Bayesian Principal Geodesic Analysis for Estimating Intrinsic Diffeomorphic Image Variability

Miaomiao Zhang, P. Thomas Fletcher

Scientific Computing and Imaging Institute, University of Utah

Abstract

In this paper, we present a generative Bayesian approach for estimating the low-dimensional latent space of diffeomorphic shape variability in a population of images. We develop a latent variable model for principal geodesic analysis (PGA) that provides a probabilistic framework for factor analysis in the space of diffeomorphisms. A sparsity prior in the model results in automatic selection of the number of relevant dimensions by driving unnecessary principal geodesics to zero. To infer model parameters, including the image atlas, principal geodesic deformations, and the effective dimensionality, we introduce an expectation maximization (EM) algorithm. We evaluate our proposed model on 2D synthetic data and the 3D OASIS brain database of magnetic resonance images, and show that the automatically selected latent dimensions from our model are able to reconstruct unobserved testing images with lower error than both linear principal component analysis (LPCA) in the image space and tangent space principal component analysis (TPCA) in the diffeomorphism space.

Keywords: Bayesian estimation, principal geodesic analysis, diffeomorphic image registration, dimensionality reduction

1. Introduction

The deformable template approach to statistical shape analysis of images is to quantify shape using deformable image registration and then compute statistics of the resulting transformations, rather than the images themselves. The first step in this process is to compute a template image, or atlas, which represents the large data set. The class of diffeomorphic transformations preserves the topology of objects in the images and provides forward and inverse mappings between the atlas and individuals. Furthermore, the requisite distance metric for atlas estimation and further statistics is provided by the setting of *Large Deformation Diffeomorphic Metric Mapping* (LDDMM) (Beg et al., 2005). Motivated by Bayesian reasoning, current approaches (Joshi et al., 2004; Twining et al., 2005; Ma et al., 2008; Vialard et al., 2011) formulate diffeomorphic atlas building as *maximum a posteriori* (MAP) optimization problems, where the image match is analogous to a log-likelihood, and the deformation regularization is analogous to a log-prior.

However, an image atlas is only a point estimate and does not encode the shape variability of a population. Extracting low-dimensional, second-order statistics of anatomical shape variability is a critical step to improve the statistical power and interpretability of further statistical analyses. The standard method for conducting dimensionality reduction and analyzing variability of Euclidean data is principal component analysis (PCA), which decomposes the data matrix into a linear combination of independent factors. Bishop (1999) introduced a Bayesian model for PCA (BPCA) that automatically learns the dimension of the latent space from data by including a sparsity-inducing prior on each component of the factor matrix. These linear factor analysis models, nevertheless, are not directly applicable to nonlinear diffeomorphic transformations.

There exist several methods for dimensionality reduction and shape variability modeling on nonlinear manifolds. Fletcher et al. (2003) generalized PCA to finite-dimensional manifolds, in a method called principal geodesic analysis (PGA), which estimates lower-dimensional geodesic subspaces by minimizing the sum-of-squared geodesic distances to the data. Based on this work, Said et al. (2007); Sommer et al. (2010) developed algorithms for exact solutions to PGA. In order to allow factor analysis on manifolds, Zhang and Fletcher (2013) recently introduced a probabilistic

model for PGA (PPGA). In the setting of diffeomorphic image registration, Vaillant et al. (2004) computed a tangent space PCA (TPCA) of the initial momenta from the atlas image. Later, Qiu et al. (2012) used TPCA as an empirical shape prior in diffeomorphic surface matching. A Bayesian model of shape variability using diffeomorphic matching of currents is also formulated by Gori et al. (2013). Their model includes an estimation of a covariance matrix of the deformations, from which they then extracted PCA modes of shape variability. Even though these methods formulate the atlas and covariance estimation as probabilistic inference problems, the dimensionality reduction is done after the fact, i.e., as a singular value decomposition of the covariance as a second stage after the estimation step.

We propose instead to treat the dimensionality reduction step as a probabilistic inference problem on discrete images, in a model called Bayesian principal geodesic analysis (BPGA), which jointly estimates the image atlas and principal geodesic modes of variation. Our model goes beyond the PPGA algorithm by introducing automatic dimensionality reduction, as well as extending from finite-dimensional manifolds to the infinite-dimensional case of diffeomorphic image registration. This Bayesian formulation has two advantages. First, it explicitly optimizes the fit of the principal modes to the data intrinsically in the space of diffeomorphisms, which results in better fits to the data. Second, by formulating dimensionality reduction as a Bayesian model with a sparsity prior, we can also infer the inherent dimensionality directly from the data.

This paper is an extension of Zhang and Fletcher (2014), with three major differences: (1) we incorporate a stronger sparsity prior, based on the adaptive sparsity method of (Figueiredo, 2003) that avoids the need for hyperparameters; (2) we provide more in-depth derivations of the statistical model and inference procedure; and (3) we expand the experimental results on brain MRI from 2D slices to full 3D volumes. We also mention the relationship of our work to manifold learning approaches and dimensionality reduction (Yan et al., 2007; Gerber et al., 2010). Unlike the non-parametric manifold learning methods, the Bayesian approach we present here is parametric and fully *generative*. The shape deformation of individuals is explicitly encoded in the model, and can be reconstructed directly in a compact space of principal modes of deformations. We show experimental results of principal geodesics and parameters estimated from both 2D synthetic data and 3D OASIS brain MRI data. To validate the advantages of our model, we reconstruct images from our estimation and compare the reconstruction errors with TPCA of diffeomorphisms and LPCA based on image intensity. Our results indicate that intrinsic modeling of the principal geodesics, estimated jointly with the image atlas, provides a better description of brain image data than computing PCA in the tangent space after atlas estimation.

2. Background

In this section, we briefly review the mathematical background for diffeomorphic atlas building. Throughout, we will consider images to be square-integrable functions defined on a d-dimensional torus domain $\Omega = \mathbb{R}^d/\mathbb{Z}^d$, that is, an image is an element of $L^2(\Omega, \mathbb{R})$. Let $T\Omega$ define the tangent space of Ω, $\tilde{V} = H^s(T\Omega)$ denote the Hilbert space of vector fields on Ω whose derivatives up to order s exist and are square-integrable. As is standard, we require that $s > (d/2) + 1$, so that V embeds continuously in $C^1(T\Omega)$, the space of continuous vector fields with continuous first derivatives. We will consider diffeomorphisms that are generated by flows of time-varying velocity fields from $\tilde{V}$. More specifically, consider a time-varying velocity field, $v_t : [0, 1] \to \tilde{V}$, then we may define the flow $t \mapsto \phi_t \in \mathrm{Diff}^s(\Omega)$ as a solution to the equation

$$\frac{d\phi_t}{dt}(x) = v_t \circ \phi_t(x). \tag{1}$$

The space of all diffeomorphisms generated in this fashion will be denoted $\mathrm{Diff}^s(\Omega)$. We use subscripts for the time variable, i.e., $v_t(x) = v(t, x)$, and $\phi_t(x) = \phi(t, x)$.

