\rho, \sigma_\rho^2] \sim N(\mu_\rho, \sigma_\rho^2) I_{[-1,1]}(\rho)$; that is, a truncated normal distribution with $[\mu_\rho] \sim N(0, 2)$ and $[\sigma_\rho^{-2}] \sim U(0, 1)$. Fixing parameters of the priors to constants gave similar results to assigning hyper-prior distributions.

The overall prior distribution for θ , $[\theta]$, is given by a product of the prior and hyper-prior distributions, with the hyper-parameters $\eta = (\mu_x, \mu_y, \sigma_x, \sigma_y, a_x, b_x, a_y, b_y, \mu_\rho, \sigma_\rho)$ integrated out. The resulting posterior distribution for θ is obtained as

$$[\theta | (\mathbf{X}, \mathbf{Y})'] = \frac{L(\theta | (\mathbf{X}, \mathbf{Y})') [\theta]}{\int L(\theta | (\mathbf{X}, \mathbf{Y})') [\theta] d\theta},$$

with the high dimensional integration in the denominator bypassed using Markov chain Monte Carlo (MCMC) sampling.

Furthermore, we use the posterior distributions to obtain posterior predictive distributions for 'new' observations from each of the 26 landmark-group combinations. Using predictive inference to assess the potential similarities and differences across the two groups at each landmark gives inference in terms of the observables, namely, the $\{(x, y)'_{(j,k)}\}$ vectors instead of making inferences on the parameters in θ used to model the data on the vectors. The predictive distributions are obtained as

$$[(x, y)'_{new,j,k} | (\mathbf{X}, \mathbf{Y})'] = \int [(x, y)'_{new,j,k} | \theta] [\theta | (\mathbf{X}, \mathbf{Y})'] d\theta$$

where $[(x, y)'_{new,j,k} | \theta]$ is a bivariate normal distribution with $(x, y)_{new,j,k}$ conditionally independent of the observed data $(\mathbf{X}, \mathbf{Y})'$. Again, MCMC is used to bypass this integration.

We used open source programs for all model implementation; specifically, R (R Development Core Team 2015) using the 'rjags'TM (Plummer 2012b) package which interface with MCMC sampling programs 'Just Another Gibbs Sampler (JAGS)'TM (Plummer 2003). The 'coda'TM package (Plummer 2012a) assists in managing and presenting sampling output. The 'emdbook'TM package (Bolker 2014) was used to produce contour plots.

References

- [1] M.F. Beg, M.I. Miller, A. Trounev, L. Younes, Computing large deformation metric mappings via geodesics flows of diffeomorphisms, *Int. J. Comp. Vis.* 61 (2) (2005) 139–157.
- [2] M.F. Beg, A. Khan, Computing an average anatomical using LDDMM and geodesic shooting, *IEEE ISBI* (2006) 1116–1119.

