Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(515): 1050–1060. doi:10.1080/01621459.2016.1164050.

Parametrization of white matter manifold-like structures using principal surfaces

Chen Yue^a, Vadim Zipunnikov^a, Pierre-Louis Bazin^d, Dzung Pham^c, Daniel Reich^b, Ciprian Crainiceanu^a, and Brian Caffo^a

^aDepartment of Biostatistics, Johns Hopkins University, Baltimore, MD, 21205

^bDepartment of Radiology and Imaging Sciences, National Institute of Health, Bethesda, MD 20892

^cCenter for Neuroscience and Regenerative Medicine, Henry Jackson Foundation, Bethesda, MD 20892

^dDepartment of Neurophysics, Max Planck Institute, Leipzig, Germany, 04103

Abstract

In this manuscript, we are concerned with data generated from a diffusion tensor imaging (DTI) experiment. The goal is to parameterize manifold-like white matter tracts, such as the corpus callosum, using principal surfaces. The problem is approached by finding a geometrically motivated surface-based representation of the corpus callosum and visualized fractional anisotropy (FA) values projected onto the surface. The method also applies to any other diffusion summary. An algorithm is proposed that 1) constructs the principal surface of a corpus callosum; 2) flattens the surface into a parametric 2D map; 3) projects associated FA values on the map. The algorithm is applied to a longitudinal study containing 466 diffusion tensor images of 176 multiple sclerosis (MS) patients observed at multiple visits. For each subject and visit the study contains a registered DTI scan of the corpus callosum at roughly 20,000 voxels. Extensive simulation studies demonstrate fast convergence and robust performance of the algorithm under a variety of challenging scenarios.

Keywords

Corpus Callosum; Principal Curves and Surfaces; Thin Plate Splines

1. INTRODUCTION

This research is motivated by the need to establish a parametric description of the corpus callosum structure, a major white matter fiber tract. The corpus callosum is a centrally located white matter structure connecting the left and right hemispheres of the brain. It is the largest white matter fiber bundle in the brain, facilitating inter-hemispheric communication. Corpus callosum neurons run left to right spanning the mid-sagittal plane.

The three panels in Figure 1 display that the corpus callosum appears as a two dimensional manifold in its principal structure. It is curved towards the inferior part of the brain on both the anterior and posterior sides. Though the corpus callosum lies in a three dimensional space, its key structure is intuitively that of a “carpet” that lies in a two dimensional manifold. Therefore, dimension reduction techniques may provide strong data compression along with novel visualization and parametrization approaches that could be easy to use in practice.

In this manuscript, a method for obtaining smooth principal surfaces of the data with a consistent parametrization is proposed. The word “consistent” implies that one would achieve a similar parametrization with the same range for assembled data structures. For example, one could obtain a comparable parametrization for corpus callosum of different subjects. To achieve this, the first step is to estimate the center surface of the corpus callosum. The second is to obtain the projection of each data point; these projections can then be used to map the diffusion properties of the corpus callosum onto a 2D manifold. The procedure, similar to Tract-Based Spatial Statistics (TBSS) (Smith et al. 2006), or other medial model based methods (Yushkevich et al. 2008), (Zhang et al. 2010), constructs a consistent mapping of DTI images of the corpus callosum which has the potential to perform pixel-wise analysis, but unlike these methods, may not require non-linear registration. At the same time, our approach also has the potential to perform pixel-wise analysis of FA value, local thickness and curvature.

Before we discuss the details of our method, other potential methods are outlined. Principal component analysis (PCA) is one of the most useful tools for dimension reduction. PCA finds directions (vectors) that explain the largest variability of the data, while constraining directions to be orthogonal.

Though PCA is a widely used method for dimension reduction, it is not suitable for nonlinear situation – as shown in the left panel in Figure 2, the true curve that generates the data is part of a circle while the first principal component is a line. Direct applications of splines, wavelets and related regression methods cannot be done, as they require the mean function to be one-to-one. This is clearly violated in the corpus callosum example and many other white matter structures/tracts. Palus and Dvorak (1992) and Palus and Dvorak (1992) illustrated the pitfalls and precautions when applying linear PCA in non-linear settings. In our case, since the corpus callosum clearly is a nonlinear structure, PCA is not a viable candidate for dimension reduction.

Many non-linear methods have been proposed for fitting non-linear data structures. As an example, Gnanadesikan (2011) proposed a non-linear extension of PCA. The core idea is to include product combinations of the variables in the data matrix. Another useful tool – the self-organizing map (SOM) – was proposed in Kohonen (1990) and Kohonen (1982). SOMs are unsupervised learning procedures which are used to discover structure in the data. Other nonlinear method methods such as non-linear principal component analysis (NLPCA) (Kramer 1992; Kramer 1991) and principal geodesic analysis (PGA) (Fletcher et al. 2004; Fletcher 2004) can also fit non-linear data structures.

An important concept of non-linear data compression is principal surfaces (Hastie 1984; Hastie and Stuetzle 1989). Principal surfaces are manifolds that pass through the middle of the data. Principal surfaces, by definition, satisfy a *self consistency* condition, in that they are the conditional expectation (local average) of the data. Hastie (1984) and Hastie and Stuetzle (1989) shows that the principal surfaces are fit via nonparametric low-dimensional manifolds that minimize the orthogonal distance from the data to themselves. The right panel of Figure 2 shows the difference between the principal curves (surfaces) and regression. It highlights that the principal curve minimizes the sum of orthogonal distances, while a spline model fit tries to minimize the sum of distances parallel to the y axis. The principal surface algorithm and its extensions (Dong and McAvoy 1996; Einbeck et al. 2010; Gerber et al. 2009; Goldsmith et al. 2011a; Jung et al. 2011; Leblanc and Tibshirani 1994; Ozertem and Erdogmus 2011), do not provide a consistent parametrization in the 2D space (see Section 3). ISOMAP (Tenenbaum et al. 2000) and Maximum Variance Unfolding (MVU) (Weinberger and Saul 2006) are methods for dimension reduction. Compared to the Hastie's algorithm (denoted as "HS"), ISOMAP and MVU provide more consistent parametrization (see Section 3).