2.1. Metrics on diffeomorphisms

A key ingredient in computational anatomy is the notion of a distance metric on the space of diffeomorphisms. Such a metric provides a means for quantifying the magnitude of the deformation between two images, and forms the mathematical foundation for estimation of statistical models, such as atlases, as least-squares minimization problems. The first step is to define an inner product on the space of velocities, $V = T_e\mathrm{Diff}^s(\Omega)$, identified with the tangent space at the identity transform, $e \in \mathrm{Diff}^s(\Omega)$. This inner product is of the form

$$\langle v, w \rangle_V = \int_\Omega \langle Lv(x), w(x) \rangle dx,$$

for $v, w \in V$, and a symmetric, positive-definite differential operator $L : V \to V^*$, mapping to the dual space, V^*. In this paper, we use $L = (-\alpha\Delta + I)^c$, for some constant $\alpha > 0$ and integer power c. The dual to the vector v is a momentum, $m \in V^*$, such that $m = Lv$ and

$v = Km$, where K is the inverse of L. Now we can define a right-invariant metric as an inner product at any other point $\phi \in \mathrm{Diff}^s(\Omega)$, by pulling back the velocities at ϕ to the identity by right composition. In other words, for $\tilde{v}, \tilde{w} \in T_\phi \mathrm{Diff}^s(\Omega)$ the right-invariant metric is given by $\langle \tilde{v}, \tilde{w} \rangle_{T_\phi \mathrm{Diff}^s(\Omega)} = \langle \tilde{v} \circ \phi^{-1}, \tilde{w} \circ \phi^{-1} \rangle_V$.

Given this definition of a right-invariant metric, we can now define a geodesic curve in $\mathrm{Diff}^s(\Omega)$ as a flow that minimizes the energy

$$E(\phi_t) = \int_0^1 \left\| \frac{d\phi_t}{dt} \circ \phi_t^{-1} \right\|_V^2 dt.$$

The geodesics that minimize this energy are characterized by the Euler-Poincaré equations (EPDiff) (Arnol'd, 1966; Miller et al., 2006),

$$\begin{aligned} \frac{\partial v}{\partial t} &= -\mathrm{ad}_v^\dagger v = -K \mathrm{ad}_v^* m \\ &= -K \left[(Dv)^T m + Dm\, v + m \,\mathrm{div}\, v \right], \end{aligned} \quad (2)$$

where D denotes the Jacobian matrix. The operator ad^* is the dual of the negative Lie bracket of vector fields,

$$\mathrm{ad}_v w = -[v, w] = Dvw - Dwv.$$

Given an initial velocity, $v_0 \in V$, at $t = 0$, the EPDiff equation (2) can be integrated forward in time, resulting in a time-varying velocity $v_t : [0, 1] \to V$, which itself is subsequently integrated in time by the rule $(d\phi_t/dt) = v_t \circ \phi_t$ to arrive at the geodesic path, $\phi_t \in \mathrm{Diff}^s(\Omega)$. This process is known as *geodesic shooting*.

2.2. Diffeomorphic atlas building

Given input images, $J^1, \ldots, J^N \in L^2(\Omega, \mathbb{R})$, a minimization of the sum-of-squared distance function is solved to estimate the atlas, $I \in L^2(\Omega, \mathbb{R})$ and the diffeomorphic transformations between the atlas and each input image as

$$E(v_t^k, I) = \sum_{k=1}^{N} \frac{1}{2\sigma^2} \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2 + \int_0^1 \left\| v_t^k \right\|_V^2 dt, \quad (3)$$

where σ^2 represents image noise variance, and the tangent vectors $\{v_t^k \in L^2([0, 1], V)\}_{k=1\ldots N}$ are time-varying velocity fields in a reproducing kernel Hilbert space, V, equipped with the metric, L. The deformation ϕ^k is defined in (1) as the integral flow of v_t^k with $\phi_0^k = Id$.

Because the distance function between images is itself a minimization problem, the atlas estimation is typically done by alternating between the minimization to find the optimal v_t^k and the update of the atlas, I. As described in Beg et al. (2005), a variational scheme is applied to simulate the evolution of velocity v_t^k and diffeomorphism ϕ_t^k at each discretized time point using gradient descent. This approach requires a large amount of memory to store the entire sequence of time-varying velocities and diffeomorphisms. To resolve this issue, Vialard et al. (2012) instead estimate only the initial velocity, v_0^k, by geodesic shooting. This requires backward integration of adjoint equations to carry gradients of the image match at the endpoint $t = 1$ back to a gradient of the initial velocity at time $t = 0$.

2.3. Decoupling images from diffeomorphisms

As shown in (Miller et al., 2006), at an optimal solution to (3), the initial momenta, $m_0^k = Lv_0^k$, are orthogonal to the level sets of the atlas. Therefore, each initial momentum m_0^k is typically represented as a scalar field P^k multiplied by the gradient of the atlas, i.e., $m_0^k(x) = \nabla I(x) P^k(x)$. This has a major disadvantage while solving the optimization problem: the coupled estimation of atlas and momenta leads to a poor convergence performance.

Singh et al. (2013) proposed to decouple the estimation of the atlas from momenta by optimizing in the full space of vector momenta, rather than restricting to scalar multiples of the image gradient. They demonstrated that this approach obtains better convergence rates and numerical stability. The vector momenta formulation also results in closed-form updates for the optimal atlas. However, perhaps the most important advantage of using the vector momenta formulation is that decoupling images from diffeomorphisms allows us to treat the diffeomorphic transformations as unobserved random variables, separate from the observed image data. Based on this critical factor, Zhang et al. (2013) treat diffeomorphisms as latent random variables in a Bayesian model and integrate them out using Monte Carlo. This provides estimation of the regularization parameter (i.e., the α in the L operator) and the image noise variance, σ^2.

