Directional Priors in Manifold Learning: A Causal Perspective PHP2530

Daniel C. Posmik

April 10, 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 $d \ll D$.

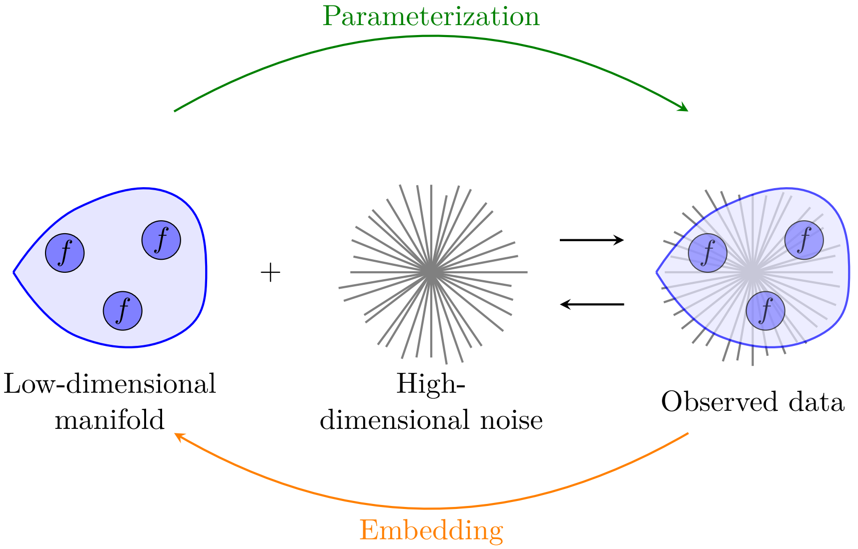


Figure 1: Manifold Learning: Parameterization vs. Embedding

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 of the fitting step is that we fit a d-dimensional manifold to our D-dimensional by minimizing the sum of squares between our data and

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

the proposed manifold. 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 settings.

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 Sobolev 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 a 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.

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.

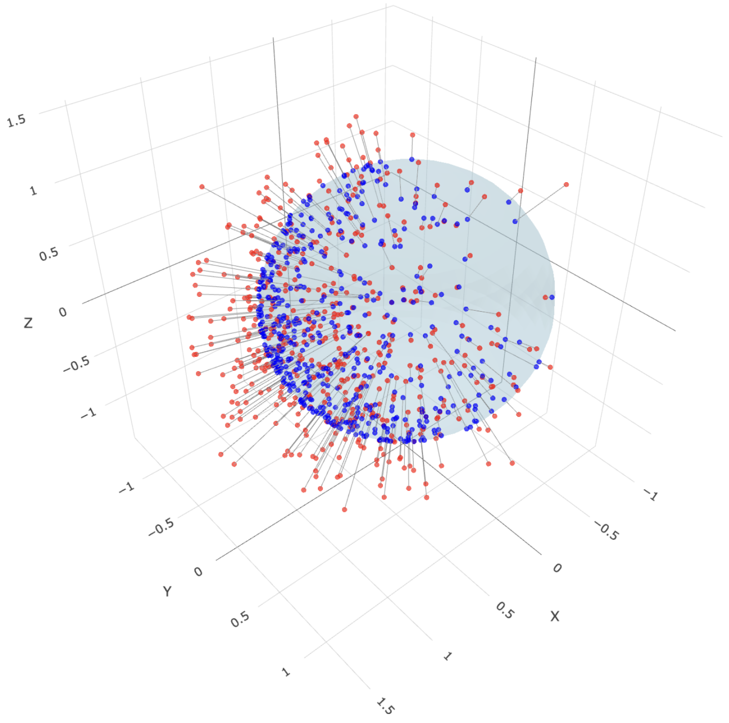


Figure 2: Noisy data projected on $\mathcal{M}_3$; the unit sphere in $\mathbb{R}^3$

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 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 $f(x_i)$.