In this manuscript, we develop a method that 1) achieves a principal surface of a target data cloud in 3D space, 2) yields a consistent parametrization in the 2D space for similar 3D data clouds. The rest of the manuscript is laid out as follows: In section 2, the principal surface concept will be introduced, and its corresponding algorithm will be shown. We will show how the principal surface algorithm works on simulated data in section 3. The algorithm then is applied to corpus callosum data and the FA maps are obtained in section 4. We will conclude the whole paper in section 5.

2. METHODS

2.1 Principal Surfaces

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})^T$, $i = 1, \dots, I$ be the data points in three dimensional space, $\mathcal{R}^3$ following an underlying distribution. $\mathbf{t}_i = (t_{i1}, t_{i2})^T$ be corresponding parametrization points in two dimensional space, $\mathcal{R}^2$. In addition, It is required (without loss of generality) that the 2D coordinate space be the unit square $[0, 1] \times [0, 1] \subset \mathcal{R}^2$. Let f be the smooth principal surface function, $f: \mathbf{t}_i \mapsto f(\mathbf{t}_i)$, that maps from $\mathcal{R}^2$ to $\mathcal{R}^3$. The principal surface function satisfies the *self-consistency* condition:

$$E(\mathbf{X} | \lambda_f(\mathbf{X}) = \mathbf{t}) = f(\mathbf{t}) \quad \text{for all } \mathbf{t}, \quad (1)$$

where $\lambda_f(\mathbf{x}) = \sup_{\mathbf{t}} \{ \mathbf{t} : \|\mathbf{x} - f(\mathbf{t})\| = \inf_{\mu} \|\mathbf{x} - f(\mu)\| \}$ is the projection function with respect to f . The projection function maps a data point on to the closest principal surface point having the largest parametrization. Intuitively, the self-consistency condition implies that the principal surface is the local average of the original data cloud. Here local means that data points have the same 2D parametrization. In developing an algorithm of achieving the principal surface of a data cloud, we have found that there are two main distinctions between different fitted principal surfaces: the degree of smoothness and the method of parametrization. In most algorithms, these properties will be controlled by the specific

smoother being used in the algorithm and its tuning parameters. The details of the specific algorithm will be demonstrated in the next section.

2.2 Algorithm

In this paper, we put forward an algorithm that is partially based on Hastie and Stuetzle (1989) but heavily modified for applications where underlying structures are smooth manifolds. This algorithm allows one to find the surface coordinate for each data point $(\mathbf{t}_i, i = 1, \dots, J)$, which will be used later to create parametric summaries. However, the original principal surface algorithm can only yield surfaces which are locally flattened. Therefore, instead of local planar smoothers, thin-plate splines (TPS) are employed for fitting the surface. (Thin-plate splines were proposed by Duchon (1977) and are now widely used for bivariate smoothing.) The TPS penalize the least squares error by a high-order derivative term in order to achieve a desired degree of smoothness. Wood (2003) and Wood (2003) improved the computational efficiency when fitting TPS by using an optimal approximating basis that we employ.

Initializing—Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_J]^T$ be the $I \times 3$ matrix that contains the centered coordinates. There are a few ways to initialize our algorithm:

PCA: Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the singular value decomposition of $\mathbf{X}$. Then $\mathbf{U}\tilde{\mathbf{\Sigma}}$ are the first two principal scores of the data matrix, where $\tilde{\mathbf{\Sigma}}$ is a submatrix of $\mathbf{\Sigma}$ containing the first two columns. Both scores are standardized to be in $[0, 1]$. and then used as an initial 2D parametrization.

ISOMAP: Construct a 3D to 2D ISOMAP on $\mathbf{X}$ and use its results as the initial 2D parametrization. ISOMAP usually provides a more natural initializing parametrization, but requires more computational time.

Smooth Local Averages—Bivariate thin plate splines (TPS) (Wood 2003) are applied to the original data points using the projection coordinates obtained from the previous step. In this step, the model

$$x_{id} = f_{d1}(t_{i1}) + f_{d2}(t_{i2}) + f_{d3}(t_{i1}, t_{i2}) + \varepsilon_{id}, \quad d=1, 2, 3, \quad (2)$$

is fit and bivariate TPS smoothing is obtained. Here

$$f(\mathbf{t}_i) = \begin{bmatrix} \hat{f}_{11}(t_{i1}) + \hat{f}_{12}(t_{i2}) + \hat{f}_{13}(t_{i1}, t_{i2}) \\ \hat{f}_{21}(t_{i1}) + \hat{f}_{22}(t_{i2}) + \hat{f}_{23}(t_{i1}, t_{i2}) \\ \hat{f}_{31}(t_{i1}) + \hat{f}_{32}(t_{i2}) + \hat{f}_{33}(t_{i1}, t_{i2}) \end{bmatrix} \quad (3)$$

is the current principal surface mapping from the 2D parametrization space to 3D coordinate space. The TPS are fit using *mgcv* package in R (R Development Core Team 2008).

Projection—Each data point was then projected onto the current principal surface and a new 2D parametrization was thus obtained. A grid search method was used to find the projection and constrained within the domain $[0, 1] \times [0, 1]$. Therefore, there will typically be some data points being projected onto the boundary, which raises some issues when further analyzing the 2D parametrization, see Section 5 for more discussion.