3. Bayesian Principal Geodesic Analysis

Before introducing our BPGA model for diffeomorphisms, we first review BPCA Bishop (1999) for Euclidean data. The main idea of BPCA is to formulate a generative latent variable model for PCA that automatically selects the appropriate dimensionality of the model. Consider a set y of n-dimensional Euclidean random variables $\{y_j\}_{j=1,\ldots,N} \in \mathbb{R}^n$, the relationship between

each variable y_j and its corresponding q-dimensional ($q < n$) latent variable x_j is

$$y_j = \mu + Bx_j + \epsilon, \tag{4}$$

where μ is the mean of dataset $\{y_j\}$, x_j is conventionally defined as a random variable generated from $N(0, I)$, B is an $n \times q$ factor matrix that relates x_j and y_j, and $\epsilon \sim N(0, \sigma^2 I)$ represents error. This definition gives a data likelihood as

$$p(y \mid x; B, \mu, \sigma) \propto \prod_{j=1}^{N} \exp\left(-\frac{\left\|y_j - \mu - Bx_j\right\|^2}{2\sigma^2}\right).$$

To automatically select the principal components from data, BPCA includes a Gaussian prior over each column of B, which is known as an automatic relevance determination (ARD) prior. Each such Gaussian has an independent variance associated with a precision hyperparameter γ_i, so that

$$p(B \mid \gamma) = \prod_{i=1}^{q} \left(\frac{\gamma_i}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\gamma_i B_i^T B_i\right),$$

where B_i denotes the ith column of B.

The value of γ_i is estimated iteratively as $\frac{n}{\|B_i\|^2}$ in this model, and thus enforces sparsity by driving the corresponding component B_i to zero. More specifically, if γ_i is large, B_i will be effectively removed in the latent space. This arises naturally because the larger γ_i is, the lower the probability of B_i will be. Notice that the columns of B define the principal subspace of standard PCA, therefore, inducing sparsity on B has the same effect as removing irrelevant dimensions in the principal subspace.

3.1. Probability Model

Likelihood. We formulate the random initial velocity for the kth individual as $v_0^k = Wx^k$, where W is a matrix with q columns of principal initial velocities, and $x^k \in \mathbb{R}^q$ is a latent variable that lies in a low-dimensional space, with

$$p(x^k \mid W) \propto \exp\left(-\frac{1}{2}\left\|Wx^k\right\|_V^2\right). \tag{5}$$

Compared to BPCA, the difference of this latent variable prior is incorporating W as a conditional probability, which guarantees smoothness of the geodesic shooting path. Notice that we shift from the momenta space Zhang and Fletcher (2014) to a nicely smooth velocity space, which gains more stable computations.

Our noise model is based on the assumption of i.i.d. Gaussian at each image voxel, much like (Ma et al., 2008; Qiu et al., 2012; Zhang et al., 2013). This can be varied under different conditions, for instance, spatially dependent model for highly correlated noise data. In this paper, we will focus on the commonly used and simple Gaussian noise model, with the likelihood given by

$$p(J^k \mid I, \sigma, x^k) = \frac{1}{(2\pi)^{M/2}\sigma^M} \exp\left(-\frac{\left\|I \circ (\phi^k)^{-1} - J^k\right\|_{L^2}^2}{2\sigma^2}\right), \tag{6}$$

where M is the number of voxels, and the norm inside the exponent is the $L^2(\Omega, \mathbb{R})$ norm. Note that for a continuous image domain, $\Omega = \mathbb{R}^d/\mathbb{Z}^d$, this is not a well-defined probability distribution due to its infinite measure in the Hilbert space $L^2(\Omega, \mathbb{R})$ on images. Therefore, we consider the input images as well as diffeomorphisms to be defined on a finite discretized grid.

Prior. The prior on W is a sparsity prior that suppresses the small principal initial velocity to zero. This prior is analogous to the hierarchical sparsity prior proposed by Figueiredo (2003), with the difference that we use the natural Hilbert space norm for the velocity. The prior is based on Laplacian distribution, a widely used and exploited way to achieve sparse estimation. It presses the irrelevant or redundant components exactly to zero. As first introduced by Andrews and Mallows (1974), the Laplace distribution is equivalent to the marginal distribution of a hierarchical-Bayes model: a Gaussian prior with zero mean and exponentially distributed variances. Let i denote the ith principal component of W. We define each component W_i as a random variable with the hierarchical model distribution

$$p(W_i \mid \tau_i) \sim N(0, \tau_i),$$
$$p(\tau_i \mid \gamma_i) \sim \mathrm{Exp}(\frac{\gamma_i}{2}),$$

After integrating out τ_i, we have the marginalized distribution as

$$p(W_i \mid \gamma_i) = \int_0^\infty p(W_i \mid \tau_i) p(\tau_i \mid \gamma_i) d\tau_i = \frac{\sqrt{\gamma_i}}{2} \exp\left(-\sqrt{\gamma_i}\,\|W_i\|_1\right),$$

which is a Laplacian distribution with scale parameter $\gamma_i/2$. The degree of sparsity is controlled by the hyperparameter γ_i on the l_1 penalty. However, the sparsity parameter is specified in an ad hoc manner. Figueiredo (2003) proposed an effective model to remove γ_i by adopting a Jeffreys' noninformative hyperprior as $p(\tau_i) \sim 1/\tau_i$. This has the advantages that (1) the improper hyperprior is scale-invariant, (2) the model is parameter-free. Using this hierarchical sparsity prior

on the columns of W for the automatic dimensionality selection, we formulate the problem as

$$p(W, x \mid \tau) \propto \exp\left(-\frac{1}{2}\sum_{k=1}^{N}\left\|Wx^k\right\|_V^2 - \sum_{i=1}^{q}\frac{\|W_i\|_V^2}{2\tau_i}\right), \quad (7)$$
$$p(\tau) \propto \frac{1}{\tau},$$

where $x = [x^1, ..., x^k]$, $\tau = [\tau_1, ..., \tau_q]$. We will later integrate out the latent variable τ using expectation maximization.

We can express our model for the kth subject using the graphical representation shown in Figure 1.

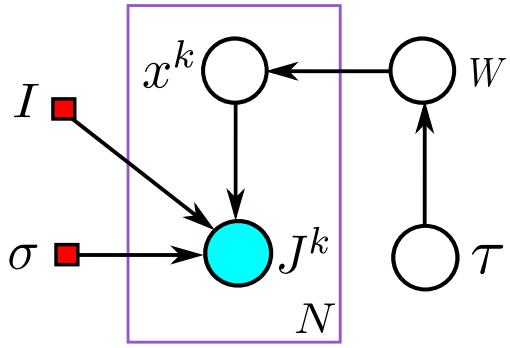


Figure 1: Graphical representation of BPGA for the kth subject J^k.

