ANALYZING SAMPLES OF PRODUCT MANIFOLDS USING DIFFUSION MAPS

SHARON ZHANG

Abstract.

1. Introduction

Suppose we have some data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ ~~$\in \mathcal{M}$~~ which are independent identically distributed samples from a distribution over some manifold $\mathcal{M} \subset \mathbb{R}^D$. The D dimensions encode information about all variables present in our data, but often times these variables need not be related ~~on~~ each other. In particular, if we have several independent variables in our data, we might assume that our ambient manifold $\mathcal{M}$ in $\mathbb{R}^D$ is the product of several independent submanifolds with dimension less than D. Our observations of $\{\mathbf{x}_i\}_{i=1}^n$ lie in the ambient space $\mathbb{R}^D$, but it is desirable to be able to analyze certain independent variables in their lower dimensional intrinsic manifold rather than the D-dimensional ambient manifold. Moreover, we might also want to recover the geometry of these lower-dimensional manifolds, so that we can learn more information about the indepdendent variable to which it is associated.

1.1. Cryogenic-electron microscopy.

Cryogenic-electron microscopy (cryo-EM) is a technique commonly used in biology for imaging molecules. Before the image is taken, samples are frozen at cryogenic temperatures, enabling them to be captured in their natural environment. The frozen particles are then visualized using an electron microscope to produce a *micrograph*, which contains many instances of image patches with molecules. The images produced by cryo-EM are two-dimensional tomographic projections of the particle's electrostatic potential. One goal of collecting these two-dimensional images is to create a three-dimensional reconstruction of the particle.

Due to the low signal-to-noise (SNR) ratio of these micrographs, many images of identical molecules are required in order to reconstruct a single molecule. Each sample is prepared by being frozen extremely rapidly in a thin layer of ice. During this process, the molecules are trapped in a random position and orientation. Thus the resulting image data for a single molecule contains information about many possible variables. One predominant variable is the inherent orientation of the particular molecule in the image, which may affect the appearance of identical molecules over different images.

1.2. Two reconstruction problems.

We can model the entire electrostatic potential of the particle as a function $\phi : \mathbb{R}^3 \to \mathbb{R}$. An image I of a molecules is obtained by applying the following transformations on ϕ:

(i) Applying a rotation $R \in$ ~~SO_3~~;

(ii) projecting in the z-direction;
(iii) convolving with a point-spread function H;
(iv) sampling on an $L \times L$ Cartesian grid of pixels;
(v) and adding noise.

Formally, we can write the final image as

$$I(x, y) = H \star PR \circ \phi + \omega \tag{1.1}$$

where P is the tomographic projection operator in the z-direction, $Pf(x, y) = \int_{-infty}^{\infty} f(x, y, z)dz$, and $R \circ \phi = \phi(R^T r)$, $r = (x, y, z)^T$. The classic cryo-EM reconstruction problem is posed as follows: given n images $I_1, \dots I_n$ and the corresponding point spread functions $H_1, \dots, H_n$, how can we recover ϕ without knowing $R_1, \dots, R_n$?

An even harder problem is known as the *heterogeneity* cryo-EM problem. The challenge in this variation arises from potential structural variations within the molecules. Namely, for different images I_i we may potentially capture different molecular conformations $\phi_1, \dots, \phi_n$. The heterogeneity cryo-EM problem requires restrictive assumptions in order to ensure that the number of output parameters in $\{\phi_i\}_{i=1}^n$ can be solved for using the number of input parameters given by $\{I_i\}_{i=1}^n$, $\{H_i\}_{i=1}^n$. One reasonable assumption is that the variability of the conformations $\{\phi_i\}_{i=1}^n$ is continuous, and thus we are able to model the conformations by a manifold. We may also assume *discrete heterogeneity*, in which we assume that $\{\phi_i\}_{i=1}^n$ come from a small fixed set of conformations.

In this paper, we explore a potential technique to isolate information about the intrinsic manifolds from the ambient manifold via spectral analysis methods. We first provide a brief exposition of diffusion maps and the Laplacian operator, then we describe the method of separation, and finally we present the results of our method on some concrete examples.

2. Theory

We begin by reviewing the definitions of the continuous Laplacian, as well as a few of its properties. We then examine how the Laplacian decomposes over product manifolds. Lastly, we go over results which relate the discrete Laplacian to the continuous Laplacian, and some spectral techniques for data analysis which make use of all these results.

2.1. The Laplace Operator.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice-differentiable function in Euclidean space. The Laplacian (or "Laplace operator") of a manifold, often denoted Δ, is a differential operator defined by

$$\Delta f = \nabla \cdot \nabla f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i^2} \tag{2.1}$$

where ∇ is the divergence operator and ∇f is the gradient of f. Informally, the Laplacian can be interpreted as a measure of the density of gradient flow at a certain point. As such, it is the backbone of many important physical equations, such as Poisson's Equation, which relates the electric potential to the electric charge density of a medium, and the wave equation, which describes the propagation of oscillations.

The *spectrum* of the Laplacian is defined as the set of eigenvalues and associated eigenfunctions which satisfy the Helmholtz equation,

$$-\Delta f = \lambda f, \tag{2.2}$$

also known as the "Laplace eigenproblem." In particular, the Helmholtz equation, together with any imposed boundary conditions, can be posed as a Sturm-Liouville problem, which is of the form

$$\frac{d}{dx}\left(p(x)\frac{df}{dx}\right) + q(x)f = -\lambda w(x) f. \tag{2.3}$$

The main results of Sturm-Liouville theory tell us that the Laplacian has an infinite set of eigenvalues

$$0 = \lambda_1 \leq \lambda_2 \leq \ldots \to \infty, \tag{2.4}$$

each with a corresponding eigenfunction, which is unique up to a constant multiple. These eigenfunctions are known as the Laplace eigenfunctions over the manifold, and they are extremely nice functions to work with because they are infinitely differentiable over the interior of the manifold. Moreover, if $\Omega \subset \mathbb{R}^n$ is compact, then the Laplace eigenfunctions form an orthogonal basis of $L^2(\Omega)$ with respect to the standard inner product.