After the projection step, the procedure is iterated. The algorithm is illustrated in pseudocode below:

Algorithm 1

Principal surface algorithm

Input: Data in 3D coordinate, $\mathbf{X}_{I \times 3}$

Output: Principal surface function, $f_{\mathcal{R}^2 \rightarrow \mathcal{R}^3}$, and the 2D parametrization of all data points, $\mathbf{T}_{I \times 2}$

Initialization: De-mean $\mathbf{X}$ for each column;

set the initial 2D parametrization as $\mathbf{T}^{(0)}$;

Set $err = 1$, $i = 1$;

while ($i < max.iter$ **and** $err > thres$) **do**

(1). Fit $\mathbf{X} = f(\mathbf{T}) + \epsilon$;

(2). $\mathbf{t}_i^{(new)} \leftarrow \arg \min_{\mathbf{t}} \|\mathbf{x}_i - f(\mathbf{t})\|^2$;

(3). $err \leftarrow \|\mathbf{T}^{old} - \mathbf{T}^{(new)}\|_2^2$, $i \leftarrow i + 1$;

end

To summarize, the algorithm iterates between two main stages: a smoothing step (Steps 1) and a projection step (Step 2). The TPS fitting step provides local averages with desired smoothness as well as a principal surface to be projected onto. Step 3 gives us a criterion for convergence.

The only parameter that needs to be chosen is the number of basis functions during TPS fitting step. This controls the complexity of the surface.

Generally, a principal surface is not unique. By varying the number of basis functions in the TPS fitting, one can achieve different principal surfaces with different curvature (see Section 3.3 for more discussion).

3. SIMULATIONS

3.1 Simulation Settings

To investigate the performance of our algorithm, four simulations studies were conducted. The number of data points was set to $I = 1500$ in all simulation settings and 500, 1000, 1500 points in each scenarios were sub-sampled to assess the accuracy of the algorithm. We apply

the proposed principal surface fitting algorithm with both PCA and ISOMAP initializations. In addition, since ISOMAP and MVU naturally produce a 2D parametrization, we apply both of them¹ to each simulation case for comparison. HS's original principal surface algorithm was also applied. The data were centered around (0, 0, 0) beforehand.

Scenario 1—The data points in the first simulation scenario are uniformly distributed around a cylinder with an open seam. Set $\theta_i \stackrel{i.i.d.}{\sim} U(0, 2\pi-0.5)$, $\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3} \stackrel{i.i.d.}{\sim} N(0, 0.15^2)$ and $z_i \stackrel{i.i.d.}{\sim} U(-3, 3)$, then let

$$\mathbf{x}_i = \begin{bmatrix} \cos\theta_i + \varepsilon_{i1} \\ \sin\theta_i + \varepsilon_{i2} \\ z_i + \varepsilon_{i3} \end{bmatrix}.$$

Scenario 2—The second scenario is a simulated corpus callosum. Set

$\theta_i \stackrel{i.i.d.}{\sim} U(\pi/2, 3\pi/2)$, $z_{i1} = \cos(\theta_i)$, $z_{i3} = -1 + \sin(\theta_i)$ for $i = 1, \dots, I/3$. $z_{i1} \stackrel{i.i.d.}{\sim} U(0, 3)$ and $z_{i3} = 0$ for $i = I/3 + 1, \dots, 2I/3$. Set $\theta_i \stackrel{i.i.d.}{\sim} U(-\pi/2, \pi/2)$, $z_{i1} = 3 + \cos(\theta_i)$, $z_{i3} = -1 + \sin(\theta_i)$ for $i = 2I/3 + 1, \dots, I$. $z_{i2} \stackrel{i.i.d.}{\sim} U(0, 5)$ for all i . Then let $\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3} \stackrel{i.i.d.}{\sim} N(0, 0.16^2)$ for all i .

$$\mathbf{x}_i = \begin{bmatrix} z_{i1} + \varepsilon_{i1} \\ z_{i2} + \varepsilon_{i2} \\ z_{i3} + \varepsilon_{i3} \end{bmatrix}.$$

Scenario 3—In the third scenario the data points form a flatten surface at the beginning and then begin to bent over towards the bottom, which looks like a half of a corpus

callosum. Set $z_{i1} \stackrel{i.i.d.}{\sim} U(0, 2)$, $z_{i3} = 0$ when $i = 1, \dots, I/2$; $z_{i1} = \cos(\theta_i) + 2$, $z_{i3} = -1 + \sin(\theta_i)$ for $i = I/2 + 1, \dots, I$, where $\theta_i \stackrel{i.i.d.}{\sim} U(-\pi/2, \pi/2)$. Let $z_{i2} \stackrel{i.i.d.}{\sim} U(0, 10)$ and $\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3} \stackrel{i.i.d.}{\sim} N(0, 0.4^2)$. All the random numbers are generated independently and

$$\mathbf{x}_i = \begin{bmatrix} z_{i1} + \varepsilon_{i1} \\ z_{i2} + \varepsilon_{i2} \\ z_{i3} + \varepsilon_{i3} \end{bmatrix}.$$

Scenario 4—In the last scenario, a two dimensional stretched digit “5” was used as a basis for simulation. Let $z_{i1} \stackrel{i.i.d.}{\sim} U(0, 1)$, $z_{i3} = 0$ when $i = 1, \dots, 3I/10$; $z_{i1} = 0$, $z_{i3} \stackrel{i.i.d.}{\sim} U(-1, 0)$ when $i = 3I/10 + 1, \dots, 4.5I/10$; $z_{i1} \stackrel{i.i.d.}{\sim} U(0, 0.5)$, $z_{i3} = -1$ when $i = 4.5I/10 + 1, \dots, 6I/10$; $z_{i1} = \frac{1}{2} + \frac{1}{2}\cos(\theta_i)$, $z_{i3} = -\frac{3}{2} + \frac{1}{2}\sin(\theta_i)$ for $i = 6I/10 + 1, \dots, 8.5I/10$ where

¹R package “vegan” was used to run ISOMAP. Landmark MVU with one time TPS fitting was fit in Matlab Toolbox for Dimensionality Reduction.

