Representing and Learning High Dimensional Data with the Optimal Transport Map from a Probabilistic Viewpoint

Serim Park
Department of Electrical
and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, 15213
serimp@andrew.cmu.edu

Matthew Thorpe
Department of Applied Mathematics
and Theoretical Physics
The University of Cambridge
Cambridge, CB3 0WA
m.thorpe@maths.cam.ac.uk

Abstract

In this paper, we propose a generative model in the space of diffeomorphic deformation maps. More precisely, we utilize the Kantarovich-Wasserstein metric and accompanying geometry to represent an image as a deformation from templates. Moreover, we incorporate a probabilistic viewpoint by assuming that each image is locally generated from a reference image. We capture the local structure by modelling the tangent planes at reference images.

Once basis vectors for each tangent plane are learned via probabilistic PCA, we can sample a local coordinate, that can be inverted back to image space exactly. With experiments using 4 different datasets, we show that the generative tangent plane model in the optimal transport (OT) manifold can be learned with small numbers of images and can be used to create infinitely many 'unseen' images. In addition, the Bayesian classification accompanied with the probabilist modeling of the tangent planes shows improved accuracy over that done in the image space. Combining the results of our experiments supports our claim that certain datasets can be better represented with the Kantarovich-Wasserstein metric. We envision that the proposed method could be a practical solution to learning and representing data that is generated with templates in situatons where only limited numbers of data points are available.

1. Introduction

Optimal Transport based techniques for signal and data analysis have received increased attention recently [10]. Given their abilities to provide accurate generative models for signal intensities and other data distributions, they have been used in a variety of applications including content-based retrieval, cancer detection, image super-resolution, and statistical machine learning, to name a few, and shown to produce state of the art results in several applications.

Manifolds arise naturally as the appropriate representations for images. For example, when representing face images, the linear average of two faces often does not resemble a face. One more reasonable representation, and the one we adopt in this work, is to use diffeomorphic deformation maps to capture the nonlinear characteristics innate in image data. Here, geodesics are given by 'optimal rearrangements' of one image into another, a notion made precise in the optimal transport framework.

In previous works, Fletcher et al. [5] suggested a principal component analysis for Lie Groups and computed the approximate principal geodesics by minimizing the sum of squared geodesic distances to the data. Ziezold et al. formulated PCA for a Riemannian manifold in [7] based on geodesics of the intrinsic mean. Boissard et al. [2] defined principal geodesic components with respect to the Wasserstein metric assuming that each input measure has been generated from a single template density. Wang et al. [26] proposed to find an approximate principal geodesic in the tangent plane of the Wasserstein-Kantorovich space for a single template. Cuturi et al. [18] proposed a new algorithm to compute approximate geodesics for the Wasserstein space by regularizing with entropy.

In addition, diffeomorphic maps have proven to be useful in modeling shape space [23, 1], in reconstructing images from under-sampled data [11], learning the geometrical transformations between the images [20, 17], visualizing the smooth deformations between the images [23, 27], and differentiating different classes of shapes [25, 21].

Here, along the line of previous attempts to learn and represent data in the diffeomorphic space, we propose to utilize the geometric characteristic of diffeomorphic space based on the Kantorovich-Wasserstein metric. We utilize the geometric transforms learned between the images to create even more images that can be utilized for various

applications, i.e. classification. Additionally, we embody the probabilistic viewpoint in modeling the diffeomorphic space and generalize it to Bayesian classification that is more natural with the data generation process. For the image set that is created with a few templates, we suggest that learning deformation maps are a better solution than a machine-learning based approach such as using variational auto-encoders [9], [16].

Our method is similar to the work by Simard et al. [20] which synthesized images using random deformation maps and to the work by Hauberg et al. [6] which learned the diffeomorphic mapping. However, our work differs from previous approaches in that we define multiple templates which characterize multiple tangent planes, and associate it with latent variables that governs which tangent plane the data belongs to. We emphasize that our work is the first work to address that a set of tangent planes accompanied with the Kantorovich-Wasserstein metric can be used to formulate a generative model for a image set, associate it with probabilistic view point, and generalize this concept to generate more data and apply it to perform classification of images.

The paper is structured as follows. In Sec. 2, we introduce the notations and preliminaries. In Sec. 3, we describe the method for modeling and learning the manifold with tangent planes. In Sec. 4, we show experimental validation, and a useful application of our method. Sec. 5 wraps up with conclusion, limitations, and future studies.

2. Preliminaries

2.1. The Optimal Transport Metric and Geometry

Here we consider the optimal transport framework in discrete settings but we note that it is usually described in terms of measures which can include both discrete and continuous settings. The optimal transport distance is based around the cost of transporting 'mass' from one image to another. Images are normalized so that intensities of all pixels sum to one, i.e. each image contains an equal amount of mass.

Let $c : \Omega \times \Omega \to [0, \infty)$ be the cost function, so that $c(\omega_1, \omega_2)$ is the cost of transporting one unit of mass at $\omega_1 \in \Omega$ to $\omega_2 \in \Omega$. A transport plan between a template image $\mathbf{r}$ and a target image $\mathbf{x}_i$ is any matrix π that transports r to x_i, mathematically we write this as

$$\begin{cases} \sum_j \pi(\omega, \omega_j) = \mathbf{r}(\omega) & \forall \omega \in \Omega, \\ \sum_j \pi(\omega_j, \omega) = \mathbf{x}_i(\omega) & \forall \omega \in \Omega. \end{cases} \tag{1}$$