The following examples derive the Laplace eigenfunctions for some simple one-dimensional manifolds.

Example 2.5 (A line)**.** Consider the Laplace eigenproblem over the line $l = [0, a]$

$$\Delta u + \lambda u = u'' + \lambda u = 0, \quad 0 < x < a \tag{2.6}$$

with Neumann boundary conditions

$$u'(0) = u'(a) = 0. \tag{2.7}$$

Both $u(x) = \cos(\alpha x)$ and $u(x) = \sin(\alpha x)$ are solutions to (2.6). By (2.7), we can eliminate $\sin(\alpha x)$. Furthermore, we must have

$$-\alpha \sin(\alpha a) = 0 \tag{2.8}$$

so $\alpha a = k\pi$ for all integers k. Then $\alpha = \frac{k\pi}{a}$, so our eigenfunctions are

$$\cos\left(\frac{k\pi}{a}x\right), \quad k = 0, 1, 2, \ldots \tag{2.9}$$

with eigenvalues

$$\lambda_k = \frac{k^2}{a^2\pi^2}. \tag{2.10}$$

Example 2.11 (A disk)**.** Let Ω be a disk with radius R. We reparametrize the Laplcian to polar coordinates

$$r = \sqrt{x^2 + y^2}, \quad \theta = \tan^{-1}\left(\frac{y}{x}\right) \tag{2.12}$$

to get

$$u_{rr} + \frac{1}{2}u_r + \frac{1}{r}u_{\theta\theta} + \lambda u = 0, \quad 0 \leq r < R \tag{2.13}$$

with Neumann boundary conditions

$$u_r(R, \theta) = 0. \tag{2.14}$$

A natural starting point is separation of variables, as we have two independent variables r and θ. This gives us $u(r,\theta) = R(r)\Theta(\theta)$, so

$$R''\Theta + \frac{R'}{r}\Theta + \frac{R}{r^2}\Theta'' = -\lambda R\Theta. \tag{2.15}$$

We can collect variables to get

$$r^2\frac{R''}{R} + r\frac{R'}{R} + \lambda r^2 = \frac{\Theta''}{\Theta} = \alpha^2. \tag{2.16}$$

Note that the the constant α^2 must be nonnegative by our discussion above on (2.4). The solutions to the left side of (2.16) are of the form $A_k\cos(k\theta)+B_k\sin(k\theta)$. There are no boundary conditions on Θ, but we must also have $A_k\cos(k\theta) + B_k\sin(k\theta) = A_k\cos(k(\theta+2k\pi)) + B_k\sin(k(\theta+2k\pi)) = A_k\cos(k\theta+2km\pi) + B_k\sin(k\theta+2km\pi)$, so k must be an integer. Thus the solutions for Θ are

$$\Theta_k(\theta) = A_k\cos(k\theta) + B_k\sin(k\theta), \quad k = 0,1,2,\dots \tag{2.17}$$

The right hand side of the equation can be rewritten as

$$r^2R'' + rR' + (\lambda r^2 - \alpha^2)R = 0. \tag{2.18}$$

We consider substitutions of the form $R(r) = J(\sqrt{\lambda}r)$ which rescale r. This gives us

$$R(r) = J(\sqrt{\lambda}) \tag{2.19}$$
$$R'(r) = \sqrt{\lambda}J'(\sqrt{\lambda}r) \tag{2.20}$$
$$R''(r) = \lambda J''(\sqrt{\lambda}r) \tag{2.21}$$

Then (2.18) is

$$\lambda r^2J''(\sqrt{\lambda}r) + \sqrt{\lambda}rJ' + (\lambda r^2 - \alpha^2)J = 0. \tag{2.22}$$

We can substitute $\beta = \sqrt{\lambda}r$, so (2.18) becomes

$$\beta^2J''(\sqrt{\lambda}r) + \beta rJ' + (\beta^2 - \alpha^2)J = 0. \tag{2.23}$$

The above is known as Bessel's equation. The solutions to Bessel's equation with the imposed Neumann boundary conditions are the kth Bessel functions of the first-kind,

$$J_k(x) = \sum_{l=0}^{\infty}\frac{(-1)^l}{l!(k+l)!}\left(\frac{x}{2}\right)^{k+2l}, \tag{2.24}$$

and either $J_k'(\sqrt{\lambda}r) = 0$ or $\lambda = 0$.

Combining (2.17) with (2.24), we get our eigenfunctions

$$u(r,\theta) = \Theta_k(\theta)J_k(\sqrt{\lambda}r). \tag{2.25}$$

2.2. **The Laplace operator over product spaces.** Suppose we have a Euclidean manifold $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$. Let $f : \mathcal{M}_1 \to \mathbb{R}^n$, $g : \mathcal{M}_2 \to \mathbb{R}^n$. We can then embed f and g as functions over $\mathcal{M}$ using the canonical embeddings of $\mathcal{M}_1$ and $\mathcal{M}_2$ into $\mathcal{M}$. If f and g are eigenfunctions of $\mathcal{M}_1$ and $\mathcal{M}_2$, i.e. $\Delta_{\mathcal{M}_1} f = \lambda f$ and $\Delta_{\mathcal{M}_2} g = \mu g$, then

$$\begin{aligned}
\Delta_{\mathcal{M}}(fg) &= f\Delta_{\mathcal{M}} g + g\Delta_{\mathcal{M}} f + 2\langle \Delta_{\mathcal{M}} f, \Delta_{\mathcal{M}} g\rangle_{\mathcal{M}} && (2.26)\\
&= f\Delta_{\mathcal{M}_\in} g + g\Delta_{\mathcal{M}_\infty} f && (2.27)\\
&= \lambda fg + \mu fg && (2.28)\\
&= (\lambda + \mu) fg. && (2.29)
\end{aligned}$$

Moreover, if $\mathcal{M}_1$ and $\mathcal{M}_2$ are compact, then $\{f_i\}_{i=0}^{\infty}$ and $\{g_i\}_{i=0}^{\infty}$ form orthonormal bases of $L^2(\mathcal{M}_1)$ and $L^2(\mathcal{M}_2)$. Then the set of functions $\{f_i, g_j\}_{i,j}$ forms a basis of eigenfunctions for $L^2(\mathcal{M})$, with eigenvalues $\{\lambda_i + \mu_j\}_{i,j}$.