$\theta_i \stackrel{i.i.d.}{\sim} U(-\pi/2, \pi/2); z_{i1} \stackrel{i.i.d.}{\sim} U(0, 0.5), z_B = -2$ when $i = 8.5I/10 + 1, \dots, I$. Let $z_{i2} \stackrel{i.i.d.}{\sim} U(0, 5)$ and $\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3} \stackrel{i.i.d.}{\sim} N(0, 0.1^2)$ when $i = 1, \dots, I$. All the random numbers are generated independently and

$$\mathbf{x}_i = \begin{bmatrix} z_{i1} + \varepsilon_{i1} \\ z_{i2} + \varepsilon_{i2} \\ z_{i3} + \varepsilon_{i3} \end{bmatrix}.$$

3.2 Simulation Results

For all four scenarios, the proposed algorithm converged in less than 10 steps and took under two minutes on a i7-2.4GHz PC machine with 8Gb RAM memory. The results of our principal surface fitting algorithm for four scenarios are shown in Figure 3. In the upper right panel in Figure 3, one can see that the algorithm reconstructs the cylinder very well. Since no constraints were used in fitting the surface, a closed cylinder is not obtained. Finding principal surfaces for closed cylinders or spheres remains an interesting topic for future research. The other panels of Figure 3 show the fitting results of the non-function shaped carpet (half corpus callosum) data cloud, the corpus callosum-shaped structure and the simulated digit “5” data cloud. The results are excellent for all scenarios. Given that the desired corpus callosum model fit is simpler and smoother than all considered examples, the simulations produce substantial evidence of the viability of the principal surface algorithm as a robust method for a wide range of problems.

Figure 4 compares the fitting result using (1) proposed method with PCA initialization, (2) proposed method with ISOMAP initialization, (3) ISOMAP induced surface² and (4) HS’s original method. The proposed principal surface algorithm generates reasonable fitting results with both initialization methods. Landmark MVU algorithm did not converge in all simulation cases, so the results are not reported.

Figure 5 shows the results for different number basis functions in TPS fitting. Starting from upper left panel, the surface fitting with $k = 20$ yields a fitted surface with less curvature. As k increases, the fitted surface becomes increasingly wiggly. Table 1 shows the curvature³ and the corresponding thickness⁴ of the fitted surface. It is very clear that as the number of basis functions increases, the curvature of fitted surface also increases and the corresponding thickness decreases. In Figure 5, it is clear that a desired surface for this particular case is with k less than or equal to 20.

To account for a 3D (spatial) error structure, we propose a new cross-validation method for finding the optimal number of basis functions. For each k , MSE(k) is calculated in four steps: i) a reference principal surface is fitted using the entire sample; ii) the sample is splitted into $n=3$ folds and n principal surfaces are fitted leaving a corresponding fold out;

²ISOMAP induced surface is created using ISOMAP 3D to 2D parametrization with TPS smoothing with each coordinates. It can be considered a surface fitting of ISOMAP result or the proposed method with only one step without any projection and further iterations.

³Integral of the second order derivatives

⁴A moving maximum of the distance from 3D data points to the fitted surface.

iii) the points from each fold are projected onto the principal surface obtained by leaving the fold out and the fold-specific MSE is calculated as the squared distance between these projections and the projections obtained at step one; iv) $MSE(k)$ is calculated as the average of the fold-specific MSEs. The optimal k is the smallest number that minimizes $MSE(k)$.

The key difference of the proposed cross-validation approach is step three when MSE is calculated as a difference between fold-specific principal surface projections and projections onto the all data principal surface. Note that a traditional MSE would be proportional to a thickness estimates and, in contrast to the MSE at step three, would not be informative about the stability of the fit. However, one should be careful when the data are very dense so that sub-sampling results in surfaces similar to the original surface. In such case, our suggestion is to implement the algorithm with different number of basis functions and choose the most suitable one by visually exploring the fits.

In six scenarios in Figure 5, the leave-one-fold surface MSEs are 0.076, 0.082, 0.083, 0.101, 0.132 and 0.126, respectively. This is consistent with our observation that no more than twenty basis functions are usually enough to provide a good fit to data. For four cases in our simulation, the optimal number of basis functions are four, four, four and five for corpus callosum shaped structure, curved carpet structure, cylinder structure and digit “five” shaped structure respectively.

3.3 Mean Square Error Comparison

In evaluating the goodness of fit, we used mean squared error (MSE) that measures the distance from the fitted surface to the underlying true surface. The MSE was calculated by: 1) projecting each data point onto the fitted surface, 2) averaging the squared distance from the fitted surface to the true surfaces. Two hundred simulations were conducted for different number of sub-sampling points and the average MSE in each scenario is shown in Table 2. The MSE generally decreases with the increase of the sub-sample size. Of course, in each scenario, full dataset yields the lowest MSE. In the first three cases, ISOMAP initialization yields lower MSE compared to PCA initialization. However, PCA has faster computing time and lower memory demands.

The MSE comparison between the proposed method and ISOMAP and HS is reported in Table 3. For each scenario, the proposed method with ISOMAP initialization has the lowest MSE. HS has a significantly higher MSE in each case compared to the proposed method. ISOMAP has the highest MSE among all cases. LMVU algorithm did not converge in four cases, so no results for LMVU are reported in Table 3. The reason for that is that though both ISOMAP and MVU find low dimensional representations of high dimensional structures, neither of them is suitable for finding a principal surface. Intuitively, both algorithms try to preserve the local geometry such as distances or angles. However, 2D projections on the principal surface clearly do not preserve these properties. Therefore, it would be more appropriate to use ISOMAP or MVU on zero thickness surface structures, not on data cloud representing 3D structures such as corpus callosum. This is the reason why MVU, while preserving both local distance and angles between neighborhood data points, does not converges in most simulation cases.

