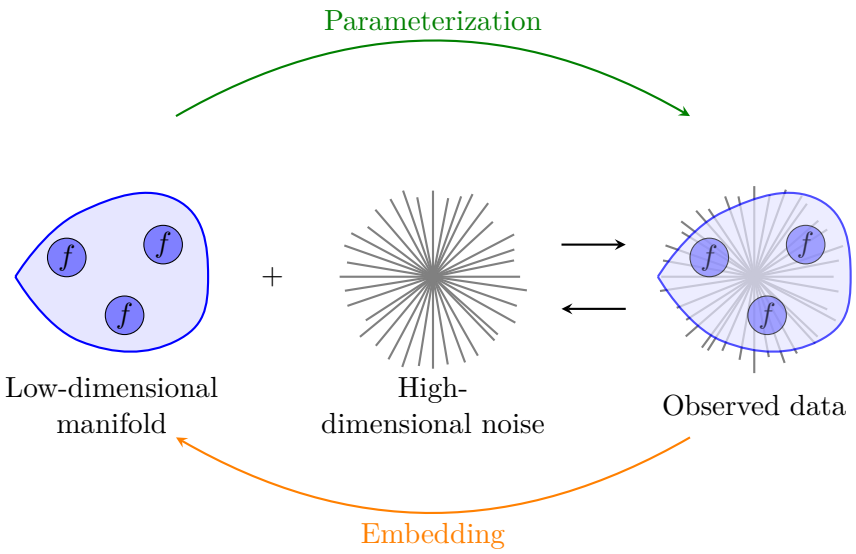

Figure 1: Manifold Learning: Parameterization vs. Embedding

Directional Priors in Manifold Learning

PHP2530

Daniel C. Posmik

April 2, 2025

1. Introduction

Manifold learning is dimensionality reduction technique that has proven useful in settings where data are high-dimensional and non-linear. Often, manifold learning algorithms are used when the topological structure of the data are to be preserved in a statistical learning task[1]. Within the manifold learning framework, data are assumed to live on a lower dimensional manifold and are corrupted by high-dimensional noise. We say that D-dimensional data can be embedded in d-dimensional manifold where $d \leq D$ but generally $d \ll D$.

Within the principal manifolds framework (Meng and Eloyan, 2021) – a replicable and flexible framework for manifold learning – the process of fitting a manifold to our data contains multiple steps. The key idea is that we fit a d-dimensional manifold to our D-dimensional by minimizing the sum of squares between our data and the proposed manifold.

1. For an introduction, see Meilă and Zhang (2023)

An important extension to linear dimensionality reduction, i.e. the principal components algorithm (PCA), is that we allow our proposed manifold to preserve underlying topological structure of our data. In a way, manifold learning reduces the dimensionality of data with an explicit focus on the topology of it. We note that – although certainly intuitive – this topological structure is not only limited to spatial abstraction, but may be extended to arbitrary dimensions of interest. This framework was pioneered as an extension to the PCA algorithm with curves (Hastie and Stuetzle (1989), Tibshirani (1992)) and has since found a myriad of applications in higher-dimensional extensions.

We now propose a method for incorporating prior distributional information into the principal manifolds framework.

2. Background

Consider a setting where we have fit a manifold $\mathcal{M}_d$ to our D-dimensional data by means of minimizing the orthogonal distance between the data and the manifold. We consider this manifold fixed and will not touch on the fitting procedure itself. Given $\mathcal{M}_d$, for each data point, i.e. the row vector $[x_{11} \cdots x_{1D}]^T$, we can now define the point on $\mathcal{M}_d$, say $f\left([x_{11} \cdots x_{1D}]^T\right)$. This point minimizes the distance between x_i and $f(x_i)$. We want to stress again that this procedure does not mean we are fitting the manifold to the data, we are simply retrieving the distance-minimizing projection point. We write

$$\arg\min_{f \in \mathcal{F}} \|x^* - f(x^*)\|_2$$

where we consider each projection function f to be a member of an arbitrary function space $\mathcal{F}$. We define the distance metric as the L^2 distance.