3.2. Inference

We use MAP estimation to determine the model parameters $\theta = \{I, \sigma\}$. After defining the likelihood (6) and prior (7) in the previous section, we now arrive at the joint posterior for BPGA as

$$\prod_{k=1}^{N} p\left(W, x, \tau \mid J^k; \theta\right) \propto \left[\prod_{k=1}^{N} p(J^k \mid x^k, \theta)\, p(x^k \mid W)\right] p(W \mid \tau)\, p(\tau). \quad (8)$$

In order to treat the W, x^k and τ as latent random variables with the log posterior given by (8), we would ideally integrate out the latent variables, which are intractable in closed form for W, x^k. Instead, we develop an expectation maximization algorithm to compute a closed-form solution to integrate out τ first, and then use a mode approximation for W, x^k to the posterior distribution. It contains two alternating steps:

E-step. Using the current estimate of the parameters $\hat{\theta}$, we compute the expectation Q of the complete log-posterior of (8) with respect to the latent variables τ as

$$Q(W, x^k, \theta \mid \hat{\theta}, \hat{W}) \propto -\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left\|I \circ (\phi^k)^{-1} - J^k\right\|_{L^2}^2 - \frac{MN}{2}\log\sigma$$
$$- \frac{1}{2}\sum_{k=1}^{N}\left\|Wx^k\right\|_V^2 - \sum_{i=1}^{q}\frac{\|W_i\|_V^2}{2\|\hat{W}_i\|_V^2}. \quad (9)$$

Note that we use the same approach to integrate out τ in Figueiredo (2003). Details are in Appendix A.

M-step: Gradient Ascent for W, x^k. We introduce a gradient ascent scheme to estimate W, x^k, and $\theta = (I, \sigma)$ simultaneously. We need to compute the gradient with respect to the initial velocity v_0^k of the diffeomorphic image matching problem in (3), and then apply the chain rule to obtain the gradient term w.r.t. W and x^k. Following the optimal control theory approach in Vialard et al. (2012), we add Lagrange multipliers to constrain the kth diffeomorphism ϕ_t^k to be a geodesic path, which is done by introducing time-dependent adjoint variables, $\hat{I}_t^k$, $\hat{m}_t^k$, $\hat{v}_t^k$ for transported image I_t^k, momentum m_t^k, and velocity v_t^k, respectively. To make the calculation simple to read, we drop the notation t and denote $\partial_t f$ as $\dot{f}$ for any function f. We then write the augmented energy

$$\tilde{Q}(W, x^k, \theta \mid \hat{\theta}, \hat{W}) = Q + \sum_{k=1}^{N}\int_0^1 \Big[\langle \hat{v}^k, \dot{v}^k + K\mathrm{ad}^*_{v^k} m^k\rangle_{L^2}$$
$$+ \langle \hat{I}^k, \dot{I}^k + \nabla I^k \cdot v^k\rangle_{L^2} + \langle \hat{m}^k, m^k - Lv^k\rangle_{L^2}\Big] dt, \quad (10)$$

where Q is the expectation function from (9), and the other terms correspond to Lagrange multipliers enforcing: a) the geodesic constraint, which comes from the EPDiff equation (2), b) the image transport equation, $\dot{I}^k = -\nabla I^k \cdot v^k$, and c) the constraint, $m^k = Lv^k$ respectively.

Dropping out the terms that are not related to W, x^k and I_0 in (10), we have

$$\tilde{Q}(W, x^k, \theta \mid \hat{\theta}, \hat{W}) \propto -\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left\|I \circ (\phi^k)^{-1} - J^k\right\|_{L^2}^2 - \frac{1}{2}\sum_{k=1}^{N}\left\|Wx^k\right\|_V^2$$
$$- \sum_{i=1}^{q}\frac{\|W_i\|_V^2}{2\|\hat{W}_i\|_V^2} + \sum_{k=1}^{N}\int_0^1 \Big[\langle \hat{v}^k, \dot{v}^k + K\mathrm{ad}^*_{v^k} m^k\rangle_{L^2}$$
$$+ \langle \hat{I}^k, \dot{I}^k + \nabla I^k \cdot v^k\rangle_{L^2} + \langle \hat{m}^k, m^k - Lv^k\rangle_{L^2}\Big] dt. \quad (11)$$

The gradient of $\tilde{Q}$ with respect to the kth initial velocity is $\nabla_{v_0^k}\tilde{Q} = v_0^k - K\hat{v}_0^k$ (details are in Appendix A). Applying the chain rule, the gradient term of (11) for updating W is

$$\nabla_W \tilde{Q} = -\sum_{k=1}^{N}(v_0^k - K\hat{v}_0^k)(x^k)^T - W\Lambda,$$

where Λ is a diagonal matrix with diagonal element $\frac{1}{\|\hat{W}_i\|_V^2}$. The gradient with respect to x^k is

$$\nabla_{x^k} \tilde{Q} = -W^T(v_0^k - K\hat{v}_0^k).$$

Closed-form solution for θ. We now derive the maximization for updating the parameters θ. This turns out to be a closed-form update for the atlas I, noise variance σ^2. For updating I and σ, we set the derivative of the expectation with respect to I, σ to zero (see Appendix A). The solution for I, σ^2 gives an update

$$I = \frac{\sum_{k=1}^N J^k \circ \phi^k |D\phi^k|}{\sum_{k=1}^N |D\phi^k|}, \quad \sigma^2 = \frac{1}{MN} \sum_{k=1}^N \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2.$$

4. Results

We demonstrate the effectiveness of our proposed model and MAP estimation routine using both 2D synthetic data and real 3D MRI brain data.

4.1. Synthetic data

Because we have a generative model, we can forward simulate a random sample of images from a distribution with known parameters $\theta = (I, \sigma)$. We tested if we can recover those parameters using our BPGA inference procedure. We simulated a 2D synthetic dataset with 40 subjects starting from an atlas image, I, of a binary circle with resolution 100×100. We then generated random samples of W with two principal modes and x^k from the prior distribution, $p(W, x^k \mid \tau)$, defined in (7), setting $\alpha = 0.2$ for the Laplacian operator L. To generate a deformed circle image, we shot the initial velocity constructed by Wx^k, and transformed the atlas by the resulting diffeomorphisms. Finally, we added i.i.d. Gaussian noise according to our likelihood model (6). We used a standard deviation of $\sigma = 0.05$, which corresponds to a SNR of 20 (which is more noise than typical structural MRI). Figure 2 compares the ground truth atlas I and principal geodesics with our estimation. In addition, our estimation of the noise variance $\sigma = 0.051$ is also close to the ground truth $\sigma = 0.05$. It shows that our method recovers the model parameters fairly well. Figure 3 demonstrates the shape variation of our synthetic data set from the atlas by $a_i = -3, -1.5, 0, 1.5, 3$.