We say that $\pi \in \Pi(\mathbf{r}, \mathbf{x}_i)$ if π satisfies (1), $\Pi(\mathbf{r}, \mathbf{x}_i)$ is the set of all mass preserving transportation plans. The cost of a transport plan π between $\mathbf{r}$ and $\mathbf{x}_i$ is given by $\sum_{j,k} \pi(\omega_j, \omega_k) c(\omega_j, \omega_k)$. We will use the quadratic cost $c(\omega_1, \omega_2) = |\omega_1 - \omega_2|^2$ in which case we can define the Wasserstein distance (often called the Kantorovich-Wasserstein distance) by

$$d_W(\mathbf{r}, \mathbf{x}_i) = \left(\min_{\pi \in \Pi(\mathbf{r}, \mathbf{x}_i)} \sum_{j,k} \pi(\omega_j, \omega_k) |\omega_j - \omega_k|^2 \right)^{\frac{1}{2}}. \tag{2}$$

The minimum is attained and d_W defines a metric [24]. Furthermore the metric space is a Riemannian manifold [4] that we describe now.

Suppose the optimal transport plan, i.e. π^* which achieves the minimum in (2), is unique and sends mass from each pixel $\omega \in \Omega$ to a unique location $\phi(\omega)$ in Ω. Then ϕ is called the *optimal transport map*. One can also write $d_W(\mathbf{r}, \mathbf{x}_i) = \left(\sum_j |\omega_j - \phi(\omega_j)|^2 \mathbf{r}(\omega_j) \right)^{\frac{1}{2}}$ and define the vector map $\mathbf{v}(\omega) = \phi(\omega) - \omega$ which gives the deformation of each pixel. Clearly $d_W(\mathbf{r}, \mathbf{x}_i) = \left(\sum_j |\mathbf{v}(\omega_j)|^2 \mathbf{r}(\omega_j) \right)^{\frac{1}{2}}$ and in fact the set of vector maps

$$T_r = \left\{ \mathbf{v} : \Omega \to \mathbb{R}^2 \ : \ \sum_j |\mathbf{v}(\omega_j)|^2 \mathbf{r}(\omega_j) < \infty \right\}$$

is the tangent plane at r. The Wasserstein distance $d_W(r, x_i)$ is the length of the shortest curve (geodesic) containing r and x_i.

Given a vector map $\mathbf{v} \in T_r$ one can define a 'new' image by $\mathbf{x}_{\text{new}}(\omega) = \phi_{\#}\mathbf{r}(\omega) := \sum_{i \text{ s.t. } \phi(\omega_i) = \omega} \mathbf{r}(\omega_i)$ where $\phi = \mathbf{v} + \mathbb{I}$ is the transport map. The construction is such that $\mathbf{x}_{\text{new}}$ lies on the geodesic from $\mathbf{r}$ in the direction $\mathbf{v}$, in particular, $d_W(\mathbf{r}, \mathbf{x}_{\text{new}}) = \left(\sum_i |\mathbf{v}(\omega_i)|^2 \mathbf{r}(\omega_i) \right)^{\frac{1}{2}}$. In the sequel, the idea is that the tangent plane is restricted to a low dimensional space spanned by a small number of basis vectors, i.e. we restrict the tangent plane to $\{\mathbf{v} = W\boldsymbol{\alpha} + \boldsymbol{\mu} = \sum_{i=1}^{\ell} \mathbf{w}_i \alpha_i + \boldsymbol{\mu} \ : \ \alpha_i \in \mathbb{R}\}$ where $\{\mathbf{w}_i\}_{i=1}^{\ell}, \boldsymbol{\mu}$ are vector maps.

2.2. Parameterizing Tangent Plane with Probabilistic Framework

As we model the image manifold with tangent planes, we parameterize the tangent planes as a joint distribution over observed and hidden variables therefore embodying a probabilistic setting:

$$P(\mathbf{v}, \boldsymbol{\alpha}_z, z) = P(\mathbf{v}|z, \boldsymbol{\alpha}_z) P(\boldsymbol{\alpha}_z|z) P(z) \tag{3}$$

where z indexes the tangent plane/reference image, $\boldsymbol{\alpha}_z \in \mathbb{R}^{\ell_z}$ are local coordinates, and $\mathbf{v} \in \mathbb{R}^{2d}$ is a deformation map of an image (d being the number of pixels in images). The variable $\mathbf{v}$ is observed, whilst α_z and z are hidden.

The tangent planes $\mathcal{T}_{r_z}$ have tangent planeial points r_z which also serve as template images. The tangent planes are indexed by a discrete hidden variable $z \in \{1, \cdots, K\}$.

The model assumes that an image is sampled from a tangent plane $\mathcal{T}_{r_z}$ with prior $P(z) = p_z$.

Each image has corresponding local coordinate ('features') $\boldsymbol{\alpha}_z \in \mathbb{R}^{\ell_z}$ in each tangent plane. And in each tangent plane the deformation map $\mathbf{v}$ can be represented with its local coordinate $\boldsymbol{\alpha}_z$ and tangent planes' basis vectors (i.e. column vectors of W_z)

$$\mathbf{v} = W_z \boldsymbol{\alpha}_z + \boldsymbol{\mu}_z + e_z, \tag{4}$$

where e_z is Gaussian random noise with distribution $N(0, \Psi_z)$. In addition, we assume that local coordinates are independently normally distributed:

$$P(\boldsymbol{\alpha}_z|z) = \frac{1}{\sqrt{(2\pi)^{\ell_z}}} \exp^{-\frac{1}{2}\boldsymbol{\alpha}_z^T \boldsymbol{\alpha}_z}.$$

Therefore, $P(\mathbf{v}|\boldsymbol{\alpha}_z, z)$ is normally distributed and $P(\mathbf{v})$ is a mixture of normal distributions:

$$\begin{aligned} P(\mathbf{v}|\boldsymbol{\alpha}_z, z) &\sim N\left(\boldsymbol{\mu}_z + W_z \boldsymbol{\alpha}_z, \Psi_z\right) \\ P(\mathbf{v}) &\sim \sum_z p_z N(\boldsymbol{\mu}_z, W_z W_z^T + \Psi_z) \end{aligned} \tag{5}$$

This is also well known as a Factor Analysis (FA) model with Gaussian prior for $P(\boldsymbol{\alpha}_z)$ and prior $P(z) = p_z$.