Using this, we can generalize the result in (2.26) through (2.29) to the product of many manifolds. That is, given a Euclidean manifold $\mathcal{M} = \prod_{i=1}^{m} \mathcal{M}_i$, where $\{f_j^{(i)}\}_j$ are the Laplace eigenfunctions of $\mathcal{M}_i$ (for given boundary conditions), then the set

$$\bigcup_{i=1}^{m} \{f_j^{(i)}\}_j \tag{2.30}$$

forms a basis of eigenfunctions for $\mathcal{M}$. The eigenvalue of $f^{(k_1,\ldots,k_m)} = \prod_{i=1}^{m} f_{k_j}^{(i)}$ is $\lambda^{(k_1,\ldots,k_m)} = \sum_{i=1}^{m} \lambda_{k_j}^{(i)}$. Furthermore, since $\lambda_0^{(i)} = 0$ and $f_0^{(i)} = \mathbf{1}$ for all $i = 1, \ldots, m$, so $\lambda_0 = 0$ and $f_0 = \mathbf{1}$ as well. This is consistent with (2.4).

In the following examples, we derive the Laplace eigenfunctions of some simple product manifolds.

Example 2.31 (Rectangle)**.** Consider a rectangle $\Omega = [0, a] \times [0, b]$. The Laplace eigenproblem is

$$\Delta u + \lambda u = u_{xx} + u_{yy} + \lambda u = 0, \quad 0 < x < a, \quad 0 < y < b \tag{2.32}$$

with Neumann boundary conditions

$$u_x(0, y) = u_x(a, y) = 0, \quad u_y(x, 0) = u_y(x, b) = 0. \tag{2.33}$$

We start with separation of variables. Let $u = X(x)Y(y)$, so the Laplace equation becomes

$$X''Y + Y''X = -\lambda XY \tag{2.34}$$

and so

$$\frac{X''}{X} = -\frac{Y''}{Y} - \lambda = -\alpha^2. \tag{2.35}$$

Thus

$$\frac{X''}{X} = -\alpha^2, \quad \frac{Y''}{Y} = \lambda - \alpha^2 = -\beta^2 \tag{2.36}$$

and this gives us

$$X'' = -\alpha^2 X, \quad Y'' = -\beta^2 Y \tag{2.37}$$

subject to boundary conditions

$$X'(0) = X'(a) = 0, \quad Y'(0) = Y'(b) = 0. \tag{2.38}$$

The familiar solutions are $X = \cos(kx)$, $Y = \cos(ly)$. Furthermore, our boundary conditions require that $ka = m\pi$ and $lb = n\pi$ for $m, n = 0, 1, 2, \ldots$ so we have $k = \frac{m\pi}{a}$ and $l = \frac{n\pi}{b}$ for integers m, n. Therefore our separated solutions are

$$X_m(x) = A\cos\left(\frac{m\pi}{a}x\right), \quad Y_n(y) = B\cos\left(\frac{n\pi}{b}y\right) \tag{2.39}$$

with eigenvalues

$$\lambda_m = \frac{m^2\pi^2}{a^2}, \quad \lambda_n = \frac{n^2\pi^2}{b^2}, \tag{2.40}$$

respectively. Combining these solutions gives us our final solutions

$$u_{m,n}(x, y) = A_{m,n}\cos\left(\frac{m\pi}{a}x\right)\cos\left(\frac{n\pi}{b}y\right) \tag{2.41}$$

with eigenvalues

$$\lambda_{m,n} = \lambda_m + \lambda_n = \pi^2\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right) \tag{2.42}$$

for all possible combinations of nonnegative integers m and n. In particular, these are precisely the combinations of eigenfunctions sampled over two independent lines of length a and b.

In Figure 1, we show the first 100 nontrivial eigenvectors calculated from $n = 10000$ uniform random samples over the rectangle $[0, 1] \times [0, \sqrt{\pi + 2}]$. As discussed before, these eigenvectors are discrete approximations of the eigenfunctions computed above.

Example 2.43 (Hollow cylinder)**.** The Laplace eigenproblem for a hollow cylinder is similar to that of the rectangle, but we have slightly different boundary conditions. The problem is stated as

$$\Delta u + \lambda u = u_{xx} + u_{\theta\theta} + \lambda u = 0, \quad 0 < x < a, \quad 0 < y < b \tag{2.44}$$

with Neumann boundary conditions

$$u_x(0, \theta) = u_x(a, \theta) = 0. \tag{2.45}$$

Again, we first use separation of variables to write $u(x, \theta) = X(x)\Theta(\theta)$. The solution for X is identical to the example over a rectangle, and the solutions for Θ are the same as those in Examples 2.11. This gives us

$$X_m(x) = \cos\left(\frac{m\pi}{l}x\right), \quad \Theta_n(\theta) = A_n\cos(n\theta) + B_n\sin(n\theta) \tag{2.46}$$

with eigenvalues

$$\lambda_m = \frac{m^2\pi^2}{l^2}, \quad \lambda_n = n^2, \tag{2.47}$$

respectively. The Laplace eigenfunctions of the hollow cylinder are then

$$u_{m,n}(x, \theta) = A_n\cos\left(\frac{m\pi}{l}x\right)\cos(n\theta) + B_n\cos\left(\frac{m\pi}{l}x\right)\sin(n\theta) \tag{2.48}$$

with eigenvalues

$$\lambda_{m,n} = \lambda_m + \lambda_n = \frac{m^2\pi^2}{l^2} + n^2. \tag{2.49}$$

Figure 2 shows the first 100 nontrivial eigenvectors calculated from $n = 10000$ uniform random samples over the hollow cylinder of radius r and height l.