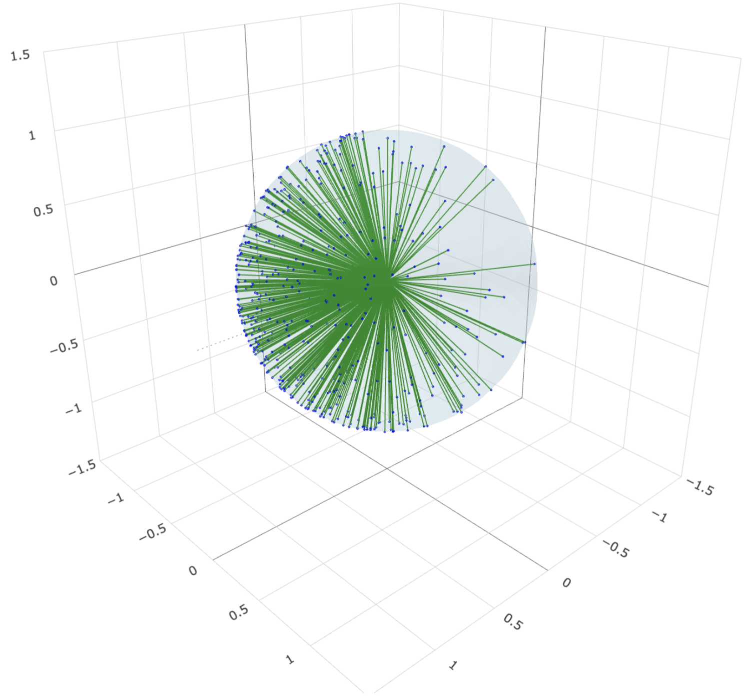


Figure 3: Polar parameterization with $\{r_i, \theta_i, \phi_i\}$ of projected data

3.2 Formulating Directional Priors

Formulating a directional prior requires careful consideration of reference points. All angles must be measured from the same reference point—for simplicity, we choose the origin. A directional prior with mean direction $\mu \in [0, \pi]$ allows us to incorporate information about where projected points should concentrate on the manifold. We only posit a prior on the random set $\{\theta_i\}_{i=1}^n$, expressing prior distributional beliefs on point concentrations in the XY-plane. For this example, we consider angles $\{\phi_i\}_{i=1}^n$ fixed, making our example univariate. Extension to the multivariate case is straightforward using multivariate analogs of our likelihood and prior distributions.

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

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

where μ is our mean vector with prior $\mu \sim f_\mu(\mu)$. For our likelihood, we initially consider the wrapped normal distribution with density:

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

However, this infinite series presents computational challenges. Instead, we use the von Mises distribution, which provides a more tractable circular analog to the normal distribution:

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

where θ is the angle, μ is the mean direction, $\kappa \geq 0$ is the concentration parameter (analogous to precision, or inverse variance), and $I_0(\kappa)$ is the modified Bessel function of first kind of order 0. As κ increases, the distribution concentrates more tightly around μ. When $\kappa = 0$, the distribution becomes uniform on the circle. The von Mises distribution scales naturally for hyperspheres in $\mathbb{R}^p$, making it ideal for spherical settings in manifold learning applications.

3.3 Variance Encodes Curvature

For our concentration parameter κ, naive selection (e.g., a constant) would be possible but would negate the benefits of manifold learning as a topology-preserving dimensionality reduction technique. Instead, we propose a principled approach that encodes local topological structure into the variance parameter. Intuitively, in regions with little topological variation, we might update our projection more liberally. Conversely, if a small change in projection angle moves our data point into a topologically distinct region, we would be more conservative with updates.

In spherically parameterized manifolds, the inverse absolute Gaussian curvature provides a natural variance encoding. This approach has an elegant connection to the score test in likelihood-based inference, which uses the slope at restricted MLE estimates as a similarity measure. Thus, our solution has both topological intuition and established statistical parallels in traditional inference 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, due to perfect symmetry, both principal curvatures equal $\frac{1}{R}$, giving a Gaussian curvature of $K = \frac{1}{R^2}$. For a unit sphere, the Gaussian curvature is $K = 1$, an intrinsic property related to the total curvature integrated over the entire sphere equaling 4π by the Gauss-Bonnet theorem.