2.3. Probabilistic Principal Component Analysis (PPCA)

Additionally, as we assume normally distributed noise $e_z \sim N(0, \sigma_z^2 I)$, Eq. (5) simplifies to the PPCA model. The variables W_z, $\boldsymbol{\mu}_z$, and $\Psi_z = \sigma_z^2 I$ in Eq. (3) can be found via eigen-decomposition when $2d > \ell_z$, i.e. the dimension of local coordinates $\boldsymbol{\alpha}_z$ is smaller than that of deformation map $\mathbf{v}$ [22]. Consider a set deformations $V_z = [\mathbf{v}_1|\cdots|\mathbf{v}_{N_z}]$ in tangent plane $\mathcal{T}_{r_z}$. Let U_z be an orthonormal matrix of eigenvectors and Λ_z a diagonal matrix of eigenvalues from eigen-decomposition on the mean centered covariance matrix $\tilde{V}_z \tilde{V}_z^T$, i.e.

$$U_z^{-1} \tilde{V}_z \tilde{V}_z^T U_z = \Lambda_z.$$

The maximum likelihood (ML) estimator of W_z, $\boldsymbol{\mu}_z$, and σ_z are

$$W_z = U_{\ell_z} \Lambda_{\ell_z}^{1/2}, \quad \boldsymbol{\mu}_z = \frac{1}{N_z} \sum_{k=1}^{N_z} \mathbf{v}_k, \quad \sigma_z^2 = \frac{1}{2d - \ell_z} \sum_{i=\ell_z+1}^{2d} \lambda_{ii}$$

where U_{ℓ_z} is the orthonormal matrix with ℓ_z largest eigenvectors in columns and Λ_{ℓ_z} is the diagonal matrix with ℓ_z largest eigenvalues in descending order at its diagonal.

Once the feature $\boldsymbol{\alpha}_z$ is drawn from the normal distribution, a new OT deformation can be sampled according to (5) via

$$\mathbf{v}_\alpha = \frac{1}{N_z} \sum_{i=1}^{N_z} \mathbf{v}_i + U_{\ell_z} \Lambda_{\ell_z}^{\frac{1}{2}} \boldsymbol{\alpha}_z + e_z.$$

3. Methods

3.1. How to find the template image

Consider a set of images $\{\mathbf{x}_i\}_{i=1}^N$. We assume that each image is deformed with a smooth mass preserving map with respect to a template image. We additionally assume that the template image is an element in the set of templates $\{\mathbf{r}_z\}_{z=1}^K$. In other words, the images can be clustered into different groups that share the same template, and the corresponding optimal transport maps of images that share the same template are denoted as $\{\mathbf{v}_i^{(z)}\}_{i=1}^{N_z}$ with the template index z made explicit.

The question arises how to select the set of template images. In order to do this we briefly recap the linearized-OT (LOT) distance [26]. Given a tangent planeial point u we define $\chi_u(\mathbf{x}_i) = \mathbf{v}_i$ to be the OT deformation between $\mathbf{x}_i$ and u (u would often be called a template point but in order to minimise confusion with the template images $\mathbf{r}_z$ we will use the terminology tangent planeial point here). One has $d_W(\mathbf{x}_i, u) = \|\chi_u(\mathbf{x}_i)\|_u = \|\chi_u(\mathbf{x}_i) - \chi_u(u)\|_u$ where $\|\mathbf{v}\|_u^2 = \sum_i \mathbf{v}(\omega_i)^2 u(\omega_i)$. The LOT distance is defined by $d_{u,LOT}(\mathbf{x}_i, \mathbf{x}_k) = \|\chi_u(\mathbf{x}_i) - \chi_u(\mathbf{x}_k)\|_u$. Heuristically the LOT distance projects $\mathbf{x}_i$ and $\mathbf{x}_k$ onto the tangent plane at the point u and computes the Euclidean distance in the tangent plane. When there does not exist a transport map (i.e. if mass is split) then the situation is more complicated and we refer to [26] for more details.

Before moving on to the algorithmic details, we mention that for the tangent plane of the OT manifold at $\mathbf{r}_z$, the intrinsic mean with respect to the LOT distance and the extrinsic mean (defined by averaging transport maps) coincide. In particular, the intrinsic mean of the LOT distance with tangent planeial point u is given as:

$$\begin{aligned} \mathbf{r}_{\text{intrinsic}} &= \arg\min_{\mathbf{r}} \sum_{i=1}^N d_{u,LOT}(\mathbf{x}_i, \mathbf{r}) \\ &= \arg\min_{\mathbf{r}} \sum_{i=1}^N \sum_j \left(\mathbf{v}_{\mathbf{x}_i}(\omega_j) - \mathbf{v}_{\mathbf{r}}(\omega_j)\right)^2 u(\omega_j) \\ &= \arg\min_{\mathbf{r}} \sum_j \sum_{i=1}^N \left(\mathbf{v}_{\mathbf{x}_i}(\omega_j) - \mathbf{v}_{\mathbf{r}}(\omega_j)\right)^2 u(\omega_j) \end{aligned}$$

where $\mathbf{v}_{\mathbf{x}}$ is the OT deformation between $\mathbf{x}$ and u. Simple calculus gives us that $\mathbf{v}_{\mathbf{r}} = \frac{1}{N} \sum_{i=1}^N \mathbf{v}_{\mathbf{r}_i}$ and the intrinsic mean corresponds to the density that is deformed from u by the mass preserving map $\mathbf{v}_{\mathbf{r}}$. This is exactly the extrinsic mean.