4.2. OASIS brain dataset

To demonstrate the effectiveness of our proposed model and MAP estimation, we applied our BPGA model to a set of brain magnetic resonance images (MRI) from the 3D OASIS brain database. The data consists of MRI from 130 subjects between the age of 60 to 95. The MRI have a resolution of $128 \times 128 \times 128$ with an image spacing of $1.0 \times 1.0 \times 1.0$ mm^3 and are skull-stripped, intensity normalized, and co-registered with rigid transforms. To set the parameters in L operator, we did initial step of estimating $\alpha = 0.1$ using the procedure in Zhang et al. (2013). We used 15 time-steps in geodesic shooting and initialize the template I as the average of image intensities, with W as the matrix of principal components from TPCA.

The proposed BPGA model automatically determined that the latent dimensionality of the data was 15. Figure 4 displays the automatically estimated modes, $i = 1, 2$, of the brain MRI variation. We forward shoot the constructed atlas, I, by the estimated principal momentum $a_i W_i$ along the geodesics. For the purpose of visualization, we demonstrate the brain variation from the atlas by $a_i = -3, -1.5, 0, 1.5, 3$. We also show the log determinant of Jacobians at $a_i = 3$, with red representing regions of expansion and blue representing regions of contraction. The first mode of variation clearly shows that ventricle size change is a dominant source of variability in brain shape. Our algorithm also jointly estimated the image noise standard deviation parameter as $\sigma = 0.04$.

Image reconstruction accuracy. We validated the ability of our BPGA model to compactly represent the space of brain variation by testing how well it can reconstruct unseen images. After estimating the principal initial velocity and parameters from the training subjects above, we used these estimates to reconstruct another 20 testing subjects from the same OASIS database that were not included in the training. We then measured the discrepancy between the reconstructed images and the testing images. Note that our reconstruction only used the first 15 principal modes, which were automatically selected by our algorithm.

We use the first fifteen dimensions to compare our model with LPCA and TPCA. Examples of the reconstructed images and their error maps from these models are shown in Figure 5 and 6. Table 1 shows the comparison of the reconstruction accuracy as measured by the average and standard deviation of the mean squared error (MSE). The table indicates that our model outperforms both LPCA and TPCA in the diffeomorphic setting. We also display the reconstruction error with increasing number of principal modes. Figure 7 shows that TPCA requires approximately 32 principal modes, more than twice as much as our model does, to achieve the same level of reconstruction accuracy. LPCA cannot match the BPGA reconstruction accuracy with even 40

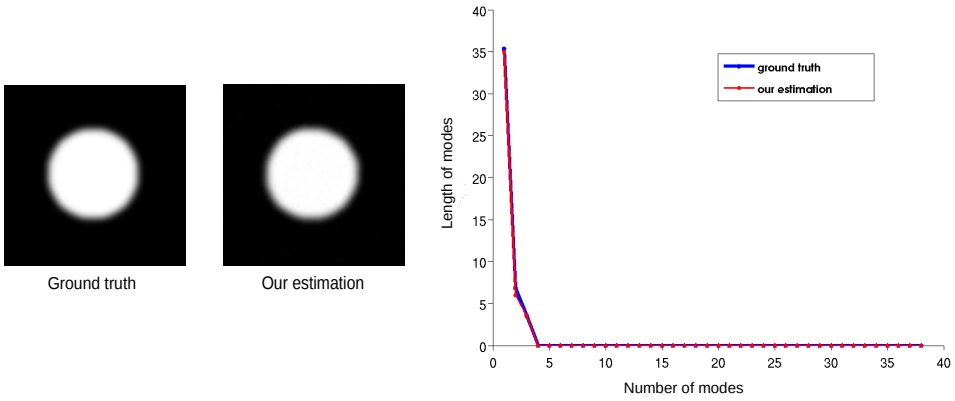


Figure 2: Left to right: ground truth of atlas I; our estimation of atlas; ground truth of the length of all principal geodesics and our estimation.

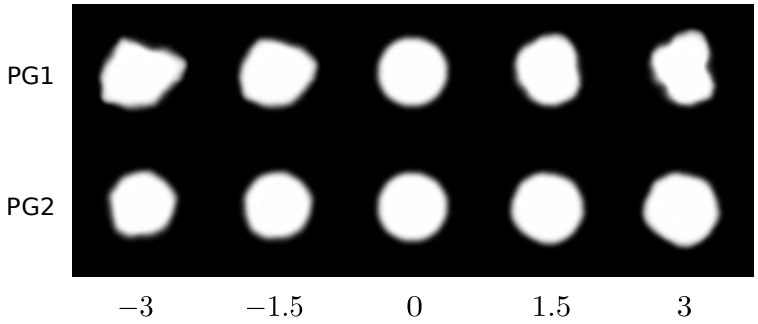


Figure 3: Top to bottom: shooting atlas by the first and second principal modes. Left to right: BPGA model of image variation evaluated at $a_i = -3, -1.5, 0, 1.5, 3$.

Table 1: Comparison of mean squared reconstruction error between LPCA, TPCA, and BPGA models. Average and standard deviation over 20 test images.

	LPCA	TPCA	BPGA
Average MSE	4.2×10^{-2}	3.4×10^{-2}	2.8×10^{-2}
Std of MSE	1.25×10^{-2}	4.8×10^{-3}	4.2×10^{-3}

principal modes. This reflects that our model BPGA gains a more compact representation than TPCA and LPCA.

5. Conclusion and Future Work

We presented a generative Bayesian model of principal geodesic analysis in diffeomorphic image registration. Our method is the first probabilistic model for automatic dimensionality reduction for diffeomorphisms. We developed an inference strategy based on MAP to estimate parameters, including the noise variance and image atlas, simultaneously. The estimated low-dimensional latent variables provide a compact representation of the anatomical variability in a large image database, and they can be used for further statistical analysis of anatomical shape in clinical studies. Reducing the dimensionality to the inherent modes of shape variability has the potential to improve hypothesis testing, classification, and mixture models.