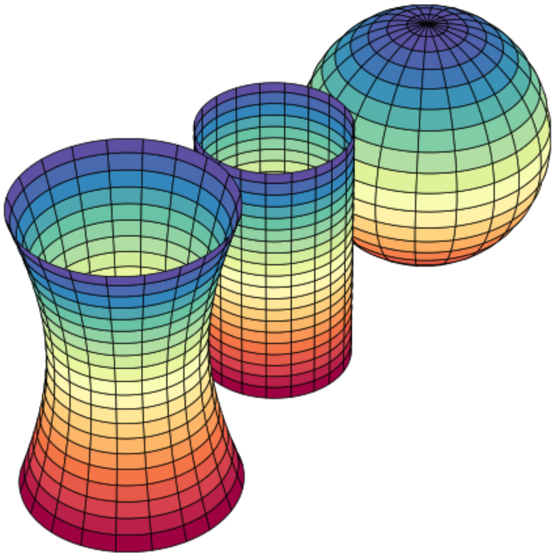


Figure 4: From left to right: a surface of negative Gaussian curvature (hyperboloid), a surface of zero Gaussian curvature (cylinder), and a surface of positive Gaussian curvature (sphere).

This curvature-based approach provides a meaningful way to weight observations based on their location in the manifold's topological structure. Areas with high curvature (significant bending of the manifold) will contribute less to updating our beliefs than areas with lower curvature (flatter regions), which aligns with our geometric intuition about confidence in projections.

3.4 Conjugacy Results

A key advantage of the von Mises distribution is conjugacy, which allows for straightforward Bayesian updates of our directional priors. Following the work of Mardia and El-Atoum (1976), a von Mises prior with mean direction μ and concentration parameter κ is conjugate to the likelihood. After observing angles $\theta_1, \ldots, \theta_n$, the posterior distribution takes the form:

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

where angles $\{\theta_i\}_{i=1}^n \sim \mathcal{VM}(\mu_i, \kappa)$ and the prior is $\mu_i \sim \mathcal{VM}(\mu^*, \kappa^*)$. In our case, the data were generated with $\mu_i = \mu = \text{circular}(0)$.

The posterior form reveals a shrinkage estimator structure. The component involving observed angles $\{\theta_i\}_{i=1}^n$ is weighted by κ, which in our framework represents the curvature at the projected point. The difference between the data mean and prior mean is weighted by the prior concentration parameter κ^*. This structure extends shrinkage estimation to our topological interpretation: when the data concentration parameter κ is low, more weight is attributed to our prior information.

For our 3D spherical example, given a prior $\boldsymbol{\theta} \sim \mathcal{VM}(\theta_0, \tau_0 = \frac{1}{\kappa_0})$, the posterior mean is:

$$\underbrace{\frac{\tau_0}{\tau_0 + \hat{\tau}} \cdot \theta_0}_{\text{Weighted Prior Vector}} \quad + \quad \underbrace{\frac{\hat{\tau}}{\tau_0 + \hat{\tau}} \cdot \bar{\theta}}_{\text{Weighted Data Vector}}$$