If there exists only one projection point $f(x_i)$ for every x_i, every $f \in \mathcal{F}$ is one-to-one and onto ("bijective") mapping. We find it interesting to highlight that the projection functions in the PCA algorithm are inherently bijective, and for inferential purposes, this is a property that is often taken for granted[2]. In a manifold learning framework, this is no longer the case. Albeit highly interesting, due to the limited scope of this paper, we shall treat this scenario as an edge case, reserving rigorous treatment for the blissful times that follow the author's qualifying exam.

Now, given the data $[\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n]$, we can reparameterize our space into polar coordinates to obtain a vector representation of the collection $f \in \mathcal{F}$. Converting a Cartesian parameterization in space with D dimensions into polar coordinates yields the d-dimensional vector $[r_i^*; \theta_{i,1}, \cdots, \theta_{i,D-1}]$, i.e. one radius r_i^* and a set of $D-1$ angles suffice to characterize each point x_i's location in space.

Recognize that the parameter r^* is not random. This is because it is simply the result from our previous projection distance-minimizing procedure. Usually, polar parameterizations assume that all angles and radii are centered at the origin. Luckily, simple vector addition and subtraction readily generalizes our parameterizations in space. For instance, to obtain the vector from the point x_i and $f(x_i)$, we simply subtract $f(x_i) - x_i$. It is impor-

2. This is because a principal axes is a straight line, i.e. neither convex or concave. Although a point's distance to its projection may be co-minimal across ≤ 2 dimensions, it only has one $f(x_i)$ in one principal axis.

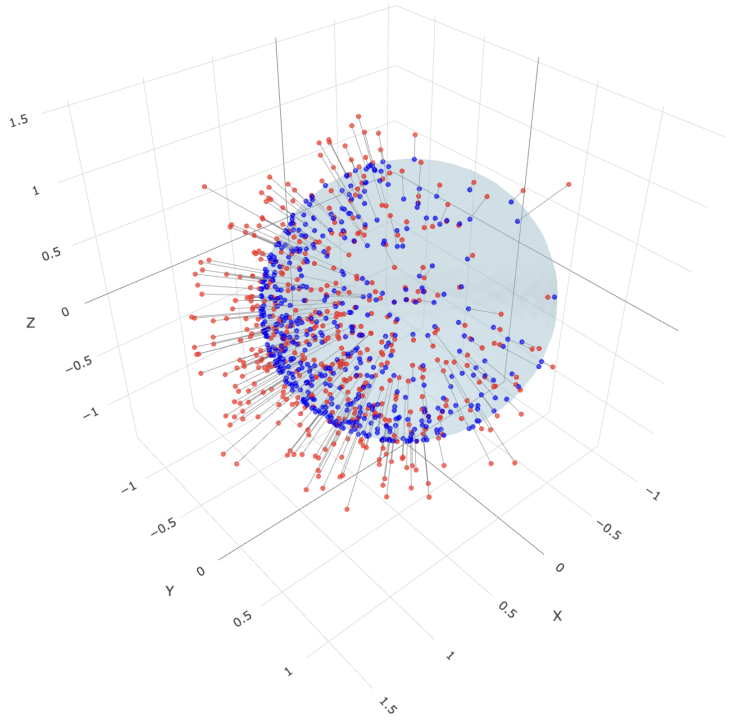

Figure 2: Noisy data projected on $\mathcal{M}_3$; the unit sphere in $\mathbb{R}^3$

tant that the issue of defining the origin is explicitly clarified when dealing with directional information. For simplicity, we will henceforth consider data centered at the origin.

3. Directional Priors

In contrast to the radii, the $\{D-1\}$-dimensional vector of angles, $\boldsymbol{\theta_i} := [\theta_{i,1}, \cdots, \theta_{i,D-1}]^T$ is random. In simple terms, $\{r_i, \boldsymbol{\theta}_i\}$ is what parameterizes the realized sample $\boldsymbol{X_i} = x_i$ in space. If we draw multiple samples, the random sampling variation in θ_i is what captures the randomness. In biomedical applications, such as cancer medicine, we may have reliable prior information (i.e. from previous trials or expert knowledge) on directional trends of malignant growths. Within a Bayesian framework, using our data to update these prior directional information offers a principled, probabilistic solution to complex inference problems in settings where directionality is a key piece of information.