Now we describe how to find the multiple template images from a set of images. Put simply, K-means clustering with Euclidean distance is performed in the tangent plane with global template u. The K cluster centers are then mapped back to the image space, yielding 'K' template im-

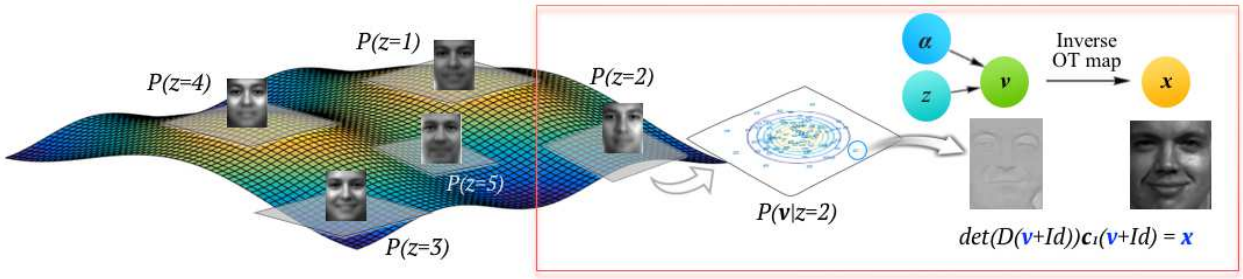

Figure 1: Representing face image space with the optimal transport manifold which can be approximated with tangent planes. The facial image corresponding to the tangent planeial point of each tangent plane is shown.

ages. As stated above the extrinsic mean (cluster center) for each cluster is identical to the intrinsic mean. We note that K-means clustering with Euclidean distance in the tangent plane is an approximation of doing K-means clustering with the Wasserstein distance. Also, we note that clustering can be performed using alternative methods, e.g. Gaussian mixture models, or classification can be used when the associated labels that characterize the image sets are available.

The number of cluster centers, K, determines how many templates govern the data generation process. K can be determined via cross-validation. The cluster centers will serve as the template when we compute the mass preserving map for each image belonging to that cluster. Note that this *local* template is different from the global template u that was used to perform K-means clustering in the tangent plane with the tangent planeial point u. After K-means clustering, we now have K different tangent planes at tangent planeial points $\{\mathbf{r}_z\}_{z=1}^K$.

3.2. Probabilistic deformation model

Consider a set of images $\{\mathbf{x}_i\}_{i=1}^N$, and corresponding optimal transport (OT) maps $\{\mathbf{v}_i\}_{i=1}^N$, as before, generated with respect to the template images $\{\mathbf{r}_z\}_{z=1}^K$ found via K-means as described in the previous section. We now introduce the probabilistic framework of the data generation process. We assume that a discrete latent random variable $z \in \{1, \ldots, K\}$ is involved in the data generation process, which governs how probable it is that the data is generated from the template $\mathbf{r}_z$. We can model the distribution of OT maps with a mixture of factor analyzers (MFA)

$$P(\mathbf{v}) = \sum_{z=1}^{K} p_z \int P(\mathbf{v}|\boldsymbol{\alpha}_z, z) P(\boldsymbol{\alpha}_z|z) d\boldsymbol{\alpha}_z,$$

where $P(\mathbf{v}|\boldsymbol{\alpha}, z)$ and $P(\boldsymbol{\alpha}|z)$ are normally distributed as in Sec. 2.2.

The latent variable z indexes the tangent planes (or equivalently the templates, $\mathbf{r}_z$). Fig. 1 illustrates an OT manifold represented with 5 tangent planes, each associated with prior $P(z)$. The tangent planeial points ($\mathbf{r}_z$) are shown on top of each tangent plane.

The red box in Fig. 1 draws an image generation pipeline. Once z is given, an OT map $\mathbf{v}$ is drawn from $P(\mathbf{v}|z)$. And then, $\mathbf{v}$ is converted back to an image $\mathbf{x}$ by pushing forward the template measure by $\mathbf{v} + \mathbb{I}$, i.e. $\mathbf{x} = (\mathbf{v} + \mathbb{I})_{\#} r_z$. The relationship between an image space ($\mathbf{x}$), an OT map space ($\mathbf{v}$), and a local coordinate ($\boldsymbol{\alpha}$) space is shown.

The learning consists of two folds. In the first step one finds the tangent plane assignments for every image $\mathbf{x}_i$ and the template images $\mathbf{r}_z$ (via K-means clustering). At the second step, the statistics for the tangent planes are collected, i.e. W_z, $\boldsymbol{\mu}_z$, and σ_z in Eq. 4.

Procedure Generating unseen images using OT Space

Learning Step: Find $\mathcal{T}_{\mathbf{r}_z}, z \in \{1, \cdots K\}$.
1 Set the number of tangent planes K.
2 Find the tangent planeial points $\mathbf{r}_z$.
3 **foreach** *tangent plane* $\mathcal{T}_{\mathbf{r}_z}$ **do**
4 For image $\mathbf{x}_i$ in cluster z, compute the OT map $\mathbf{v}_i$ between $\mathbf{r}_z$ and $\mathbf{x}_i$.
5 Learn W_z, $\boldsymbol{\mu}_z$, and σ_z via PPCA.
6 Find p_z.
7 **end**
Generation Step: Generate Unseen Images.
8 Draw $z \sim p(z)$.
9 Draw $\alpha^* \sim N(0, I)$.
10 Compute the unseen OT map, i.e. $\mathbf{v}^* = W_z \alpha^*$.
11 Compute the unseen image via inverse OT mapping i.e. $(\mathbf{v}^* + \mathbb{I})_{\#} r_z$.