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

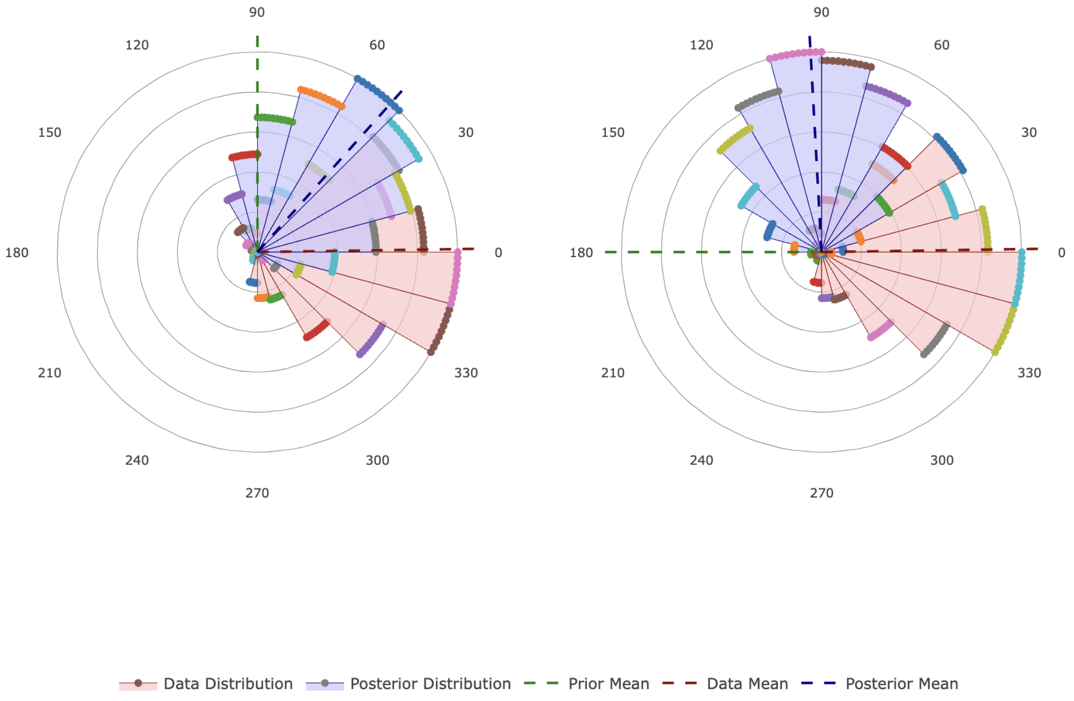


Figure 5: Posterior distribution under prior θ_0: Left $\theta_0 = \pi/2$; Right: $\theta_0 = \pi$

Figure 5 demonstrates that the posterior mean effectively balances the prior and data means. This occurs because our manifold is a sphere with equivalent curvature at every point. For manifolds with heterogeneous curvature, we would observe less reprojection with respect to the prior mean in regions where curvature is more extreme, thus preserving the topological characteristics of the data.

The conjugacy relationship for von Mises distributions has an elegant geometric interpretation. For a von Mises likelihood $f(x|\mu) \propto \exp(\kappa \cos(x - \mu))$ and a von Mises prior on the mean direction $\pi(\mu) \propto \exp(\kappa_0 \cos(\mu - \mu_0))$, the posterior distribution is also von Mises: $\pi(\mu|x) \propto \exp(\kappa_n \cos(\mu - \mu_n))$. The posterior parameters can be expressed in terms of resultant vectors:

$$\kappa_n \cos(\mu_n) = \kappa_0 \cos(\mu_0) + \kappa \cos(x)$$

$$\kappa_n \sin(\mu_n) = \kappa_0 \sin(\mu_0) + \kappa \sin(x)$$

These equations can be rewritten in vector form as:

$$\kappa_n = \sqrt{(\kappa_0 \cos(\mu_0) + \kappa \cos(x))^2 + (\kappa_0 \sin(\mu_0) + \kappa \sin(x))^2}$$

and

$$\mu_n = \text{atan2}(\kappa_0 \sin(\mu_0) + \kappa \sin(x), \kappa_0 \cos(\mu_0) + \kappa \cos(x))$$

This representation yields a powerful geometric interpretation: the posterior mean direction and concentration are determined by the vector sum of the prior and likelihood information vectors in $\mathbb{R}^2$. Each vector has direction given by unit vectors pointing at angles μ_0 and x respectively, with magnitude determined by concentration parameters κ_0 and κ. The resultant vector's length determines the posterior concentration parameter κ_n, while its direction determines the posterior mean direction μ_n.

This insight demonstrates that when prior and likelihood point in similar directions, the posterior becomes more concentrated (higher κ_n), reflecting increased certainty. Conversely, conflicting directions lead to a less concentrated posterior (lower κ_n), reflecting uncertainty.