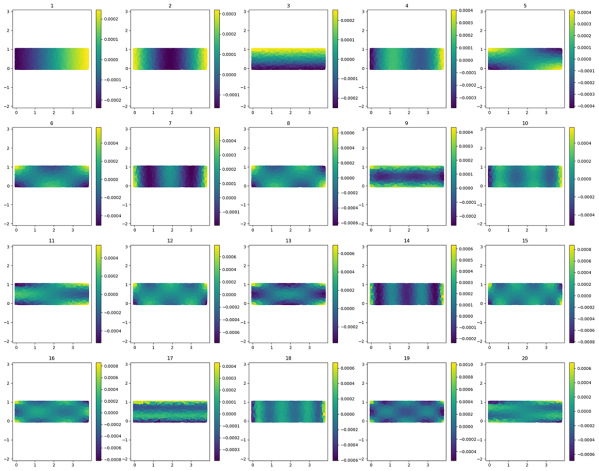


FIGURE 1. The 20 most significant (lowest frequency) Laplace eigenvectors of data sampled over a rectangle $\Omega = l_1 \times l_2$. We can easily pick out the base eigenvectors of both underlying lines which form the rectangle, as they distinctively approximate cosines of increasing frequency, which are exactly the Laplace eigenfunctions of the line. The remaining eigenvectors are mixture eigenvectors of base eigenvectors from l_1 and l_2. For example, eigenvector 5 is a mixture of eigenvector 1 and eigenvector 3.

2.3. **Graph Laplacians and diffusion maps.** The results of the previous sections hold for the continuous Laplacian, but in practice we are rarely truly in the continuous case. In this section, we introduce the graph Laplacian, which is a discrete version of the continuous Laplacian. While data is inherently discrete, we will show that for large sample sizes the discrete analogs of the spectrum of the Laplacian provide good approximations of the actual spectrum. We also show the relationship between the graph Laplacian and diffusion maps, which is a spectral method of dimensionality reduction.

The discrete Laplacian is defined over a graph (G, E, V), which we can interpret as a representation of $\mathbf{x}_1, \dots \mathbf{x}_n \in \mathbb{R}^D$. In order to capture the relationships within our data, we define a weight matrix W where

$$W_{ij} = k\left(\frac{||x_i - x_j||^2}{2\sigma}\right), \tag{2.50}$$

with k being the Gaussian kernel. We can then define the random walk matrix $A = D^{-1}W$, where D is the diagonal degree matrix defined by

$$D_{ii} = \sum_{j=1}^{m} W_{ij}. \tag{2.51}$$

The matrix A describes the probability cloud of a random walk on G, which can be used as a model for the connectivity between points in the data. The idea is that a

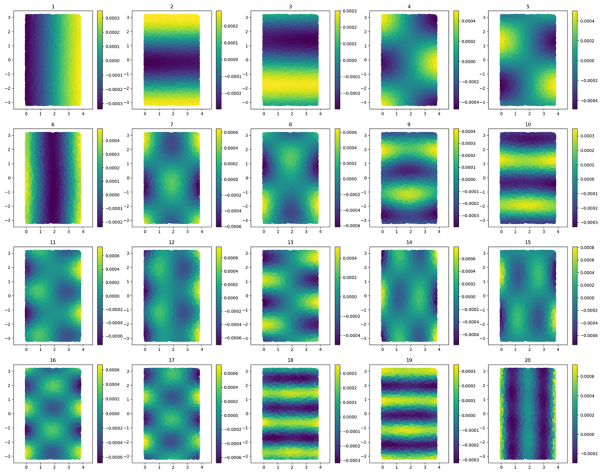


FIGURE 2. The 20 most significant eigenvectors of data sampled over a hollow cylinder, which can be represented by $\Omega = l \times \theta$. The base eigenvectors of l approximate cosine functions from left to right, and the base eigenvectors of θ approximate sine and cosine functions.

random walker on this graph is always more likely to move from its current position to another point that is strongly similar, rather than another point that is less similar. Hence, after a long time, the distribution of the random walk describes the underlying geometry of the data. A lower-dimensional embedding of the data using the diffusion map essentially reorganizes the data so that the Euclidean distance between data points in the lower-dimensional space approximates their connectivity in the kernel space.

The diffusion map is a mapping from each point $i \mapsto A^t[i,:]$, which describes the distribution of a random walker's location at time t, given that the initial location was at point v_i. To compute A, we instead consider the symmetric matrix

$$S = D^{\frac{1}{2}} A D^{-\frac{1}{2}} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}, \tag{2.52}$$

as A is not symmetric. We then take the eigendecomposition

$$S = V \Sigma V^T \tag{2.53}$$

and substitute it back into (2.52) to obtain

$$A = D^{-\frac{1}{2}} S D^{\frac{1}{2}} = D^{-\frac{1}{2}} V \Sigma V^T D^{\frac{1}{2}} = \left(D^{-\frac{1}{2}} V\right) \Sigma \left(D^{\frac{1}{2}} V\right)^T. \tag{2.54}$$

Define $\Phi = D^{-\frac{1}{2}} V = [\varphi_1, \ldots, \varphi_d]$ and $\Psi = D^{\frac{1}{2}} V = [\psi_1, \ldots, \psi_d]$, where $\phi_i, \psi_i \in \mathbb{R}^n$. This gives us the decomposition

$$A = \Phi \Sigma \Psi^T. \tag{2.55}$$

Now, $A\varphi_i = \lambda_i \varphi_i$ and $\psi_i^T A = \lambda \psi_i^T$, i.e., ϕ_i and ψ_i are the right and left eigenvectors of A, so we can write

$$M = \sum_{j=1}^{n} \lambda_j \varphi_j \psi_j^T, \quad M^t = \sum_{j=1}^{n} \lambda_j^t \varphi_j \psi_j^T. \tag{2.56}$$

Then our mapping can be written as

$$i \mapsto M^t[i,:] = \sum_{j=1}^{n} \lambda_j^t \phi_j(i) \psi_j^T \tag{2.57}$$

and so each point can be represented entirely by the basis $\{\psi_j\}_{j=1}^d$ and its coefficients $(\lambda_1^t \phi_1(i) \cdots \lambda_n^t \phi(n))$.