3.3. Generation Step

Here a step for synthesizing 'unseen' images is described. First, z is drawn from $p(z)$, which determines the tangent plane $\mathcal{T}_{\mathbf{r}_z}$. Then, $\boldsymbol{\alpha}^*$ is drawn from $N(0, I)$. The

'unseen' OT map $\mathbf{v}^*$ is synthesized by

$$\mathbf{v}^* = W_z \boldsymbol{\alpha}^* + \boldsymbol{\mu}_z, \tag{6}$$

i.e. a linear combination of 'deformation patterns' in column vectors of W_z with Gaussian $\boldsymbol{\alpha}$. Once $\mathbf{v}^*$ is generated, an unseen image $\mathbf{x}^*$ can be uniquely identified by inverse OT mapping with respect to the template $\mathbf{r}_z$ (tangent planeial points for $\mathcal{T}_z$), i.e. $\mathbf{x}^* = (\mathbf{v}^* + \mathbb{I})_{\#} r_z$.

3.4. Bayesian Classification

Since we assume that the data is generated in a probabilistic framework, it arises as a natural choice to formulate Bayesian classification. Given a set of images $\mathbf{x}_{i=1}^N$ with labels $y_{i=1}^N$, let's assume that we have learned the tangent planes $\mathcal{T}_{\mathbf{r}_z}$, $z = 1, \ldots, K$ such that each tangent plane represents the subset of images that belong to the same label. When a new test data $\mathbf{x}$ comes in, we can determine the label of the data by finding the most probable tangent plane. More specifically, we can find the tangent plane that yields the highest posterior probability given the test image $\mathbf{x}$:

$$\begin{aligned} z^* &= \underset{z=1,\ldots,K}{\operatorname{argmax}}\, p(z|\mathbf{x}) \\ &= \underset{z=1,\ldots,K}{\operatorname{argmax}}\, p(z|\mathbf{v}) \\ &= \underset{z=1,\ldots,K}{\operatorname{argmax}}\, \frac{p(\mathbf{v}|z)p(z)}{\sum_{z=1}^K p(\mathbf{v}|z)p(z)} \\ &= \underset{z=1,\ldots,K}{\operatorname{argmax}}\, p(\mathbf{v}|z)p(z) \end{aligned}$$

where $p(\mathbf{v}|z)$ is normal distribution with mean $\mu_z + W_z \boldsymbol{\alpha}_z$ and $p(z)$ is the learned prior.

4. Experiments

4.1. Datasets

We test how accurately the tangent plane approximation represents the image manifold on four datasets: MNIST, FERET, ADNI PET, and the Thyroid Nuclei dataset.

MNIST digits: MNIST dataset [13] consists of 70,000 images of 10 digits (0-9) (of size 28×28). In the subsequent experiment, we randomly selected a subset of MNIST dataset, 600 images per each digit and 6000 images in total.

FERET face images: The FERET dataset [14, 15] consists of face images photographed from different angles. For the experiment, frontal views were selected, and cropped and aligned apriori, in total we used 2137 images (of size 130×160).

ADNI PET Scans: Alzheimer's Disease Neuroimaging Initiative (ADNI) database [1] [8] was set up to define the progression of Alzheimer's disease, which includes MRI (Magnetic resonance imaging) images, PET (Positron emission tomography) images, genetics, cognitive tests, blood biomarkers, etc. The single axial slice from 18F-florbetapir brain PET volumes were used for the experiment. The dataset consists of 264 images (of size 218×182) which are labeled either as Amyloid positive or negative.

Thyroid Nuclei images: The Thyroid Nuclei dataset consists of segmented thyroid nuclei [3] from 47 patients with two types of follicular lesions: follicular adenoma (FA, 27 patients) and follicular carcinoma (FTC, 20 patients) tissue blocks, which were obtained from the archives of the University of Pittsburgh Medical Center. The dataset consists of a total of 500 nuclei images (of size 192×192), either labeled as FA or FTC based on its tissue block.

[1] http://adni.loni.usc.edu

-	MNIST	FERET	ADNI	NUCLEI
K	10	20	4	4
# tangent planes	10	20	4	2
d	784	20800	39676	36864
ℓ_z	9	20	40	60

Table 1: Number of clusters and tangent planes

4.2. Finding the tangent planes

All aforementioned image datasets are chosen with the consideration that i) images consist of different classes (i.e. digits, identity of face, malign vs benign cells, Amyloid positive vs negative brains) and ii) that the same classes of images are more likely to be deformed from the shared class templates (i.e. digit '2' is highly likely to be deformed from another digit '2' not digit '3').

The templates of each class are found via K-means clustering. When label information is present (which is true for all except the FERET dataset) the K-means clustering with LOT distance is performed within the class so that the templates are learned per class not jointly. For example, $K = 2$ in the ADNI dataset, and therefore in total 4 tangent planes (2 classes $\times$ 2 clusters) are used to represent the image manifold. For the MNIST dataset, $K = 10$, and the mean image was computed and used as a template image. The number of tangent planes, the number of clusters (K), the dimension for the image space (d), and the dimension for the tangent planes (ℓ_z) are summarized in Table 1. Across all datasets, $\ell_z \ll d$, implying that a d dimensional image space can be represented with much lower ℓ_z dimensional tangent planes.