There are several avenues for future work to build upon our BPGA model. In this paper, we precomputed the regularization parameter using simple atlas building model in Zhang et al. (2013). Since different parameters can lead to different principal modes, atlas, etc., ideally we would estimate the regularization parameter simultaneously with all other parameters. Doing this would require a more computationally-expensive approach that integrates out the latent x variables, rather than the mode approximation used here. Such an approach has been done for PPGA on finite-dimensional

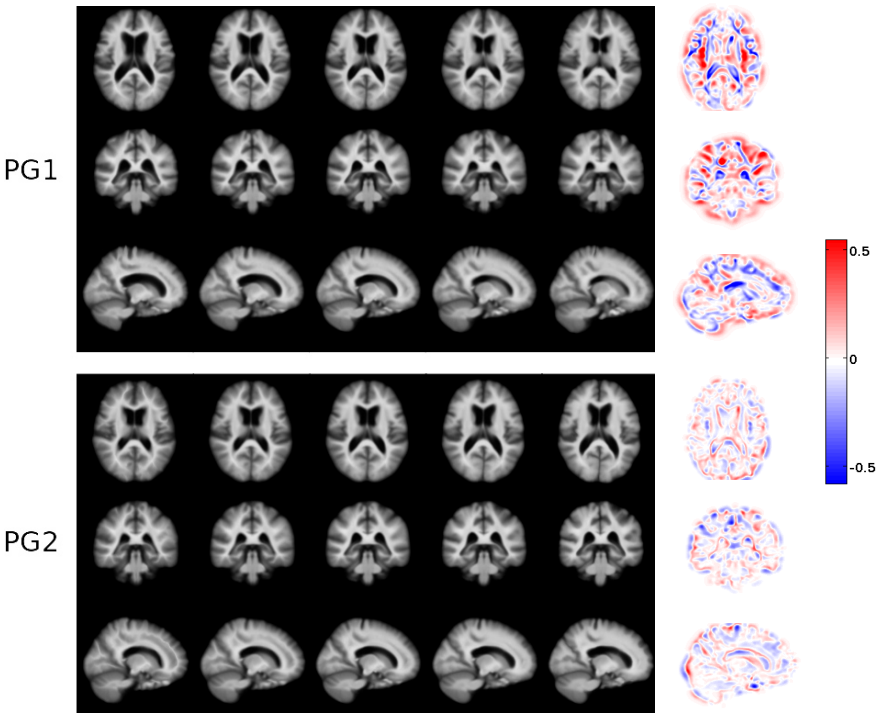
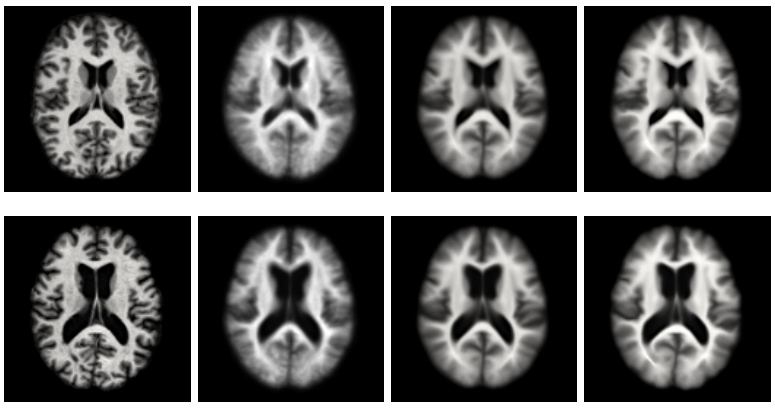


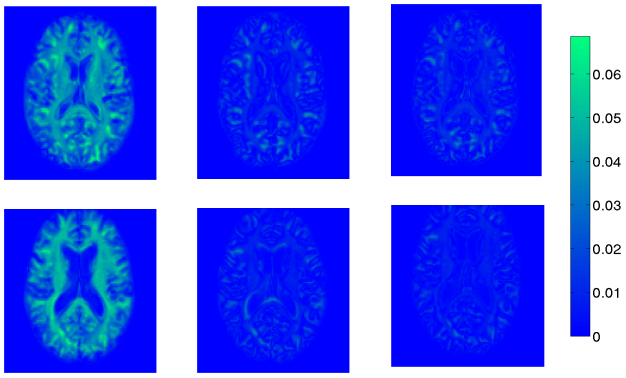
Figure 4: Top to bottom: axial, coronal and sagittal views of shooting the atlas by the first and second principal modes. Left to right: BPGA model of image variation evaluated at $a_i = -3, -1.5, 0, 1.5, 3$, and log determinant of Jacobians at $a_i = 3$.



(a) Observed (b) LPCA (c) TPCA (d) BPGA

Figure 5: Left to right: original data, reconstruction by LPCA, TPCA, and BPGA.

manifolds (Zhang and Fletcher, 2013). This would be related to several other approaches that integrate out deformations in image atlas building. For instance, Allassonnière and Kuhn (2010) proposed a fully generative Bayesian model of small elastic deformation in which the latent image transformations are marginalized from the distribution. Markov chain Monte Carlo (MCMC) methods for sampling elastic deformations in Bayesian atlas models have been introduced by Van Leemput (2009), Risholm et al. (2010), and Iglesias et al. (2012). Furthermore, Simpson et al. (2012) inferred the regularization parameter from a hierarchical Bayesian model, although their work was in the elastic deformation setting as well. Zhang et al. (2013) were the first to develop



(a) LPCA (b) TPCA (c) BPGA

Figure 6: Left to right: absolute value of reconstruction error map by LPCA, TPCA, and BPGA.

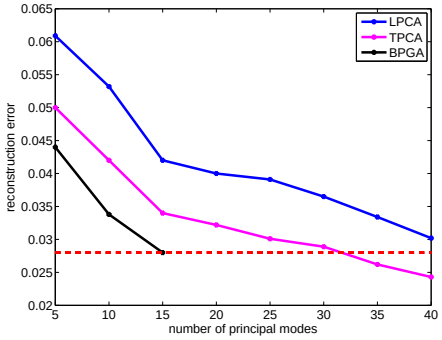


Figure 7: Averaged mean squared reconstruction error with different number of principal modes by LPCA, TPCA, and BPGA over 20 test images.

a truly Bayesian model for diffeomorphic atlas building and regularization parameter estimation by integrating out latent random diffeomorphisms.

In addition, much like the Euclidean BPCA model (Bishop, 1999), we did not enforce that the principal modes be orthogonal. This can be achieved by optimization in the Stiefel manifold of orthonormal frames, as is done in Zhang and Fletcher (2013). However, the high-dimensionality of velocity fields makes this a difficult problem to implement directly.

Acknowledgments. This work was supported by NIH Grant 5R01EB007688 and NSF CAREER Grant 1054057.