Observe that $A\mathbf{1} = \mathbf{1}$ by virtue of its probabilistic construction, so one of the eigenvalues will always be 1. Moreover, it can be shown that $|\lambda_j| \leq 1$ for all j. Thus λ_1 and ϕ_1 are uninteresting, and we can define a truncated diffusion map

$$i \mapsto (\lambda_1^t \phi_1(i), \cdots, \lambda_{d+1}^t \phi(d+1)). \tag{2.58}$$

The graph Laplacian is given by $L = I - A$, where I is the $n \times n$ identity matrix. The largest eigenvalues and corresponding eigenvectors of L can be computed for further spectral analysis methods such as spectral clustering and dimensionality reduction. It can be easily seen that the eigenvectors of $L = I - A$ are precisely the eigenvectors of A. Therefore, calculating Φ will give us the d most significant Laplace eigenvectors of L.

2.4. **Limit of the graph Laplacian.** Ideally, we would like our Laplace eigenvectors to act as discrete approximations of the actual Laplace eigenfunctions of the data manifold. Fortunately, this is the case for a sufficiently large sample of data. If the data $\{\mathbf{x}_i\}_{i=1}^n$ is uniformly distributed over $\mathcal{M}$, it follows that the graph Laplacian converges to the continuous Laplace operator Δ_M over $\mathcal{M}$. If the data instead has a general distribution $p(x)$, the graph Laplacian converges to the backward Fokker-Planck operator,

$$\mathcal{L}f = \Delta_M f - \nabla U \cdot \nabla f \tag{2.59}$$

where $U(x) = -2 \log p(x)$ is a potential function. Moreover, the eigenvectors of L are discrete approximations of the eigenfunctions of the continuous Laplacian with homogeneous Neumann boundary conditions,

$$\Delta_M f(x) = \lambda f(x), \quad x \in \mathcal{M} \tag{2.60}$$

$$\frac{\partial f}{\partial \nu} = 0, \quad x \in \partial\mathcal{M} \tag{2.61}$$

where ν is the normal vector to the boundary $\partial\mathcal{M}$ at x.

While the results above assume a Euclidean manifold, this discussion can be extended to general manifolds as well. For the purposes of our exploration, we will not consider these generalizations. In the remainder of this paper, we only consider ambient manifolds $\mathcal{M}$ which are the product of two independent sub-manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$. Thus every eigenfunction of $\mathcal{M}$ can be written as $f_j g_k$ with eigenvalue $\lambda_j + \mu_k$, where f_j is a Laplace eigenfunction of λ_j over $\mathcal{M}_1$ and g_k is a Laplace eigenfunction of μ_k over $\mathcal{M}_2$.

3. Method

3.1. **Predicting base eigenvectors.** As we have shown, any Laplace eigenvector φ_k for data sampled from the product manifold $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$ can be written as the element-wise product

$$\varphi^{(i,j)} = \varphi_i^{(1)} \varphi_j^{(2)}, \tag{3.1}$$

where $\varphi_i^{(1)}$ is an eigenvector of $\mathcal{M}_1$ and $\varphi_j^{(2)}$ is an eigenvector of $\mathcal{M}_2$. If $\varphi_k^{(b)}$ is the trivial eigenvector $\mathbf{1}$ for $b = 1, 2$, we call $\varphi^{(i,j)}$ a *base* eigenvector of $\mathcal{M}_{3-b}$, otherwise we say it is a *mixture* eigenvector. We now present a method for determining the base eigenvectors included in the first N Laplace eigenvectors of $\mathcal{M}$.

We begin by finding all triplets of eigenvectors $(\varphi_i, \varphi_j, \varphi_k)$ such that φ_k best matches the element-wise product of φ_i and φ_j. To do so, we iterate through all triplets $(\varphi_i, \varphi_j, \varphi_k)$ where $i < j < k \leq N$, and check how "close" $\varphi_i \varphi_j$ is to φ_k. Here, "closeness" is measured by scaling $\varphi_i \varphi_j$ and φ_k exactly into the range $[-1, 1]$ and computing the L1-norm of the difference between the two scaled vectors. We make sure to check both φ_k and $-\varphi_k$, as the sign of the first two eigenvectors relative to the third eigenvector is not known. The φ_k that is closest to $\varphi_i \varphi_j$ is recorded, as well as the distance. Finally, since we only want to work with reliable triplets, we filter out all of the resulting $\binom{N}{2}$ triplets which have a distance above a certain threshold. The algorithm is summarized in Algorithm 1 below.

Algorithm 1: Searching for reliable triplets

```
Result: Constructs a list of eigenvector triplets (φ_i, φ_j, φ_k)
bestTriplets ← Array();
for i ← 1 to N − 2 do
    for j ← i + 1 to N − 1 do
        product ← φ_i φ_j;
        closestDist ← 0, closest ← None;
        for k ← j + 1 to N do
            candidateDist ← | ± φ_k − product|;
            if candidateDist < closestDist then
                closestDist ← candidateDist, closest ← φ_k;
            end
        end
        if closestDist < d then
            add (φ_i, φ_j, φ_k) to bestTriplets;
        end
    end
end
return bestTriplets;
```

The result of the algorithm is a list of triplets indicating which eigenvectors are likely to be base eigenvectors (the first two eigenvectors in each triplet), and which eigenvectors are likely to be mixture eigenvectors (the third eigenvector in each triplet).