Once the templates are found, the deformation maps between each image and the templates are learned. The templates serve as tangent planeial points of the tangent planes, and eigenvectors of deformation maps will represent the tangent planes. For example, sample eigenvectors computed from the deformation maps are shown in Fig. 2. The direction of the arrows indicates where the masses (pixels) are being transport to in the image and from the template,

and the length of the arrows represents the amount of the masses being transported (the longer the arrows, the larger the amount of masses transported).

We sought to validate the proposed method by applying the method to solve two common problems: synthesizing more images and classification. We envision that the method would be especially advantageous when only small number of images are available. We also compared the synthesized images to the conventional way of augmenting data and applied it to train complex classifiers.

For the MNIST dataset which contains about 70k images, learning such templates and deformation maps can be carried out easily. We note that for ADNI PET and Thyroid Nuclei dataset, however, with much fewer available images compared to the dimension of each image, learning such templates and deformation maps becomes non-trivial. We show here that the proposed method is capable of generating new unseen images without requiring massive datasets and that the method can extend to the Bayesian classification method, which both accentuate the benefits of the study in applications where collecting large datasets is impractical or unlimited number of synthetic images are desired.

4.3. Synthesizing Unseen Images

Once the image manifold is modeled with tangent planes, we can sample synthetic images. Specifically, the tangent plane the new synthetic image belongs to will be determined according to its prior p_z. Then, $\boldsymbol{\alpha}$ will be drawn out from normal distribution as in (6) to determine where in tangent plane the synthetic image will be located at. Each point in the tangent plane has a correspondence with a real image, and therefore, we can generate a synthetic image by mapping from the tangent space to the image space.

The synthesized images for MNIST, FERET, ADNI, and Thyroid datasets are shown in Fig. 3.

For the MNIST dataset, the template images are shown in the first row. The synthesized images in rows 2-15 are generated by deforming the template images. For the FERET dataset, the templates are shown in the top row, and synthetic 'unseen' faces are generated by transporting masses (pixels) from template images. It is interesting to see that synthetic deformation maps are capable of generating new faces with a variety of facial expressions (smile with visible teeth, grin, frown, neutral, etc.) and facial identities (different shapes of eyebrows, eyes, and nose, presence of mustache, size of cheek bones and jaws, etc.).

For the ADNI dataset and thyroid nuclei dataset, we visualized both real and synthetic images images to help readers understand that visually there is barely any difference between true images and synthesized images.

4.3.1 How are we sure that synthesized images are not sampled from the training set?

Here we repeated the same experiment for FERET dataset, but this time with only **19 images**. We performed this exercise to make sure that synthetic images in Fig. 3 are not copies of the existing 2137 images. By reducing the training set to 19 images, we could confirm i) that the synthesized images are not replicates of existing images, and ii) that the method can synthesize richer data given a small number of training samples. Fig. 4 shows 19 real images used for the experiment (top row) and 38 synthesized images (rows 2-3). Although artifacts are noticeable due to drastically reduced initial training samples (i.e. blurred nose), the method is capable of creating images with a variety of facial expressions.

4.3.2 Comparison with synthesizing image with PPCA

In order to visualize how PPCA modeling with the Euclidean distance would work out we performed the same experiment of synthesizing images with the Euclidean distance instead of on the OT manifold. Fig. 5 shows 'unseen' images created by eigenvectors. As expected, linearly combining eigenvectors doesn't generate reasonable images because images do not lie on a linear Euclidean subspace.

4.4. Data augmentation for training CNN

Here, we tested whether enlarging the datasets with our proposed method can facilitate learning complex systems such as a convolutional neural network (CNN). For our task,

-	MNIST	ADNI PET	NUCLEI
# Train	80	211	400
# Test	20	53	100
Default	89.90%	94.34%	77.50%
/w Jittering	99.00%	92.45%	80.00%
/w PCA	95.00%	93.40%	84.50%
# Synthesized Train	900	200	500
/w added train set	**100.00**%	**94.86**%	**85.00**%

Table 2: Classification Accuracy with and without the data augmentation for CNN classifier

-	MNIST	ADNI PET	NUCLEI
# Train	800	211	400
# Test	200	53	100
Logistic Regression	89.00%	92.06%	70.20%
Bayesian Classification	**97.00**%	**96.23**%	**72.00**%