Appendix A

Deriving Expectation. The complete expectation function is

$$\begin{aligned} Q(W, x^k, \theta \,|\, \hat{\theta}, \hat{W}) &= E_{\tau \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}} \left[\sum_{k=1}^{N} \log p(W \,|\, J^k; \theta) \right] \\ &\propto -\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2 - \frac{MN}{2} \log \sigma \\ &\quad - \frac{1}{2} \sum_{k=1}^{N} \left\| W x^k \right\|_V^2 - \sum_{i=1}^{q} \frac{\|W_i\|_V^2}{2} E[\tau_i^{-1} \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}]. \end{aligned} \tag{12}$$

Since the sum of log likelihood and the prior on x^k do not depend on τ, we reduce the E-step to compute the conditional expectation $E[\tau_i^{-1} \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}]$. Observe that $p(\tau_i \,|\, J^k; \theta, W, x^k) = p(\tau_i \,|\, W, x^k)$, thus

$$p(\tau_i \,|\, J^k; \theta, W, x^k) = \frac{p(W \,|\, \tau_i, x^k) p(\tau_i)}{\int p(W \,|\, \tau_i, x^k) p(\tau_i) d\tau_i}.$$

The conditional expectation is computed by

$$\begin{aligned} E[\tau_i^{-1} \mid J^k; \hat{\theta}, \hat{W}, \hat{x^k}] &= \int \frac{1}{\tau_i} p(\tau_i \mid J^k; \hat{\theta}, \hat{W}, \hat{x^k}) d\tau_i, \\ &= \frac{\int \frac{1}{\tau_i} N(\hat{W} \mid 0, \tau_i) \frac{1}{\tau_i} d\tau_i}{\int N(\hat{W} \mid 0, \tau_i) \frac{1}{\tau_i} d\tau_i}, \\ &= \frac{1}{\|\hat{W}_i\|_V^2}, \end{aligned} \tag{13}$$

We obtain the Q function by plugging (13) into (12).

Deriving Derivatives. Now we compute the variation of $\tilde{Q}$ w.r.t. time-dependent variables I^k, m^k, v^k. Note that the following equations are equivalent for the geodesic paths of each of the subjects, so for notation simplicity, we drop the subject index k momentarily. The derivations are

$$\begin{aligned} \partial_I \tilde{Q} &= \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{I}, (\dot{I} + \epsilon \delta \dot{I}) + \nabla (I + \epsilon \delta I) \cdot v \rangle_{L^2} \\ &\quad + \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} \\ &= \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} + \int_0^1 \langle \hat{I}, \delta \dot{I} + \nabla \delta I \cdot v \rangle_{L^2} \\ &= \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} + \langle \hat{I}, \delta I \rangle_{L^2}\Big|_{t=0}^{t=1} - \int_0^1 \langle \dot{\hat{I}}, \delta I \rangle_{L^2} \\ &\quad + \int_0^1 \langle \hat{I}, \nabla \delta I \cdot v \rangle_{L^2} \\ &= \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} + \langle \hat{I}_1, \delta I_1 \rangle_{L^2} - \langle \hat{I}_0, \delta I_0 \rangle_{L^2} \\ &\quad - \int_0^1 \langle \dot{\hat{I}}, \delta I \rangle_{L^2} - \int_0^1 \langle \nabla \cdot (\hat{I} v), \delta I \rangle_{L^2}, \end{aligned}$$

$$\begin{aligned} \partial_v \tilde{Q} &= \langle L v_0, \delta v_0 \rangle_{L^2} + \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{v}, \delta \dot{v} + K \mathrm{ad}^*_{v+\epsilon\delta v} m \rangle_{L^2} \\ &\quad + \langle \hat{I}, \dot{I} + \nabla I \cdot (v + \epsilon \delta v) \rangle_{L^2} \\ &= \langle L v_0, \delta v_0 \rangle_{L^2} + \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{v}, \delta \dot{v} \rangle_{L^2} + \langle \mathrm{ad}_{v+\epsilon\delta v} K \hat{v}, m \rangle_{L^2} \\ &\quad + \langle \hat{I}, \dot{I} + \nabla I \cdot (v + \epsilon \delta v) \rangle_{L^2} \\ &= \langle L v_0, \delta v_0 \rangle_{L^2} + \langle \hat{v}, \delta v \rangle_{L^2}\Big|_{t=0}^{t=1} - \int_0^1 \langle \dot{\hat{v}}, \delta v \rangle_{L^2} \\ &\quad + \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle -\mathrm{ad}_{K\hat{v}} (v + \epsilon \delta v), m \rangle_{L^2} \\ &\quad + \langle \hat{I}, \dot{I} + \nabla I \cdot (v + \epsilon \delta v) \rangle_{L^2} \\ &= \langle L v_0, \delta v_0 \rangle_{L^2} + \langle \hat{v}_1, \delta v_1 \rangle_{L^2} - \langle \hat{v}_0, \delta v_0 \rangle_{L^2} \\ &\quad + \int_0^1 \langle -\mathrm{ad}_{K\hat{v}} \delta v, m \rangle_{L^2} + \langle \hat{I}, \nabla I \cdot \delta v \rangle_{L^2} - \langle \dot{\hat{v}}, \delta v \rangle_{L^2} \\ &= \langle L v_0, \delta v_0 \rangle_{L^2} + \langle \hat{v}_1, \delta v_1 \rangle_{L^2} - \langle \hat{v}_0, \delta v_0 \rangle_{L^2} \\ &\quad + \int_0^1 \langle -\mathrm{ad}^*_{K\hat{v}} m, \delta v \rangle_{L^2} + \langle \hat{I} \nabla I, \delta v \rangle_{L^2} - \langle \dot{\hat{v}}, \delta v \rangle_{L^2}, \end{aligned}$$

$$\begin{aligned} \partial_m \tilde{Q} &= \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{v}, K \mathrm{ad}^*_v (m + \epsilon \delta m) \rangle_{L^2} + \langle \hat{m}, (m + \epsilon \delta m) \rangle_{L^2} \\ &= \int_0^1 \langle \mathrm{ad}_v K \hat{v}, \delta m \rangle_{L^2} + \langle \hat{m}, \delta m \rangle_{L^2}, \end{aligned}$$

here $\nabla \cdot$ is the divergence operator. Since we have $\delta I_0 = 0$, $\delta v_0 = 0$, the optimality conditions for I, v are given by the following time-dependent system of ODEs, termed the *adjoint equations*:

$$\left.\begin{aligned} -\dot{\hat{I}} - \nabla \cdot (\hat{I} v) &= 0, \\ -\mathrm{ad}^*_{K\hat{v}} m + \hat{I} \nabla I - \hat{\dot{v}} &= 0, \\ \mathrm{ad}_v K \hat{v} + \hat{m} &= 0, \end{aligned}\right\} \tag{14}$$

subject to initial conditions

$$\hat{v}_1 = 0, \qquad \hat{I}_1 = \frac{1}{\sigma^2}(I_1 - J).$$

Finally, after integrating these adjoint equations backwards in time to $t = 0$, the gradient of $\tilde{Q}$ with respect to the kth initial velocity is

$$\nabla_{v_0^k} \tilde{Q} = v_0^k - K \hat{v}_0^k.$$

Deriving Closed-form Solution for θ. Notice that the gradient term of (10) w.r.t. $\theta = \{I, \sigma\}$ only relates to the image matching term

$$\tilde{Q}(I, \sigma) = -\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2 - \frac{MN}{2} \log \sigma, \tag{15}$$

we have the gradient of (15) as

$$\partial_\sigma \tilde{Q}(I, \sigma) = \frac{1}{\sigma^3} \sum_{k=1}^{N} \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2 - \frac{MN}{2\sigma}. \tag{16}$$