3.2. **Separating eigenvectors.** Let $\mathcal{T}$ be our list of triplets. We will use a voting procedure to determine which manifold each eigenvector is associated with. Since it is not guaranteed that there exists a division or a unique division that is characterized by our list of triplets, we opt for a spectral approach. First, define $T = \{\varphi_1, \ldots, \varphi_t\}$ to be the set of eigenvectors found in our list of triplets. We will construct a complete weighted graph (G, E, V) where each vertex corresponds to an eigenvector in T, and each edge has a weight w_{ij} representing the affinity between φ_i and φ_j. A higher affinity indicates that the two eigenvectors likely belong to the same intrinsic manifold. We will also keep track of the number of times an eigenvector appears as a base eigenvector or a mixture eigenvector in each triplet via a voting system. For every time that an eigenvector appears as a base eigenvector we cast it one vote, and every time that an eigenvector appears as a mixture eigenvector we subtract one vote for it. Let $v(i)$ denote the net number of votes that eigenvector φ_i receieves. We include in our affinity matrix only edges between eigenvectors which have receieved more than K net votes.

To construct the votes and weights for G, we iterate through each triplet $(\varphi_i, \varphi_j, \varphi_k)$ in $\mathcal{T}$ and collect votes for edges and vertices as follows:

- Add $e^{d(\varphi_i,\varphi_j,\varphi_k)}$ to each of w_{ij} and w_{ji}, where $d(\varphi_i, \varphi_j, \varphi_k) = |\varphi_i\varphi_j - \varphi_k|$
- Add 1 to each of $v(i)$ and $v(j)$
- Subtract 1 from $v(k)$

After going through all triplets in $\mathcal{T}$, the edges between eigenvectors which appeared most frequently together as well-paired base eigenvectors will have large distances between them, and eigenvectors which appeared most as base eigenvectors have the highest votes. The final algorithm is summarized in Algorithm 2.

Algorithm 2: Spectral procedure to separate base eigenvectors

Result: Given a list $\mathcal{T}$ of triplets and K, computes two lists of base eigenvectors associated with $\mathcal{M}_1$ and $\mathcal{M}_2$

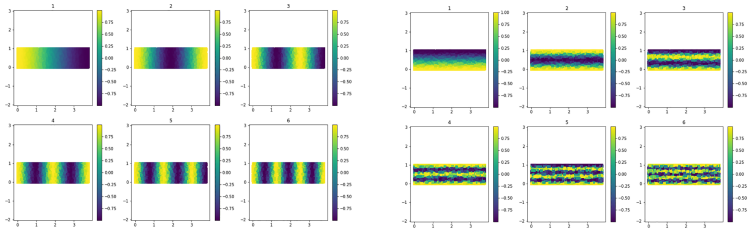
initialize $W \leftarrow \text{Array}(t, t)$;
for $(\varphi_i, \varphi_j, \varphi_k)$ *in* $\mathcal{T}$ **do**
  $W(i,j) \leftarrow W(i,j) + e^{d(\varphi_i,\varphi_j,\varphi_k)}$;
  $W(j,i) \leftarrow W(j,i) + e^{d(\varphi_i,\varphi_j,\varphi_k)}$;
  $v(i) \leftarrow v(i) + 1$;
  $v(j) \leftarrow v(j) + 1$;
  $v(k) \leftarrow v(k) - 1$;
end
Restrict W to the eigenvectors $\{\varphi_i : v(i) \geq K\}$;
`eigenvectors1`, `eigenvectors2` $\leftarrow$ `SpectralClustering`(W, *clusters*$= 2$);
return `eigenvectors1`, `eigenvectors2`;

4. Results

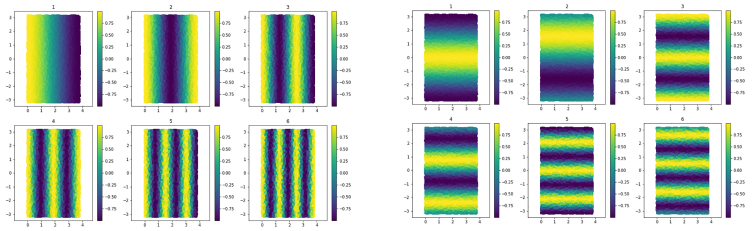
To test the algorithm, we ran it on two two-dimensional toy manifolds, a rectangle and a hollow cylinder. The ground truth Laplace eigenfunctions of both manifolds are known, as derived in Section 2. For each manifold, we collect 10,000 uniform random samples over the ambient manifold, and we examine the 100-dimensional diffusion map created from those samples, which give us the first 100

nontrivial Laplace eigenvectors. The default parameters are set to $d = 0.08, K = 2$ for the rectangle and $d = 0.12, K = 2$ for the hollow cylinder, but we also explore the effects of varying these parameters.

The results are shown in Figures 4 and 5, and the first six ground truth eigenfunctions for each manifold are shown in Figure 3. The algorithm produces two groups of base eigenvectors, each corresponding to an independent manifold. In both of these examples, the algorithm is able to accurately determine the most significant base eigenvectors associated with each independent sub-manifold, as well as some higher frequency eigenvectors. Comparisons between the results and the ground truth eigenvectors show that the algorithm consistently identifies several of the lowest frequency eigenfunctions. Note that some eigenvectors in Figures 4 and 5 are mirror images of those in Figure 3, as the algorithm takes into account both positive and negative scalar multiples of eigenvectors.



(a) Laplace eigenfunctions for the rectangle



(b) Laplace eigenfunctions for the hollow cylinder

FIGURE 3. Ground truth Laplace eigenfunctions over the (a) rectangle and (b) hollow cylinder. We reparametrize as polar coordinates for the cylinder, so the vertical axis represents θ in radians.

The two main parameters of the algorithm are d, which is the distance threshold used to filter out unreliable triplets, and K, which is the voting cutoff for an eigenvector to be considered in the affinity matrix used for manifold separation. Both parameters serve to limit the amount of information that the clustering portion of the algorithm can access, but act against each other. That is, reducing d filters out more eigenvector triplets (Algorithm 1), so that fewer triplets make it to the voting process. Conversely, lowering K increases the number of eigenvectors in the clustering process (Algorithm 2). Thus, in order to obtain comparable results any large adjustment in one parameter in one direction should be coupled with an adjustment of the other parameter in the same direction. Figures 4 and 5 show how

comparable results are as d and K vary simultaneously. Expanding the number of reliable triplets by increasing d and choosing base eigenvectors more selectively by increasing K did not alter the results significantly.