The posterior mean naturally represents a weighted compromise between prior and data, with weights proportional to their respective concentration parameters. For multiple observations, the vectors from each observation simply add to the prior vector, providing an efficient sequential updating mechanism. This vector-based interpretation elegantly connects Bayesian inference on the circle with vector addition in the plane, making it particularly suitable for our manifold learning framework where directional data naturally arises from projections onto curved surfaces.

4. A Causal Extension – Toward a Manifold RD Estimator

Make it crystal clear that the posterior variance is the Bayesian thing here!!

An interesting application of directional priors is a causal setting where we seek to estimate treatment effects after some treatment has been administered. A popular class of estimators for comparing changes in outcomes of interest close to a treatment cut-off are regression discontinuity (RD) estimators. RD estimators exploit the the balance between units close to the treatment cut-off, assuming that treatment does not depend on any unit characteristics. Recovering causal effects in RD estimators hinges on the assumption that in terms of unit characteristics, treatment assignment is quasi-random very close to the cut-off. Popular examples in literature include measuring the effect of advanced coursework on students above some cutoff, comparing their outcomes to students just below said cutoff.

In the principal manifold learning setup, directional priors offer a promising Bayesian approach to causal inference. For simplicity, assume that treatment is administered at time point t^*. Then, all units in $(-\infty, t^*)$ are pre-treatment units, and all units in $[t^*, \infty)$ are post-treatment units. In a canonical RD setting, under a set of identification assumptions, we can recover an average treatment by considering values of Y_i immediately to the right (say $Y_{it}|_{t^*+\delta}$) and left ($Y_{it}|_{t^*-\delta}$) of the cutoff. This yields the causal estimate

$$\hat{\tau} = \bar{Y}|_{t=t^*+\delta} - \bar{Y}|_{t=t^*-\delta}$$

The above estimator is commonly measured in the Local Linear Regression (LLR) framework, i.e. measure the ATE as the difference in post- and pre-treatment means, given that they are sufficiently close to the treatment cutoff t^*. Branson et al. (2019) considers a Bayesian RD setting where a Gaussian process prior is placed on the pre- and post-treatment mean response functions. We believe this work, although not directly related to the use of topological information in causal inference, encapsulates the desire for more flexible, data-driven, and principled causal inference through Bayesian paradigms.

4.1 Intuition

In a principal manifolds framework, we can also exploit this notion of quasi-randomness to recover causal effects. Specifically, we propose the estimation of causal treatment effects using the posterior density of projection angles with empirical prior variances encoding the local topological structure of pre-treatment manifolds. For this procedure, suppose we use the set of pre-treatment units $\{Y_{it}\}_{i=1}^{n}|_{t<t^*}$ to fit the d-dimensional manifold $\mathcal{M}_d^{t<t^*}$. As we have shown in Section 3, a polar reparameterization allows us to recover the projection angles $\{\theta_i\}$ as well as their mean which we shall denote as $\bar{\theta}^{t<t^*}$. This mean pre-treatment projection angle $\bar{\theta}^{t<t^*}$ has a curvature which we will denote as $\kappa^{t<t^*}$. Together, the mean pre-treatment projection angle and curvature sufficiently parameterize the first two moments of the pre-treatment outcome of interest.

After treatment happens at $t = t^*$, we can fit the post-treatment manifold $\mathcal{M}_d^{t\geq t^*}$. In the spirit of LLR-based causal estimation, we want to emphasize that $\mathcal{M}_d^{t\geq t^*}$ is *not* fit iteratively for each post-treatment unit, i.e. we do not obtain the post-treatment unit by combining the post-treatment units with pre-treatment units. Instead, our pre- and post-treatment manifolds are topological analogues to $\bar{Y}|_{t=t^*-\delta}$ and $\bar{Y}|_{t=t^*+\delta}$, respectively. That being said, it may be interesting to let the choice of manifold (e.g., the shape, highest order, etc.) from the pre-treatment case inform the post-treatment case, effectively borrowing prior topological information on a more global level. [3] Now, suppose that we observe the outcomes of m additional units in the immediate post-treatment period, then the data $\{Y_{it}\}_{i=n+1}^{m}|_{t\geq t^*}$ allows us to fit the manifold $\mathcal{M}_d^{t\geq t^*}$. The mean projection angle of $\mathcal{M}_d^{t\geq t^*}$, i.e. $\bar{\theta}^{t\geq t^*}$ will update from $\bar{\theta}^{t<t^*}$, and yield a new posterior estimate for the projection angle.