Before we formulate our approach formally, we briefly introduce directional statistics. When we reparameterized our data from Cartesian coordinates into polar coordinates, we did not address the underlying probabilistic transformations. For example, when encoding uncertainty in a spherical setting, it may be naive to parameterize a normal density with support on $\mathbb{R}^1$ since angles are defined in the interval $[0, 2\pi]$. Directional statistics offer well-defined spherical reparameterizations of distributions, such as the normal. Instead of defining a mean vector in $\mathbb{R}^1$, we can define a mean directional vector contained in $[0, 2\pi]$. To

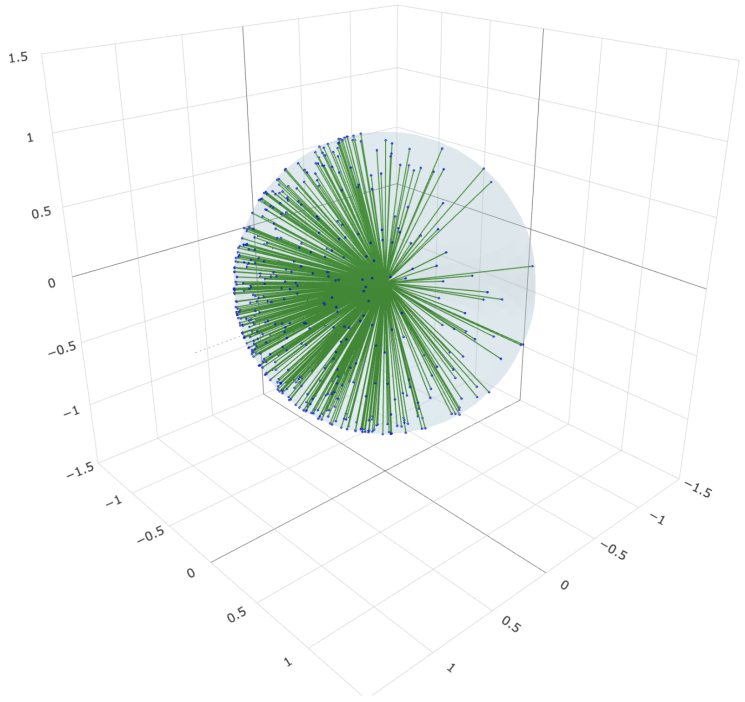

Figure 3: Polar parameterization with $\{r_i, \theta_i, \phi_i\}$ of projected data

account for the infinite support, the wrapped normal is defined on a support of $k \text{mod}(2\pi)$. Henceforth, unless stated otherwise, we will deal with the spherical parameterizations of densities to account for the spherical parameterization of our problem.

3.1 Motivating Example: 3D Sphere

We begin our discussion with an example of a 3D sphere. As discussed, this means that we have some simulated noisy data $\{x_i\}_{i=1}^n$ in 3D space. The noisy data is projected on the manifold of choice, i.e. the unit sphere. This yields the data $\{f(x_i)\}_{i=1}^n$. These data are parameterized in a polar coordinate system with $\{r_i, \theta_i\}$, where the (non-random) radius $\vec{r_i}$ is the vector from the origin to $f(x_i)$. Since we are in 3D space, we have two angles. One angle, $\{\theta\}_{i=1}^n$, which parameterizes the angle from the origin on the XY-plane. The angle $\{\phi\}_{i=1}^n$ is the angle between the XY-plane and the point $f(x_i)$.

3.2 Formulating Directional Priors

Formulating a directional prior is not trivial. For one, it is important to set a reference point for both the directional prior and the angles. It is important that all angles are measured from the same reference point. For simplicity, we choose the origin. Now, choosing a

directional prior in some direction $\mu \in [0, \pi]$ allows[3] us to incorporate information on where the projected points should be concentrated on the manifold. We only posit a prior on the random set $\{\theta_i\}_{i=1}^n$, expressing some prior distributional belief on the concentrations of points in the XY-plane. In this example, we consider the angles $\{\phi_i\}_{i=1}^n$ to be fixed, making our example univariate. An extension to the multivariate case would be straightforward by using the multivariate analogoues of our likelihood and prior distribution functions.