One potential concern is that the results may be extremely sensitive to variations within a single parameter. In particular, stability over varying values of d is desirable, as d may vary between zero and the largest distance of any triplet. To explore this, we examine how the chosen eigenvectors compare across a range over d for different fixed values of K. The results are shown in Figure 6. For $K = 1, 2$, the predictions are stable over the range $d \in [0.1, 0.2]$, but if d becomes too large then the first eigenvector may be misclassified. We suspect that this is because the first eigenvector appears as a base eigenvector in the most number of triplets, and if d is too large then Algorithm 1 does not sufficiently filter out unreliable triplets containing the first eigenvector, which skews the voting process.

5. Conclusion

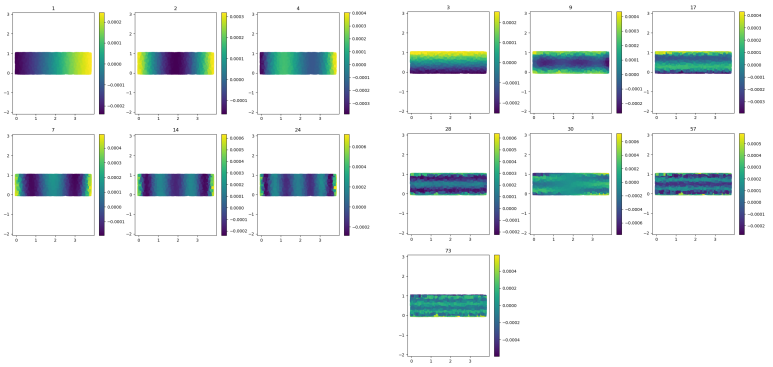
5.1. **Summary.** In this paper, ~~we explore a possible~~ algorithm for independent manifold analysis. One motivating application of such an algorithm is for the heterogeneity reconstruction problem in cryo-EM. Reconstructing three-dimensional models of molecules imaged by cryo-EM requires some knowledge of the manifold of possible conformations which that molecule may have. Isolating that manifold from the ambient manifold of the data can help extract this information.

Our algorithm leverages multiple spectral graph analysis techniques, including diffusion maps and spectral clustering, to analyze the Laplace eigenvectors of the data. For a large number of data points, these eigenvectors approximate the eigenfunctions of the ambient manifold. When tested on toy manifolds that are the product of two independent manifolds, it successfully distinguishes and separates the top most significant eigenvectors of either submanifold. Moreover, the predictions are stable over a reasonable range of the algorithm parameters.

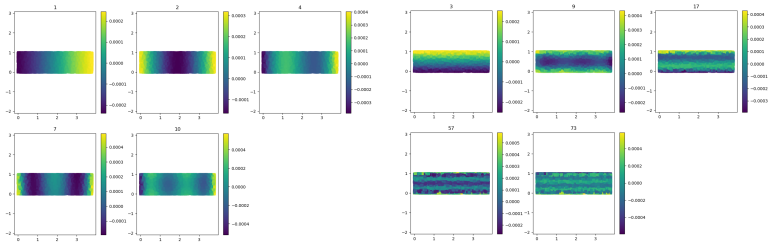
5.2. **Future work.** In future explorations, it is necessary to have a concrete evaluation method of how well the algorithm performs on a given manifold. In order to properly evaluate the algorithm, we must have a metric which measures the algorithm's performance. In this setting, a natural metric would be to simply compare the results of the algorithm against the ground truth and count how many correct eigenvectors the algorithm chooses, while penalizing incorrect choices. These counts should also be weighted, as correctly identifying the lowest frequency (most significant) eigenvectors is an easier task than correctly identifying higher frequency ones. This is because lower frequency eigenvectors are simply considered more frequently as potential base eigenvectors, as the number of triplets in which they appear is higher. Mischaracterizing a more significant mixture eigenvector is easier than mischaracterizing a less significant eigenvector for the same reason. The difficulties of these choices also depends on the dimensionality reduction performed by the diffusion map, as this gives us greater or fewer eigenvectors to examine. A good metric should combine all of these factors.

Acknowledgments.