Table 3: Classification Accuracy with Bayesian Classifier

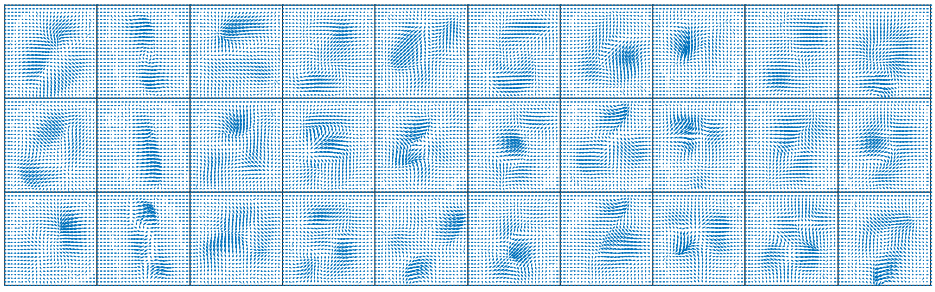

Figure 2: Sample Eigenvectors (from deformation maps) of Digit 0-9

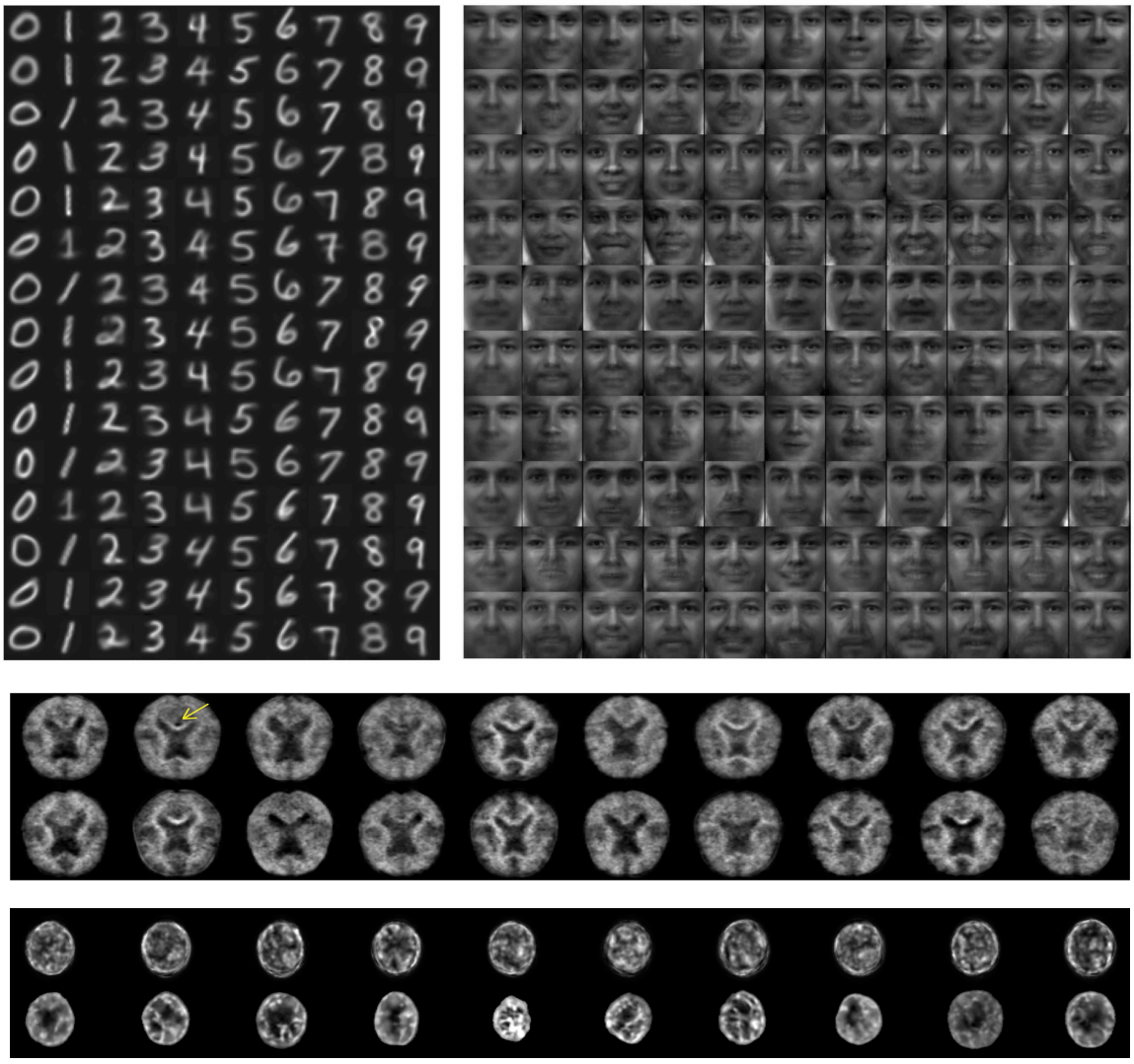

Figure 3: Synthesized images of MNIST (left top) and FERET (right top) images, synthesized and true images of ADNI pet scans and Thyroid nuclei images. For MNIST and FERET images, the top row shows the template images. For ANDI pet scans and Thyroid nuclei images, top row shows the synthesized images whereas the bottom row shows the true images.

a CNN with two convolutional (conv.) layers and two dense layers was set up. The first conv. layer consists of 48 fil-

Figure 4: 'Synthesized' faces (rows 2-3) generated only using 19 Real Faces (top row).

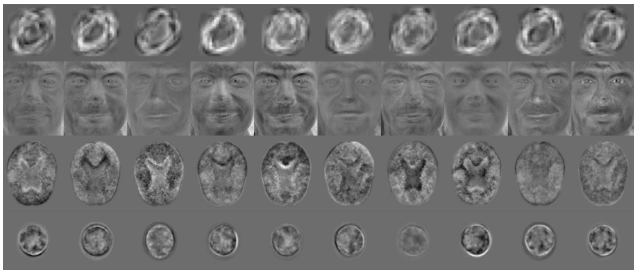

Figure 5: 'Unseen' Images generated by PPCA using Euclidean distance

ters, the second conv. layer consists of 96 filters, and the dense layer consists of 100 filters. ReLu activation layers follow conv. and dense layers, except for the final dense layer which has softmax output instead (or sigmoid for binary classification). The conv. layers' filters configuration is identical to that of Alexnet [12]. The Alexnet architecture was utilized specifically in favor of using its pretrained weights. Pretrained weights of Alexnet was loaded into our smaller network by arbitrarily choosing 48 (conv1) - 96 (conv2) filter weights out of 96 (conv1) -256 (conv2) filter weights.

For the MNIST dataset, the sample size was reduced to 100 images to emphasize the effect of how the proposed method can facilitate better learning for complex classifiers.