[3] This is a really interesting direction in the Bayesian paradigm and touches on the notion of persistent homology. Although this discussion goes beyond the scope of this paper, it highlights how nuanced various levels/ sources of topological information can be.

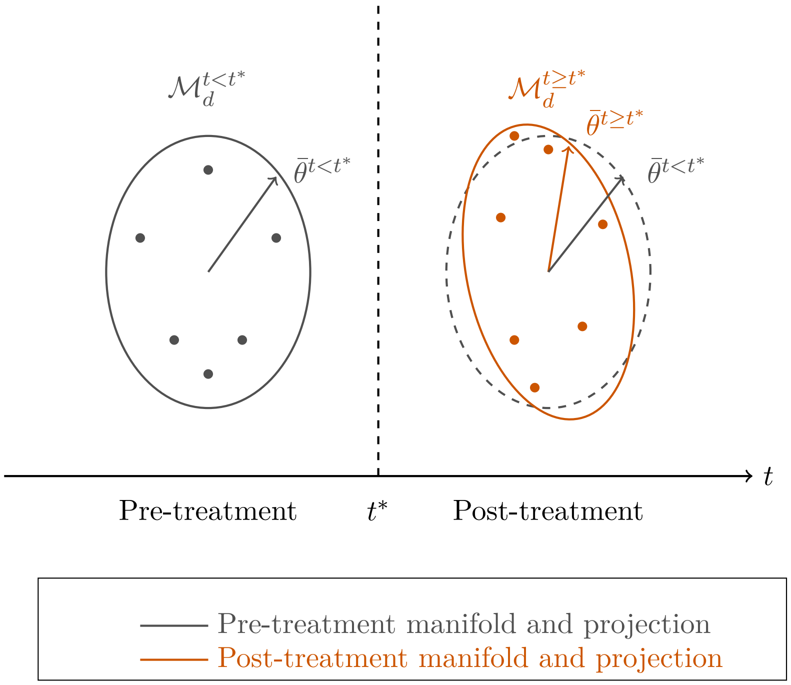

$\mathcal{M}_d^{t<t^*}$
$\bar{\theta}^{t<t^*}$
$\mathcal{M}_d^{t\geq t^*}$
$\bar{\theta}^{t\geq t^*}$
$\bar{\theta}^{t<t^*}$
t
Pre-treatment
t^*
Post-treatment
Pre-treatment manifold and projection
Post-treatment manifold and projection

Figure 6: The Manifold RD Estimator

4.2 The Estimator

Combining our insights from RD estimators with the conjugacy results of Section 3.4, we propose a Bayesian manifold RD estimator in its simplest form. After having the pre- and post-treatment manifold, consider subtracting the pre-treatment distribution of angles from the post-treatment distribution. This yields a RD estimator of the form:

$$f_{\tau|\theta} = \mathcal{VM}\left(\bar{\theta}^{t\geq t^*} - \bar{\theta}^{t<t^*}, \sigma^2\right)$$