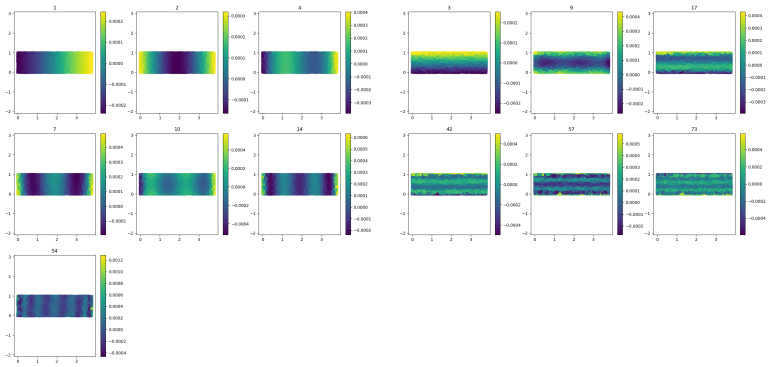
References



(a) $d = 0.05,\ K = 1$

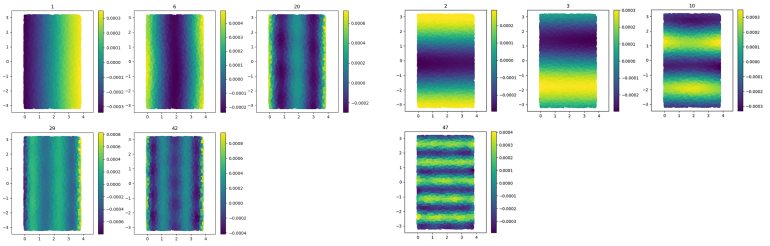


(b) $d = 0.08,\ K = 2$

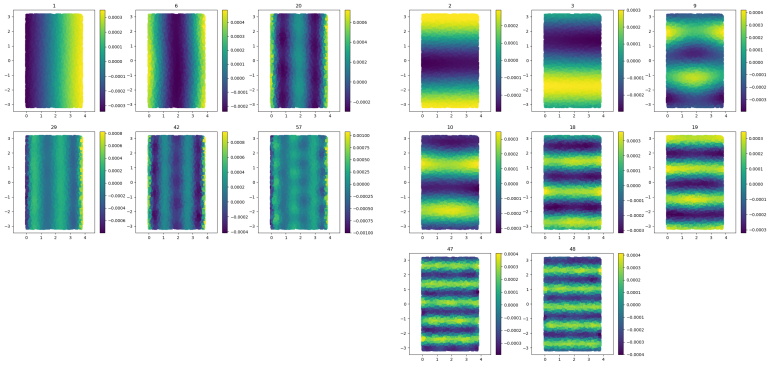


(c) $d = 0.25,\ K = 3$

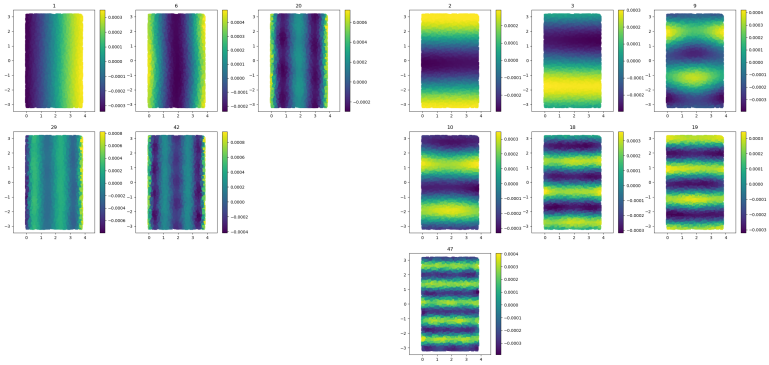
FIGURE 4. The results of the algorithm on 10,000 points sampled from the rectangle $\Omega = [0, 2 + \sqrt{\pi}] \times [0, 1]$. Varying distance and voting thresholds d and K were applied over the same set of samples, resulting in slightly different separations. On the left are the base eigenvectors associated with $l_1 = [0, 2 + \sqrt{\pi}]$, on the right are the base eigenvectors associated with $l_2 = [0, 1]$. As we can see, the algorithm is able to consistently pick out the most significant base eigenvectors for both manifolds, with only a single mixture eigenvector (20) incorrectly chosen in (A).



(a) $d = 0.08$, $K = 1$

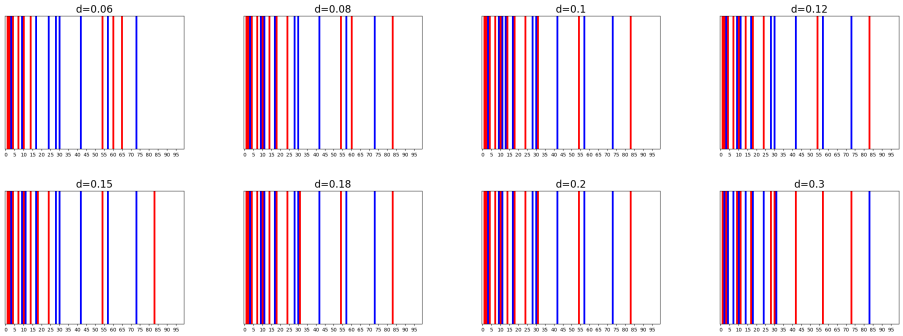


(b) $d = 0.12$, $K = 2$

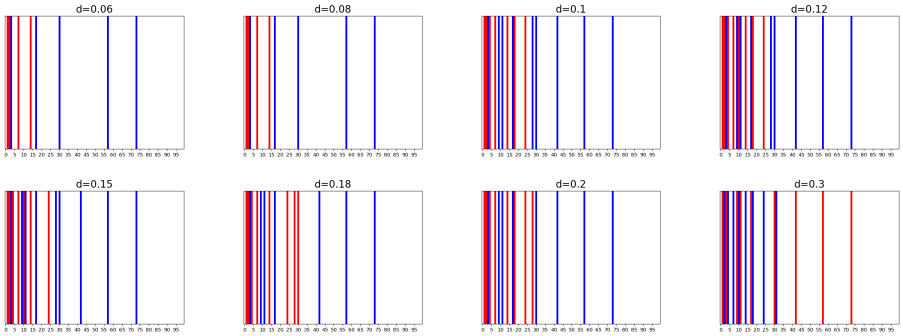


(c) $d = 0.18$, $K = 3$

FIGURE 5. The results of the algorithm on 10,000 points sampled from the hollow cylinder $\Omega = [0, 2+\sqrt{\pi}] \times [-\pi, \pi]$. Varying distance and voting thresholds d and K were applied over the same set of samples, resulting in slightly different separations. On the left are the base eigenvectors associated with $l = [0, 2 + \sqrt{\pi}]$, on the right are the base eigenvectors associated with $\theta = [-\pi, \pi]$. Like for the rectangle, the algorithm is able to consistently pick out the most significant base eigenvectors for both manifolds, with only a single mixture eigenvector (9) incorrectly chosen in (B) and (C).



(a) Rectangle, $K = 1$



(b) Rectangle, $K = 2$

FIGURE 6. The first 100 nontrivial eigenvectors of data sampled from a rectangle, separated and color-coded by manifold using varying parameters for d and K. Red and blue lines indicate base eigenvectors beloning to the two independent manifolds (the two lines whose product is the rectangle), and white indicates a mixture eigenvector. Within the range $[0.1, 0.2]$ over the distance threshold d, the results for both values of K are stable.