Our posterior distribution on the random vector θ can be expressed as follows.

$$\begin{aligned} f_{\theta|\mu}(\theta|\mu) &\propto \left\{ \prod_{i=1}^{n} f_{\theta_i|\mu}(\theta_i|\mu) \right\} \cdot f_\mu(\mu) \\ &\propto \mathcal{L}(\theta|\mu) \cdot f_\mu(\mu) \end{aligned}$$

where $\boldsymbol{\mu}$ is our mean vector that we posit a prior on, i.e. $\mu \sim f_\mu(\mu)$. To form our likelihood, we rely on the wrapped ("spherical") normal distribution. The density of a wrapped normal is given by

$$f(\theta|\mu, \sigma^2) = \frac{1}{2\pi\sigma^2} \sum_{k=-\infty}^{\infty} \exp\left(-\frac{(\theta - \mu + 2\pi k)^2}{2\sigma^2} \right)$$

with the infinite support reflected in the $mod(k)$ argument. Integrating over 2π is equivalent to integrating over $\mathbb{R}$ in the canonical parameterization. Now, a disadvantage of using the above form is the infinite series. Deriving moments is not straightforward and requires (inverse) Fourier transforms. Thus, we will use the von Mises distribution, an analogue to the spherical normal that is easier to handle. The density of a von Mises distribution is given by

$$f(\theta|\mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi I_0(\kappa)}$$

where θ is the angle, μ is the mean direction (location parameter), $\kappa \geq 0$ is the concentration parameter, and $I_0(\kappa)$ is the modified Bessel function of the first kind of order 0. The von Mises distribution is also known as the circular normal distribution and is the circular analog to the normal distribution. As κ increases, the distribution becomes more concentrated around the mean direction μ. When $\kappa = 0$, the distribution reduces to the uniform distribution on the circle. Although we will not dive further into this, the von Mises distribution scales nicely for p-dimensional hyperspheres in $\mathbb{R}^p$, making the family of von Mises distributions an attractive and flexible candidate for learning tasks in spherical settings.

3.3 Variance Encodes Curvature

In the above parameterization, it is natural to ask how to define the variance in terms of the concentration parameter κ. Choosing the variance term naively, e.g., a constant, would of course be possible, but would negate the benefits that manifold-learning gives us

3. Note that we choose prior means are only meaningful in $[0, \pi]$ This is because for any prior mean greater than π, we would consider the closer angle in the opposite direction.

as a dimensionality reduction technique which preserves local toplogical structure. Thus, we are interested in a principled manner of choosing the variance term that encodes the (dis-)similarity of the topological structure. Intuitively, if we have a directional prior on a region with little topological variation, we may be willing to update our projection more liberally. However, suppose that even a small change in the projection angle moves our data point into an area that is very different from our original projection. In that case, we not be as willing to update our projection.

In manifolds that spherically parameterized, using the inverse absolute Gaussian curvature is a sensible choice to encoding variance. Interestingly, there is an intuitive connection to the score test in likelihood-based inference, which uses the slope at restricted MLE estimates as a measure of closeness/ similarity. Thus, not only does this solution make intuitive topological sense, it has a strong analogue in well-established likelihood-based methods.

The Gaussian curvature of a surface can be defined as the product of the principal curvatures:

$$K = k_1 \cdot k_2$$

where k_1 and k_2 are the principal curvatures at a point on the surface. For a sphere of radius R, we can deduce its Gaussian curvature by considering any point on a sphere of radius R. At this point, we can find two perpendicular directions that correspond to the principal curvatures. Due to the perfect symmetry of a sphere, the curvature is the same in all directions at any given point. For a sphere, both principal curvatures are equal to $\frac{1}{R}$ (the reciprocal of the radius). Therefore, for our example, the Gaussian curvature is:

$$K = k_1 \cdot k_2 = \frac{1}{R} \cdot \frac{1}{R} = \frac{1}{R^2}$$

For a unit sphere (where $R = 1$), the Gaussian curvature[4] is $K = \frac{1}{1^2} = 1$.