- [3] F.L. Bookstein, Landmark methods for forms without landmarks: morphometrics of group differences in outline shape, *Med. Image Anal.* 1 (3) (1996) 225–243.
- [4] F.L. Bookstein, Biometrics, biomathematics and the morphometric synthesis, *Bull. Math. Biol.* 56 (2) (1996) 313–365.
- [5] R. Camassa, D. Kuang, L. Lee, Solitary waves and n -particle algorithms for a class of euler-poincaré equations, *Stud. Appl. Math.* 137 (2016) 502–546.
- [6] R. Camassa, D. Kuang, L. Lee, A geodesic landmark shooting algorithm for template matching and its applications, *SIAM J. Imaging Sci.* 10 (1) (2017) 303–334.
- [7] A. Chertock, P. Du Toit, J.E. Marsden, Integration of the EPDIFF equation by particle methods, *ESAIM: M2AN* 46 (2012) 515–534.
- [8] B. Du, L. Zhang, Random-selection-based anomaly detector for hyperspectral imagery, *IEEE Trans. Geosci. Remote Sensing* 49 (50) (2011) 1578–1589.
- [9] B. Du, L. Zhang, A discriminative metric learning based anomaly detection method, *IEEE Trans. Geosci. Remote Sens.* 52 (11) (2014) 6844–6857.
- [10] B. Du, R. Zhao, L. Zhang, L. Zhang, A spectral-spatial based local summation anomaly detection method for hyperspectral images, *Signal Process.* 124 (2016) 115–131.
- [11] P. Dupuis, U. Grenander, M.I. Miller, Variational problems on flows of diffeomorphisms for image matching, *Q. Appl. Math.* 56 (1998) 587–600.
- [12] P.T. Fletcher, C. Lu, M. Pizer, S. Joshi, Principal geodesic analysis for the study of nonlinear statistics of shape, *IEEE Trans. Medical Imaging* 23 (8) (2004) 995–1005.
- [13] P.T. Fletcher, S. Venkatasubramanian, S. Joshi, Robust statistics on riemannian manifolds via the geometric median, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
- [14] J. Gao, A.N. Evans, Expression robust 3d face landmarking using thresholded surface normals, *Pattern Recognit.* 78 (2018) 120–132.
- [15] J. Glaunes, S. Joshi, Template estimation from unlabeled point set data and surfaces for computational anatomy, in: *Proceedings of the International Workshop on the Mathematical Foundations of Computational Anatomy (MFCA-2006)*.
- [16] J. Glaunes, A. Trounev, L. Younes, Diffeomorphic matching 10 of distributions a new approach for unlabelled point-sets and submanifolds matching, in: *Proceedings of CVPR'04*.
- [17] U. Grenander, *General Pattern Theory*, Oxford University Press, Clarendon, UK, 1993.
- [18] R.A. Guler, S. Tari, G. Unal, Landmarks inside the shape: shape matching using image descriptors, *Pattern Recognit.* 49 (2016) 79–88.
- [19] <http://life.bio.sunysb.edu/morph/data/book-schizo.dta>.
- [20] Y. Han, Recognize objects with three kinds of information in landmarks, *Pattern Recognit.* 46 (2013) 2860–2873.
- [21] D.D. Holm, J.T. Ratnanather, A. Trounev, L. Younes, Soliton dynamics in computational anatomy, *Neuroimage* 23 (2004) 170–178.
- [22] S. Joshi, M.I. Miller, Landmark matching via large deformation diffeomorphisms, *IEEE Trans. Image Process.* 9 (2000) 1357–1370.
- [23] H. Karcher, Riemannian center of mass and mollifier smoothing, *Commun. Pure Appl. Math.* 30 (5) (1977) 509–541.
- [24] H. Kuhn, A note on fermat's problem, *Math. Program.* 4 (1973) 98–107.
- [25] P. Lorenzen, B. Davis, S. Joshi, in: *Unbiased atlas formation via large deformations metric mapping*, 8, 2005, pp. 411–418.
- [26] J. Ma, M. Miller, A. Trounev, L. Younes, Bayesian template estimation in computational analysis, *Neuroimage* 42 (2008) 251–262.
- [27] L. Marsh, R.L. Suddath, N. Higgins, D.R. Weinberger, Medial temporal lobe structures in schizophrenia: relationship of size to duration of illness, *Schizophr. Res.* 11 (3) (1994) 225–238.
- [28] M. Micheli, P.W. Michor, D. Mumford, Sectional curvature in terms of the co-metric, with applications to the riemannian manifolds of landmarks, *SIAM J. Imaging Sci.* 5 (1) (2012) 394–433.