Another important advantage of our method is that it obtains a more smooth parametrization than the original HS's method. In Figure 6, a color scale (from dark red to yellow) is used to represent the original data points. The color changes as the points move along the true surface from one side to another. After applying HS's, ISOMAP and the proposed methods, the 2D parametrizations are shown along with their original labeled colors. The proposed method as well as ISOMAP, in each scenario, shows more smooth, continuous parametrization. In contrast, the HS's method does not provide a natural parametrization in most scenarios, and produces a clear discontinuity in the cylinder scenario. ISOMAP produces a natural parametrization, but has a disadvantage of not providing a well fitted surface. Clearly, a comprehensive cross-algorithm study is needed to systematically explore resulting parametrizations and their properties.

4. APPLICATION

4.1 Fitting the Principal Surface of a Corpus Callosum

The MS study contained 466 scans generated from a diffusion tensor imaging (DTI) experiment performed on 176 patients. For each scan, fractional anisotropy (FA) value for each voxel of the entire corpus callosum area were calculated via tractography. The details are provided in Ozturk et al. (2010) and Reich et al. (2010). Fractional anisotropy has been associated both cross sectionally and longitudinally with multiple sclerosis diagnoses and symptoms (Goldsmith et al. 2011b; Goldsmith et al. 2012; Greven et al. 2011; Reich et al. 2010; Zipunnikov et al. 2011).

We start with the principal surface fit. For each unregistered scan, there are roughly 70,000 data points in the corpus callosum area, while for each registered scan, there are roughly 20,000 data points in the corpus callosum region of interest. For computational simplicity, 1,000 were randomly sub-sampled to build the surface. The computing time for unregistered corpus callosum is approximately 100 seconds, while the computing time for the registered ones is approximately 35 seconds. The results for two arbitrary unregistered scans are shown in Figure 7. The results for two arbitrary registered scans are shown in Figure 8. Clearly, unregistered data has a larger portion of the corpus callosum compared to the registered ones. The fitted surfaces for both registered and unregistered scans have obvious face validity and is indicative of the fits from the other scans, each inspected visually. Cross validation suggests five and seven as the optimal numbers of basis functions for registered and unregistered corpus callosum, respectively.

4.2 Flattened FA Representation

Our primary goal for the principal surface fitting is to flatten the surface to make a 2D FA image for visualization and subsequent analysis under the belief that the principal surface encodes the majority of the relevant biological information. Thus, our goal is to use DTI-based morphometric information to create 2D images of MR contrast properties, such as FA, axial diffusivity and so on. Thereby, each data point was projected onto the surface by a grid search method. Then the associated FA values were smoothed on the surface by local averaging. Interpolating this smooth images onto a grid yields the 100×100 2D FA images, Figure 9 shows a few typical examples. Top panels show the diffusion properties of the

flattened surfaces of the corresponding left panels. The similar 2D maps obtained from ISOMAP and HS are shown at the middle and bottom panels of Figure 9. For registered corpus callosum data, all three methods yield similar parametrizations. Arguably, the important FA information is retained. Thus this investigation provides the potential insight that it may be sufficient to visualize the 2D FA values instead of the original 3D FA values.

4.3 Correspondence

For corpus callosum data, it is crucial that the fitted 2D parametrization is consistent across all subjects. In other words, it is desirable that assembled structures have aligned 2D indices after the data points got projected onto the fitted principal surfaces. We admit that stable parametrization is not guaranteed in fitting the principal surface, especially, when the registration process involves complex non-linear transformation or linear transformation in a relatively large scale. However, we provided a small simulation study to explore how sensitive our algorithm is to mild registration errors. Specifically, for an arbitrary chosen image, we created 50 “unregistered” copies by applying random rotations and random scaling along each of the three axes. We, then, ran our algorithm with ISOMAP initialization on all fifty images and compared the obtained parametrizations. Mathematically, assume

$\mathbf{X}=(\mathbf{x}_1^T, \dots, \mathbf{x}_n^T)^T$ is the original data cloud. Let $\Phi(\mathbf{X}):=\mathbf{XUD}$, where $\mathbf{U}$ is a randomly generated 3×3 rotation matrix⁵ and $\mathbf{D}=\text{diag}(d_1, d_2, d_3)$ where all d'_i s, $i=1, 2, 3$ are uniformly generated from $U(0.9, 1.1)$. For one scan, we randomly generated 50

transformations Φ_k , $k=1, \dots, 50$. Let $\mathbf{T}=(\mathbf{t}_1^T, \dots, \mathbf{t}_n^T)^T$ be the 2D parametrization result of the original corpus callosum and $\mathbf{T}_k$ be the 2D parametrization result of $\Phi_k(\mathbf{X})$. Then

$MSE_k:=\frac{1}{n} \sum_i \left\{ \|\mathbf{t}_{ik}-\mathbf{t}_i\|^2 \right\}$ is calculated to evaluate the mismatch of the parametrization between the original data cloud and morphed data cloud. Ideally, zero MSE suggests zero misalignment which means the proposed method is invariant to rotation and a restricted scaling. Notice that the parametrization in the proposed method is restricted to $[0, 1]$ and minimal mismatch in each axis is 0.01, so that a 10^{-4} MSE would suggest that on average the 2D parametrization moves one unit distance in 2D space. When the MSE is large, say over 0.09, it means that averagely the parametrization moves more than 0.3 distance (30 units) when all points are flipped over, or there are large mismatches in the 2D spaces of the morphed shape. Among 50 MSE's using ISOMAP initialization, the median is 8.97×10^{-5} , the mean is 1.81×10^{-4} and the maximum is 2.26×10^{-3} . In other words, under moderate rotation and relatively mild scaling, our algorithm can achieve a good parametrization consistency. More complex situations remain to be explored.