Table 2 shows the classification accuracy with and without synthetic data, as well as conventional data-jittering (with translation, rotations, and shear transformation) method. The testing accuracy consistently improved with adding the synthesized data set. For the MNIST dataset, with 100 initial training samples and with 900 synthetic samples added, test accuracy reached 100%. For ADNI and Thyroid, adding synthetic images does not harm nor benefit the classification, therefore suggesting that synthesized images closely reflect the original images and therefore do not provide additional useful discriminant information for classifiers to utilize.

4.5. Bayesian Classification

The images are generated from a probabilistic model which gives a measure of how likely an image belongs to a tangent plane. Here we test how our model can aid binary classification for small datasets. More precisely, once the tangent planes and latent priors are learned, for each new test image, we can find the label by associating it to the tangent plane that the image is most likely generated. The details on our Bayesian classifier described in Sec. 3.4, and Table 3 show the Bayesian classification accuracy for MNIST, ADNI, and Thyroid dataset. We note that for ADNI images, this is the current best reported classification accuracy.

5. Conclusion

In this paper, we proposed to represent data with templates and diffeomorphic maps uniquely identified with Wasserstein-Kantorovich cost. Regardless of the size of the dataset, if the images share common templates, we showed that images can be represented in tangent planes and provide alternative representation of the dataset that can be utilized in synthesizing images and augmenting datasets for complex classifier training. In addition, we used a probabilitic framework by assigning each tangent plane with a latent variable, and formulated the Bayesian classifier which is demonstrated to be suitable for a dataset sharing common templates.

However, we dot believe our methodology would generalise to non-structural images, e.g. uncategorized natural images. Our method inherently assumes that images are deformed from 'template' images, and generalizing to non-structural images would require either different assumptions on the data distribution or a much larger dataset.

Nevertheless, modeling in tangent planes with optimal transport maps produces realistic local variations compared to using diffeomorphisms or Euclidean geometry. Specifically, although the optimal transport manifold is modeled as locally linear, variations in tangent planes corresponds to highly non-linear variations in the image space. We anticipate a future study to include generalizing the FA model to a fully Bayesian model and assuming different distribution for the data, expanding the method for non structured images, applying synthetic images to solve inverse estimation problems [19], generating ground truth data for quantifying accuracy of image analysis operation, and generating new samples for simulation based training.

6. Acknowledgements

MT would like to thank the Cantab Capital Institute for the Mathematics of Information at the University of Cambridge for their support during the preparation of the manuscript.

References

[1] M. Bauer, M. Bruveris, and P. W. Michor. Overview of the geometries of shape spaces and diffeomorphism groups. *Journal of Mathematical Imaging and Vision*, 50(1-2):60–97, 2014. 1

[2] E. Boissard, T. Le Gouic, J.-M. Loubes, et al. Distributions template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015. 1

[3] C. Chen, W. Wang, J. A. Ozolek, and G. K. Rohde. A flexible and robust approach for segmenting cell nuclei from 2d microscopy images using supervised learning and template matching. *Cytometry Part A*, 83(5):495–507, 2013. 5

[4] M. do Carmo. *Riemannian Geometry*. Birkhäuser Basel, translated from the second Portuguese edition by F. Flaherty edition, 1992. 2

[5] P. T. Fletcher, C. Lu, and S. Joshi. Statistics of shape via principal geodesic analysis on lie groups. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003. 1

[6] S. Hauberg, O. Freifeld, A. B. L. Larsen, J. W. Fisher III, L. K. Hansen, L. Pudwell, E. Rowland, R. Steinert, W. John, P. Sköldström, et al. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. *arXiv preprint arXiv: 1510.02795*, 2015. 2

[7] S. Huckemann and H. Ziezold. Principal component analysis for riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 38(2):299–319, 2006. 1

[8] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of magnetic resonance imaging*, 27(4):685–691, 2008. 5

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *NIPS*, 2013. 2

[10] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017. 1

[11] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015. 1

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 8

[13] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, 1998. 5

[14] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000. 5

[15] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998. 5

[16] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 2

[17] S. Rifai, Y. Dauphin, P. Vincent, Y. Bengio, and X. Muller. The manifold tangent classifier. In *NIPS*, volume 271, page 523, 2011. 1

[18] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pages 3312–3320, 2015. 1

[19] A. Shariff, R. F. Murphy, and G. K. Rohde. A generative model of microtubule distributions, and indirect estimation of its parameters from fluorescence microscopy images. *Cytometry Part A*, 77(5):457–466, 2010. 8

[20] P. Simard, B. Victorri, Y. LeCun, and J. Denker. Tangent prop - a formalism for specifying selected invariances in an adaptive network. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 895–903. Morgan-Kaufmann, 1992. 1, 2

[21] R. Sparks and A. Madabhushi. Novel morphometric based classification via diffeomorphic based shape representation using manifold learning. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 658–665, 2010. 1

[22] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. 3

[23] M. Vaillant, M. I. Miller, L. Younes, and A. Trouvé. Statistics on diffeomorphisms via tangent space representations. *NeuroImage*, 23:S161–S169, 2004. 1

[24] C. Villanu. *Topics in optimal transportation*. American Mathematical Soc., 2003. 2

[25] L. Wang, F. Beg, T. Ratnanather, C. Ceritoglu, L. Younes, J. C. Morris, J. G. Csernansky, and M. I. Miller. Large deformation diffeomorphism and momentum based hippocampal shape discrimination in dementia of the alzheimer type. *IEEE transactions on medical imaging*, 26(4):462–470, 2007. 1

[26] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013. 1, 3

[27] M. Zhang and P. T. Fletcher. Bayesian principal geodesic analysis in diffeomorphic image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 121–128. Springer, 2014. 1