3.4 Conjugacy

A key advantage of using the von Mises distribution are conjugacy results. We lean on the work of Mardia and El-Atoum (1976). A von Mises prior distribution with mean direction μ and concentration parameter κ (scaled by c) is conjugate to the likelihood. After observing angles $\theta_1, \ldots, \theta_n$, the posterior distribution is proportional to the following expression (Mardia and El-Atoum, 1976):

$$f(\mu_i, \mu^* | \{\theta_i\}_{i=1}^n) \propto \exp(\kappa \cdot \sum_{i=1}^{n} \cos(\theta_i - \mu_i) + \kappa^* \cdot \sum_{i=1}^{n} \cos(\mu_i - \mu^*))$$

where the angles $\{\theta_i\}_{i=1}^n \sim \mathcal{VM}(\mu_i, \kappa)$ and the prior on μ_i is $\mu_i \sim \mathcal{VM}(\mu^*, \kappa^*)$.[5] In our case, we generated the data with $\mu_i = \mu = \text{circular}(0)$. We can also see that the above form assumes the form of shrinkage estimator[6]. The component involving $\{\theta_i\}_{i=1}^n$ is weighted by variance term κ, in our case the curvature at the projected point. The difference between

4. This constant Gaussian curvature of 1 is an intrinsic property of the unit sphere and is related to the fact that the total curvature integrated over the entire sphere equals 4π (by the Gauss-Bonnet theorem)
5. $\mathcal{VM}(\cdot, \cdot)$ denotes the von Mises distribution
6. For the normal case, this procuedure leads to the James-Stein estimator (Guttorp and Lockhart, 1988).

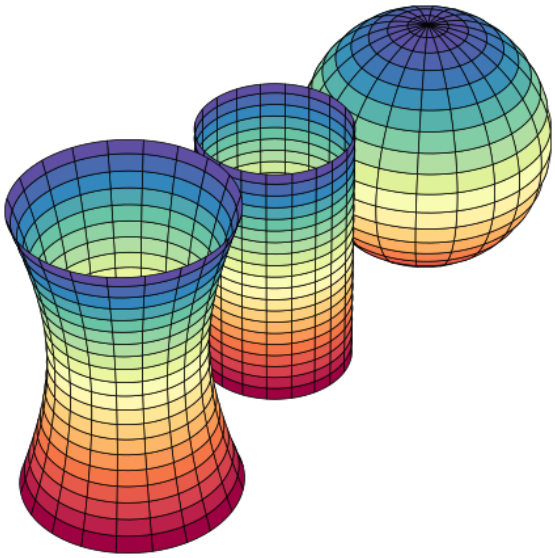

Figure 4: From left to right: a surface of negative Gaussian curvature (hyperboloid), a surface of zero Gaussian curvature (cylinder), and a surface of positive Gaussian curvature (sphere).

the mean parameter of the data and the prior mean is weighted by the prior concentration parameter κ^*. It is interesting to extend this form of a shrinkage estimator to the explicitly topological interpretation we have mentioned above. If the concentration parameter of our data (κ) is low, we attribute more weight to our "prior" information. The above result is the desired probabilistic framework we sought to establish for this directional prior case.

We are now ready to state the conjugacy result for our 3D spherical example. Given the following von-Mises parameterization of a prior on $\boldsymbol{\theta} \sim \mathcal{VM}(\theta_0, \tau_0 = \frac{1}{\kappa_0})$, we have the posterior mean

$$\frac{\tau_0}{\tau_0 + \hat{\tau}} \cdot \theta_0 + \frac{\hat{\tau}}{\tau_0 + \hat{\tau}} \cdot \bar{\theta}$$

and with posterior variance $(\tau_0 + \hat{\tau})^{-1}$. Here, $\bar{\theta} := \frac{1}{n} \sum_{i=1}^{n} \theta_i$ and $\hat{\tau} = \frac{1}{\hat{\kappa}}$ where $\hat{\kappa}$ is the observed Gaussian curvature at a point on the manifold. The shrinkage toward the prior mean is immediately evident.