⁵ 3×3 rotation matrices are generated as $\mathbf{U}=\mathbf{U}_x\mathbf{U}_y\mathbf{U}_z$ where

$$\mathbf{U}_x=\begin{Bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) \\ 0 & \sin(\theta_x) & \cos(\theta_x) \end{Bmatrix}, \mathbf{U}_y=\begin{Bmatrix} \cos(\theta_y) & 0 & \sin(\theta_y) \\ 0 & 1 & 0 \\ -\sin(\theta_y) & 0 & \cos(\theta_y) \end{Bmatrix} \text{ and } \mathbf{U}_z=\begin{Bmatrix} \cos(\theta_z) & -\sin(\theta_z) & 0 \\ \sin(\theta_z) & \cos(\theta_z) & 0 \\ 0 & 0 & 1 \end{Bmatrix}. \theta_x, \theta_y \text{ and } \theta_z \text{ are generated uniformly from } -0.1\pi \text{ to } 0.1\pi.$$

5. DISCUSSION

In this manuscript a principal surface algorithm was introduced and used to fit the corpus callosum. The goal of this work is largely developmental, creating a handy tool for dimension reduction in morphological analysis of primary brain structures. In simulations, the proposed algorithm performed superbly. While applied to the corpus callosum, the algorithm could be applied to any other three dimensional manifold-like objects where a two dimensional surface could be embedded and is of interest, for example, some of the major white matter tracts in the brain (Bazin et al. 2011). The two dimensional manifold characterizes the original data and accomplishes both dimension reduction and better visualization. The surface that was constructed is smooth and could be easily projected onto to represent other properties of the original structure, such as the FA, mean diffusivity, parallel diffusivity, local thickness and so on. The algorithm is computationally feasible and scales well to larger images and densely measured structures.

The role of the two dimensional sub-representation needs to be further explored. Before the sub-representation routinely practised in any statistics analysis, the boundary issue mentioned in Section 3 should be carefully tackled. Due to the local averaging step, the obtained surface is a shrinkage of the original data cloud. Therefore, more points are projected onto the boundary on the domain $[0, 1] \times [0, 1]$. In applications, these boundary points could be deleted since the 2D representations on the boundary are less representative.

The corpus callosum in our application is anatomically simpler than the one considered in Yushkevich et al. (2008). Therefore, it remains to be explored how our principal surface algorithm performs in structures of higher complexity. O'Donnell et al. (2013) and Jin et al. (2014) provided powerful methods for white matter tract clustering. However, by focusing on fiber clustering, these procedures differ fundamentally from ours. Thus it remains interesting to analyze the white matter tracts on the 2D representation for each cluster. For future work, we are developing functional data analysis tools (Goldsmith et al. 2011b; Goldsmith et al. 2011c; Greven et al. 2011; Zhu et al. 2010; Zhu et al. 2011) for relating the dimension reduced 2D manifold to outcomes of interest for the purpose of inference, biomarker creation and prediction. Of note, we are particularly interested in whether or not the 2D representation of the corpus callosum is less sensitive to issues of whole brain registration often used in the processing pipeline. In fact, it is possible that registering the 2D representation is preferable to whole brain registration a priori in certain applications. Note also that MFPCA (Di et al. 2009) and LFPCA (Zipunnikov et al. 2011; Zipunnikov et al. 2014), methods have been shown to isolate registration error as a part of the model (Lee et al. 2015), thus raising the intriguing possibility of DTI processing streams that dramatically decrease the need and importance of whole brain template-based registration.

Furthermore, current work could be easily modified to achieve dimension reduction for different input and output dimensions.⁶ For example, sagittal mid-line callosum average thickness is meaningful in the study of neurological diseases (Luders et al. 2009; Vidal et al. 2006). In addition, Zhang et al. (2010) used medial models for obtaining 2D thickness maps

⁶A 3D to 1D principal curve algorithm could be easily achieved by changing the input and output dimension in TPS fitting

of corpus callosum. With the proposed method, CC mid-sagittal thickness or FA curves as well as the CC thickness or FA on principal surfaces could be achieved simultaneously.

Finally, consider that the proposed procedure can be further extended as a general problem of fitting skeleton manifolds. We are also interested in fitting surfaces with fixed boundaries. Work has been done to analyze principal curves with fixed origin and the end point (Caffo et al. 2008). The extension to surfaces seems to be quite challenging. A possible route is follows. Suppose $\underline{t} = (t_1, t_2)$ is the corresponding coordinate on the surface of the original data point, $\underline{x} = (x_1, x_2, x_3)$, which lies in 3D space. Now consider a predetermined function for $\underline{x}(\underline{t}) = (x_1(\underline{t}), x_2(\underline{t}), x_3(\underline{t}))$ when the boundary values of t are linearly constrained. Such constraints would yield cylindrical fits easily though extensions to completely closed surfaces would require more elaborate constraints.

Acknowledgments

The project described was supported by the NIH grant RO1 Number R01EB012547 and NIH grant P41 EB015909 from the National Institute of Biomedical Imaging And Bioengineering. The project described was also supported by Grant Number R01NS060910 from the National Institute of Neurological Disorders and Stroke. The project described was partially supported by grant NINDS R01 NS070906. Data for this study was acquired with grants from the National Multiple Sclerosis Society and EMD Serono. In addition, this research was partially supported by the Intramural Research Program of the National Institute of Neurological Disorders and Stroke, National Institutes of Health.