- [29] R.I. McLachlan, S. Marsland, N -particle dynamics of the euler equations for planar diffeomorphisms, *Dyn. Sys.* 22 (3) (2007) 269–290.
- [30] M.I. McLachlan, S. Marsland, Discrete mechanics and optimal control for image registration, *Anziam J.* 48 (2007) 1–16.
- [31] M.I. Miller, A. Trounev, L. Younes, On metrics and the euler-lagrange equations of computational anatomy, *Annu. Rev. Biomed. Eng.* 4 (2002) 375–405.
- [32] M.I. Miller, A. Trounev, L. Younes, Geodesic shooting for computational anatomy, *J. Math Imaging* 24 (2006) 209–228.
- [33] M.I. Miller, A. Trounev, L. Younes, Hamiltonian systems and optimal control in computational anatomy: 100 years since d'arcy thompson, *Annu. Rev. Biomed. Eng.* 17 (2015) 447–509.
- [34] D. Mumford, A. Desolneux, *Pattern theory: the stochastic analysis of real world signals*, A K Peters, Ltd, Natick, MA, 2010.
- [35] P.C. Nopoulos, J.W. Ceilleys, E.A. Gailisa, N.C. Andreasen, An MRI study of mid-brain morphology in patients with schizophrenia: relationship to psychosis, neuroleptics, and cerebellar neural circuitry, *Biol. Psychiatry* 49 (1) (2001) 13–19.
- [36] X. Pennec, Intrinsic statistics on riemann manifolds: basic tools for geometric measurements, *J. Math. Imaging Vis.* 25 (2006) 127–154.
- [37] A. Srivastava, S. Joshi, W. Mio, X. Liu, Statistical shape analysis: clustering, learning and testing, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (4) (2005).
- [38] N. Thakoor, J. Gao, S. Jung, Hidden markov model-based weighted likelihood discriminant for 2-d shape classification, *IEEE Trans. Image Process.* 16 (11) (2007) 2707–2719.
- [39] J. Thewlis, H. Bilen, A. Vedaldi, Unsupervised learning of object landmarks by factorized spatial embeddings, in: *The IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3229–3238.
- [40] M. Vaillant, M.I. Miller, L. Younes, A. Trounev, Statistics on diffeomorphisms via tangent space representations, *Neuroimage* 23 (2004) 161–169.
- [41] Y. Vardi, C.H. Zhang, The multivariate L_1 -median and associated data depth, *Proc. Natl. Acad. Sci. U.S.A.* 97 (4) (2000) 1423–1426.
- [42] B.S.C. Wade, S.H. Joshi, B.A. Gutman, P.M. Thompson, Machine learning on high dimensional shape data from subcortical brain surfaces: a comparison of feature selection and classification methods, *Pattern Recognit.* 63 (2017) 731–739.
- [43] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, A. Kassim, Robust facial landmark detection via recurrent attentive refinement networks, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science*, vol 9905, Springer, Cham.
- [44] L. Younes, F. Arrate, M.I. Miller, Evolutions equations in computational anatomy, *Neuroimage* 45 (2009) 40–50.
- [45] L. Younes, *Shapes and diffeomorphisms*, Applied Mathematical Sciences, 171, Springer, 2010.
- [46] J. Zhang, Y. Gao, Y. Gao, B.C. Munsell, D. Shen, Detecting anatomical landmarks for fast alzheimer's disease diagnosis, *IEEE Trans. Med. Imaging* 35 (12) (2016) 2524–2533.

Snehalata Huzurbaza, PhD, is Professor and Chair of the Biostatistics Department at the West Virginia University. She is a fellow (elected) of the American Statistical Association.

Dongyang Kuang, PhD, is a postdoctoral researcher in the Department of Mathematics and Statistics at the University of Ottawa, Canada. He obtained his PhD degree in Mathematics from the Department of Mathematics and Statistics at the University of Wyoming in June, 2016.

Long Lee, PhD, is Professor of Mathematics in the Department of Mathematics and Statistics at the University of Wyoming. He obtained his PhD degree in Applied Mathematics from the University of Washington, Seattle in 2002, under the supervision of Professor Randall J. LeVeque. He completed his post-doctoral training at UNC-Chapel Hill with Professor Roberto Camassa in 2005.