Setting (16) to zero, We then get the closed-form formulation to update σ^2 by

$$\sigma^2 = \frac{1}{MN} \sum_{k=1}^{N} \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2.$$

Next we compute the gradient of (10) w.r.t. the atlas I by changing variables $y = \left(\phi^k\right)^{-1}(x)$, such that

$$x = \phi^k(y), \quad dx = |D\phi^k(y)| dy.$$

After dropping the normalizing constant of σ which is irrelevant to I, we expand and rewrite equation (15) as

$$\tilde{Q}(I) = \sum_{k=1}^{N} \int_\Omega \langle I(y) - J^k \circ \phi^k(y), I(y) - J^k \circ \phi^k(y) \rangle_{L^2} |D\phi^k(y)| dy.$$

This gives the derivative w.r.t I by

$$\partial_I \tilde{Q} = \sum_{k=1}^{N} (I - J^k \circ \phi^k)|D\phi^k| \tag{17}$$

Equating (17) to zero at optimal, we have

$$I = \frac{\sum_{k=1}^{N} J^k \circ \phi^k |D\phi^k|}{\sum_{k=1}^{N} |D\phi^k|}.$$

References

Allassonnière, S., Kuhn, E., 2010. Stochastic algorithm for parameter estimation for dense deformable template mixture model. ESAIM-PS 14, 382–408.

Andrews, D.F., Mallows, C.L., 1974. Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological) , 99–102.

Arnol'd, V.I., 1966. Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. Ann. Inst. Fourier 16, 319–361.

Beg, M., Miller, M., Trouvé, A., Younes, L., 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International Journal of Computer Vision 61, 139–157.

Bishop, C.M., 1999. Bayesian PCA. Advances in neural information processing systems , 382–388.

Figueiredo, M.A., 2003. Adaptive sparseness for supervised learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on 25, 1150–1159.

Fletcher, P.T., Lu, C., Joshi, S., 2003. Statistics of shape via principal geodesic analysis on Lie groups, in: Computer Vision and Pattern Recognition, IEEE. pp. I–95.

Gerber, S., Tasdizen, T., Fletcher, P.T., Joshi, S., Whitaker, R., 2010. Manifold modeling for brain population analysis. Medical image analysis 14, 643–653.

Gori, P., Colliot, O., Worbe, Y., Marrakchi-Kacem, L., Lecomte, S., Poupon, C., Hartmann, A., Ayache, N., Durrleman, S., 2013. Bayesian atlas estimation for the variability analysis of shape complexes, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 267–274.

Iglesias, J.E., Sabuncu, M.R., Leemput, K.V., ADNI, 2012. Incorporating parameter uncertainty in Bayesian segmentation models: application to hippocampal subfield volumetry, in: Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 50–57.

Joshi, S., Davis, B., Jomier, M., Gerig, G., 2004. Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23, S151–S160.

Ma, J., Miller, M.I., Trouvé, A., Younes, L., 2008. Bayesian template estimation in computational anatomy. NeuroImage 42, 252–261.

Miller, M.I., Trouvé, A., Younes, L., 2006. Geodesic shooting for computational anatomy. Journal of Mathematical Imaging and Vision 24, 209–228.

Qiu, A., Younes, L., Miller, M.I., 2012. Principal component based diffeomorphic surface mapping. Medical Imaging, IEEE Transactions on 31, 302–311.

Risholm, P., Samset, E., Wells, W.M., 2010. Bayesian estimation of deformation and elastic parameters in non-rigid registration, in: Biomedical image registration. Springer, pp. 104–115.

Said, S., Courty, N., Le Bihan, N., Sangwine, S.J., 2007. Exact principal geodesic analysis for data on SO(3), in: Proceedings of the 15th European Signal Processing Conference, pp. 1700–1705.

Simpson, I.J.A., Schnabel, J.A., Groves, A.R., Andersson, J.L.R., Woolrich, M.W., 2012. Probabilistic inference of regularisation in non-rigid registration. NeuroImage 59, 2438–2451.

Singh, N., Hinkle, J., Joshi, S., Fletcher, P.T., 2013. A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction, in: International Symposium on Biomedical Imaging, IEEE. pp. 1219–1222.

Sommer, S., Lauze, F., Hauberg, S., Nielsen, M., 2010. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 43–56.

Twining, C., Cootes, T., Marsland, S., Petrovic, V., Schestowitz, R., Taylor, C., 2005. A unified information-theoretic approach to groupwise non-rigid registration and model building, in: Information Processing in Medical Imaging, Springer. pp. 1–14.

Vaillant, M., Miller, M.I., Younes, L., Trouvé, A., 2004. Statistics on diffeomorphisms via tangent space representations. NeuroImage 23, S161–S169.

Van Leemput, K., 2009. Encoding probabilistic brain atlases using Bayesian inference. Medical Imaging, IEEE Transactions on 28, 822–837.

Vialard, F.X., Risser, L., Holm, D., Rueckert, D., 2011. Diffeomorphic atlas estimation using Kärcher mean and geodesic shooting on volumetric images, in: MIUA.

Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J., 2012. Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. International Journal of Computer Vision 97, 229–241.

Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S., 2007. Graph embedding and extensions: a general framework for dimensionality reduction. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29, 40–51.

Zhang, M., Fletcher, P.T., 2013. Probabilistic principal geodesic analysis, in: Advances in Neural Information Processing Systems, pp. 1178–1186.

Zhang, M., Fletcher, P.T., 2014. Bayesian principal geodesic analysis in diffeomorphic image registration, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 121–128.

Zhang, M., Singh, N., Fletcher, P.T., 2013. Bayesian estimation of regularization and atlas building in diffeomorphic image registration, in: Information Processing in Medical Imaging, Springer. pp. 37–48.