We can see that the posterior mean is essentially the average of the prior and data means. This is because our manifold is a sphere with equivalent curvature at every point. For manifolds with heterogeneous curvature, we would see less reprojection with respect to the prior mean when curvature is more extreme.

4. An Empirical Bayes Extension in the Service of Causality

Talk about the pre-post treatment idea. What can angles represent? Dont do any code for this, just explain.

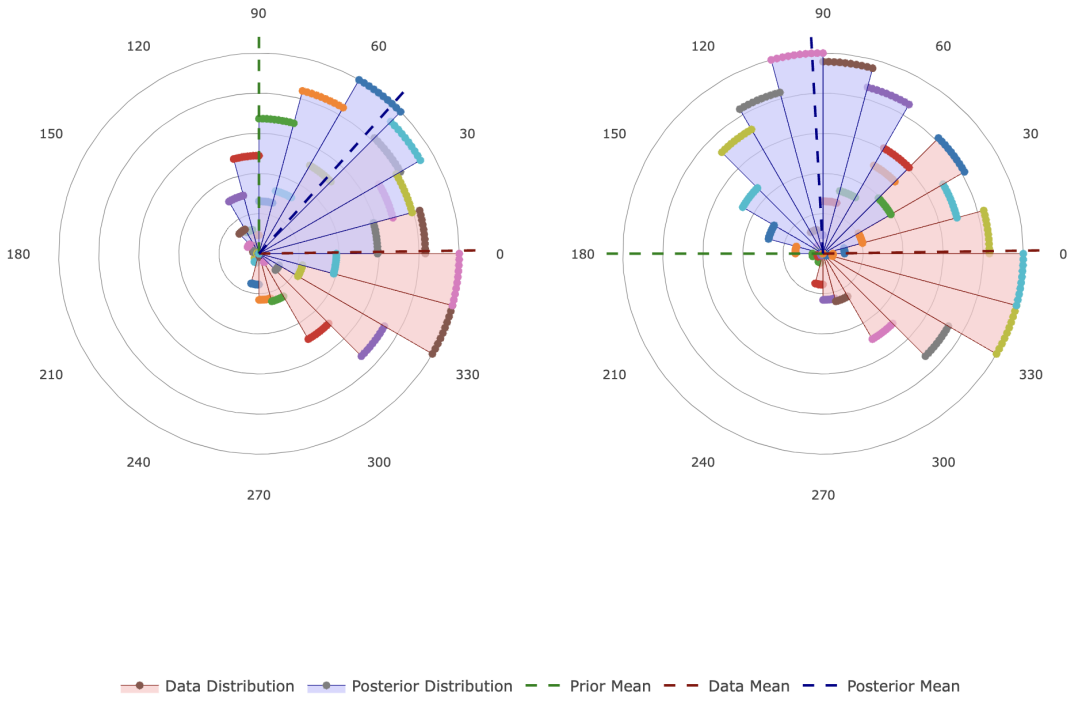

Figure 5: Posterior distribution under prior θ_0: Left $\theta_0 = \pi/2$; Right: $\theta_0 = \pi$

Could a prior on κ potentially encode anticipated topological changes?

5. Limitations and Future Direction

Appendix A.

In this appendix we prove the following theorem from Section 1

References

Peter Guttorp and Richard A. Lockhart. Finding the location of a signal: A bayesian analysis. *Journal of the American Statistical Association*, 83(402):322–330, 1988. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2288846.

Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2289936.

K. V. Mardia and S. A. M. El-Atoum. Bayesian inference for the von mises-fisher distribution. *Biometrika*, 63(1):203–206, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335106.

Marina Meilă and Hanyu Zhang. Manifold learning: what, how, and why, 2023. URL https://arxiv.org/abs/2311.03757.

Kun Meng and Ani Eloyan. Principal manifold estimation via model complexity selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):369–394, March 2021. ISSN 1467-9868. doi: 10.1111/rssb.12416. URL http://dx.doi.org/10.1111/rssb.12416.

Robert Tibshirani. Principal curves revisited. *Statistics and Computing*, 2(4):183–190, 1992. doi: 10.1007/BF01889678. URL https://doi.org/10.1007/BF01889678.