References

- Bazin P, Ye C, Bogovic J, Shiee N, Reich D, Prince J, Pham D. Direct segmentation of the major white matter tracts in diffusion tensor images. *NeuroImage*. 2011; 58(2):458–468. [PubMed: 21718790]
- Caffo B, Crainiceanu C, Deng L, Hendrix C. A case study in pharmacologic colon imaging using principal curves in single-photon emission computed tomography. *Journal of the American Statistical Association*. 2008; 103(484):1470–1480. [PubMed: 20016767]
- Di C, Crainiceanu C, Caffo B, Punjabi N. Multilevel functional principal component analysis. *Annals of Applied Statistics*. 2009; 3(1):458–488. [PubMed: 20221415]
- Dong D, McAvoy T. Nonlinear principal component analysis based on principal curves and neural networks. *Computers Chemical Engineering*. 1996; 20(1):65–78.
- Duchon J. Splines minimizing rotation-invariant semi-norms in sobolev spaces. *Lecture notes in mathematics*. 1977; 571/1977:85–100.
- Einbeck J, Evers L, Powell B. Data compression and regression through local principal curves and surfaces. *International journal of neural systems*. 2010; 20(3):177–192. [PubMed: 20556846]
- Fletcher P, Lu C, Pizer S, Joshi S. Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*. 2004; 23(8):995–1005.
- Fletcher, PT. PhD thesis. Citeseer; 2004. Statistical variability in nonlinear spaces: Application to shape analysis and DT-MRI.
- Gerber, S.; Tasdien, T.; Whitaker, R. Dimensionality reduction and principal surfaces via kernel map manifolds. *Computer Vision, 2009 IEEE 12th International Conference*; 2009. p. 529-536.
- Gnanadesikan, R. *Methods for statistical data analysis of multivariate observations*. Vol. 321. John Wiley & Sons; 2011.
- Goldsmith J, Caffo B, Crainiceanu C, Reich D, Du Y, Hendrix C. Nonlinear tube-fitting for the analysis of anatomical and functional structures. *The annals of applied statistics*. 2011a; 5(1):337–363. [PubMed: 21709824]
- Goldsmith J, Crainiceanu C, Caffo B, Reich D. Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*. 2011b; 57(2):431–439. [PubMed: 21554962]

- Goldsmith J, Crainiceanu C, Caffo B, Reich D. Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2012; 61(3):453–469.
- Goldsmith J, Wand M, Crainiceanu C. Functional regression via variational Bayes. *Electronic Journal of Statistics*. 2011c; 5:572–602. [PubMed: 22163061]
- Greven S, Crainiceanu C, Caffo B, Reich D. Longitudinal functional principal component analysis. *Recent Advances in Functional Data Analysis and Related Topics*. 2011:149–154.
- Hastie T. Principal curves and surfaces. Technical report. 1984:11.
- Hastie T, Stuetzle W. Principal Curves. *Journal of the American Statistical Association*. 1989; 84(406): 502–516.
- Jin Y, Shi Y, Zhan L, Gutman BA, de Zubicaray GI, McMahon KL, Wright MJ, Toga AW, Thompson PM. Automatic clustering of white matter fibers in brain diffusion MRI with an application to genetics. *Neuroimage*. 2014; 100:75–90. [PubMed: 24821529]
- Jung S, Foskey M, Marron J. Principal arc analysis on direct product manifolds. *The Annals of Applied Statistics*. 2011; 5(1):578–603.
- Kohonen T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*. 1982; 43:59–69.
- Kohonen T. The self-organizing map. *Proceedings of the IEEE*. 1990; 78(9):1464–1480.
- Kramer M. Nonlinear principal component analysis using autoassociative neural networks. *AICHE Journal*. 1991; 37(2):233–243.
- Kramer MA. Autoassociative neural networks. *Computers and Chemical Engineering*. 1992; 16(4): 313–328. Neural network applications in chemical engineering.
- Leblanc M, Tibshirani R. Adaptive Principal Surfaces. *Journal of the American Statistical Association*. 1994; 89(425):53–64.
- Lee S, Zipunnikov V, Reich D, Pham D. Statistical image analysis of longitudinal RAVENS images. *Frontiers in neuroscience*. 2015:9. [PubMed: 25688184]
- Luders E, Narr K, Hamilton L, Phillips O, Thompson P, Valle J, Del'Homme M, Strickland T, McCracken J, Toga A, Levitt J. Decreased Callosal Thickness in Attention-Deficit/Hyperactivity Disorder. *Biological Psychiatry*. 2009; 65(1):84–88. [PubMed: 18842255]
- O'Donnell L, Golby A, Westin C. Fiber clustering versus the parcellation-based connectome. *NeuroImage*. 2013; 80:283–289. [PubMed: 23631987]
- Ozertem U, Erdogmus D. Locally defined principal curves and surfaces. *Journal of Machine learning research*. 2011; 12:1249–1286.
- Ozturk A, Smith S, Gordon-Lipkin E, Harrison D, Shiee N, Pham D, Caffo B, Calabresi P, Reich D. MRI of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis*. 2010; 16(2):166–177. [PubMed: 20142309]
- Palus M, Dvorak I. Singular-value decomposition in attractor reconstruction: Pitfalls and precautions. *Physica D: Nonlinear Phenomena*. 1992; 55(12):221–234.
- R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2008.
- Reich D, Ozturk A, Calabresi P, Mori S. Automated vs. conventional tractography in multiple sclerosis: Variability and correlation with disability. *NeuroImage*. 2010; 49:3047–3056. [PubMed: 19944769]
- Smith S, Jenkinson M, Johansen-Berg H, Rueckert D, Nichols T, Mackay C, Watkins K, Ciccarelli O, Cader Z, Matthews P, et al. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*. 2006; 31(4):1487–1505. [PubMed: 16624579]
- Tenenbaum J, De Silva V, Langford J. A global geometric framework for nonlinear dimensionality reduction. *Science*. 2000; 290(5500):2319–2323. [PubMed: 11125149]
- Vidal C, Nicolson R, DeVito T, Hayashi K, Geaga J, Drost D, Williamson P, Rajakumar N, Sui Y, Dutton R, et al. Mapping corpus callosum deficits in autism: an index of aberrant cortical connectivity. *Biological Psychiatry*. 2006; 60(3):218–225. [PubMed: 16460701]
- Weinberger K, Saul L. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*. 2006; 70(1):77–90.

- Wood S. Thin plate regression splines. *Journal of the royal statistical society*. 2003; 65(1):95–114.
- Yushkevich P, Zhang H, Simon T, Gee J. Structure-specific statistical mapping of white matter tracts. *NeuroImage*. 2008; 41(2):448–461. [PubMed: 18407524]
- Zhang H, Awate S, Das S, Woo J, Melhem E, Gee J, Yushkevich P. A tractspecific framework for white matter morphometry combining macroscopic and microscopic tract features. *Medical image analysis*. 2010; 14(5):666–673. [PubMed: 20547469]
- Zhu H, Kong L, Li R, Styner M, Gerig G, Lin W, Gilmore J. FADTTS: Functional analysis of diffusion tensor tract statistics. *NeuroImage*. 2011; 56(3):1412–1425. [PubMed: 21335092]
- Zhu H, Styner M, Tang N, Liu Z, Lin W, Gilmore J. Frats: Functional regression analysis of dti tract statistics. *Medical Imaging, IEEE Transactions on*. 2010; 29(4):1039–1049.
- Zipunnikov V, Caffo B, Yousem D, Davatzikos C, Schwartz B, Crainiceanu C. Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics*. 2011; 20(4):852–873. [PubMed: 25960627]
- Zipunnikov V, Greven S, Shou H, Caffo B, Reich D, Crainiceanu C, et al. Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis. *The Annals of Applied Statistics*. 2014; 8(4):2175–2202. [PubMed: 25663955]

Figure 1. 3D renderings and 2D projections of the corpus callosum of one subject. Three panels represent sagittal, frontal and horizontal 2D Slices respectively. Red and blue colors indicate higher and lower FA values on 2D slices, respectively.

Figure 2.

Left panel is a illustration of different dimension reduction methods. The blue points are the original data points, the dot-dashed green line is the first principal component, the dashed black curve is the spline fitting and the solid red curve is one of the principal curves. Right panel shows the difference between the principal curve and the regression method: top panel shows that the principal curve minimizes the orthogonal distance, bottom panel shows that the spline regression minimize the distance in y axis. Both panels in the middle use the same dataset.

Figure 3.

Simulation results. In all panels, original data points are shown in green and the fitted surfaces are shown in blue. The panels from top-left to bottom-right show the results for our method with 1,000 sub-sampling points, initiated by ISOMAP with the three, four, four and five basis functions respectively implementing on (1) a non-function shaped “carpet” data cloud, (2) an cylinder with an open seam on one side, (3) the CC-shaped structure and (4) a simulated digit “5” data cloud.

Figure 4.

Principal surface fitting of digit “5” results comparison between different methods. Upper left shows the fitting result of the proposed method initiated by principal scores from PCA, upper right panel shows the result of the proposed method initiated by ISOMAP, bottom left panel shows the result of ISOMAP and bottom right panel shows the result of HS’s principal surfaces.

Figure 5.

Six panels demonstrate the fitting results of a Corpus Callosum-shaped structure with different number of basis in the TPS fitting. From top left to bottom right, the number of basis increases from $k = 20$ to $k = 25$.

Figure 6.

Parametrization results in each scenario. We colored the original points from one side to the other continuously using dark red to yellow. Each image shows the final parametrization space with each data's original labeled color. From top to bottom, *Top*: The proposed method; *Middle*: ISOMAP; *Right*: HS's method. From left to right, *Left*: Curved "Carpet" structure; *Second from left*: Cylinder structure; *Third from left*: Corpus Callosum-shaped structure; *Right*: Digit "five".

Figure 7. Principal surface fitting results of unregistered corpus callosum data. Left panels show the fitted surface with the original data cloud and right panels show the results of fitted surface.

Figure 8. Principal surface fitting results of registered corpus callosum data. Left panels show the fitted surface with the original data cloud and right panels show the results of fitted surface.

Figure 9.

2D parametrization of corpus callosum. Top row shows 2D parametrization from the proposed principal surface algorithm; middle row shows the result from ISOMAP and bottom row shows the result of HS's principal surface. Bright color indicates higher FA values and darker color indicates lower FA values.

Table 1

Curvature and thickness comparison between different number of basis functions. Thickness is in parenthesis.

Number of Basis functions	Scenarios			
	Carpet	Cylinder	Corpus	Five
4	26.174 (0.152) 0.204	41.703 (0.035) 0.054*	46.051 (0.031) 0.058	30.065 (0.017) 0.041*
5	32.824 (0.146) 0.264	77.866 (0.031) 0.147	48.592 (0.031) 0.058*	31.228 (0.016) 0.046
6	69.522 (0.124) 0.252	94.234 (0.023) 0.138	56.851 (0.030) 0.063	36.454 (0.012) 0.075
7	132.800 (0.093) 0.196	118.111 (0.020) 0.127	60.712 (0.027) 0.170	38.121 (0.012) 0.050
8	197.029 (0.076) 0.193*	136.875 (0.018) 0.150	59.546 (0.027) 0.154	43.552 (0.011) 0.064
9	259.639 (0.063) 0.214	164.889 (0.017) 0.128	64.241 (0.026) 0.167	45.822 (0.010) 0.045
10	308.302 (0.056) 0.237	189.778 (0.016) 0.127	65.219 (0.026) 0.143	49.801 (0.010) 0.065

Table 2

MSE comparisons between different number of sub-sampling points. The unit is 10^{-3} .

Sub-sample	Carpet		Cylinder		Corpus		Five	
	PCA	ISOMAP	PCA	ISOMAP	PCA	ISOMAP	PCA	ISOMAP
500	56.6	33.9	44.4	31.6	25.2	12.2	10.4	10.9
1,000	55.4	27.3	32.6	27.2	20.5	11.3	9.9	10.4
1,500	55.0	27.1	33.3	28.9	19.5	10.6	9.8	10.2

Table 3

MSE comparison between our method and HS, ISOMAP for all four scenarios. The unit is 10^{-3} .

Methods	Scenarios			
	Carpet	Cylinder	Corpus	Five
PS (PCA)	16.9	9.6	11.3	4.3
PS (ISOMAP)	16.3	11.5	2.5	2.1
HS	35.4	50.4	66.2	21.4
ISOMAP	94.3	476.7	296.9	164.6