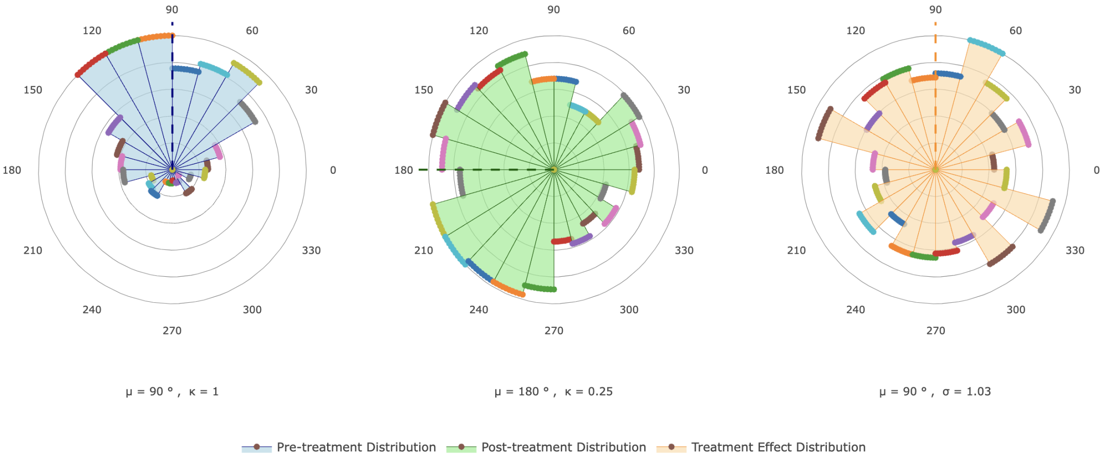
where

$$\sigma^2 = \sqrt{\left[\kappa^{t<t^*}\cos(\bar{\theta}^{t<t^*}) + \kappa^{t\geq t^*}\cos(\bar{\theta}^{t\geq t^*})\right]^2 + \left[\kappa^{t<t^*}\sin(\bar{\theta}^{t<t^*}) + \kappa^{t\geq t^*}\sin(\bar{\theta}^{t\geq t^*})\right]^2}$$

by the conjugacy results of Section 3.4.

Here, we estimate the Average Treatment Effect (ATE) as the difference between the mean projection angles of the post- and pre-treatment manifold, respectively. As illustrated in Figure 6, this difference captures the directional shift in data geometry following treatment intervention. The posterior variance, expressed as the Euclidean norm of the difference between weighted directional vectors, quantifies our uncertainty about this effect. Intuitively, a larger norm indicates greater topological dissimilarity between pre- and post-treatment manifolds, suggesting a more substantial causal impact.

To illustrate this, let us recycle the toy example from Section 3.4. Suppose we have a post-treatment mean projection angle, $\bar{\theta}^{t\geq t^*} = \pi$ and a pre-treatment mean projection angle, $\bar{\theta}^{t<t^*} = \pi/2$. We consider both of these manifolds to be circles, but the post-treatment manifold has radius 2. Thus, the manifold we have fit grows from a unit circle (radius = 1) to a post-treatment circle with radius 2. This means that while the pre-treatment Gaussian curvature stays 1, i.e. $\kappa^{t<t^*} = 1$, our post-treatment curvature is now $\kappa^{t\geq t^*} = \frac{1}{R^2} = \frac{1}{4}$. We simulate 500 draws and observe the following results:



μ = 90 ° , κ = 1

μ = 180 ° , κ = 0.25

μ = 90 ° , σ = 1.03

Figure 7: Simulation of Treatment Effect $\hat{\tau}$

We can see that the resulting treatment effect estimate $\hat{\tau}$ is indeed very close to $\frac{\pi}{2}$. The variance term is slightly above 1, i.e. greater than the individual parts.

When compared to well-established LLR methods, the mean projection angles are simply a spherical reparameterization of the outcome variable of interest. This formulation naturally preserves identification assumptions required for causal interpretation while accommodating the topological abstractions within the principal manifold framework. The magnitude of the posterior variance directly reflects how drastically treatment has altered the local geometry of our outcome space, e.g., if we have a large difference, we want to encode that topological uncertainty into our ATE.

An important point of emphasis is what this procedure is not. Namely, we are not simply updating our fitted manifold with the new data, and letting the posterior angle update iteratively with each new post-treatment observation. Rather, we are using the pre-treatment manifold to borrow topological information for our post-treatment mean vector. In that sense, after having fit $\mathcal{M}_d^{t\geq t^*}$ to the post-treatment data only, we assign the prior mean (i.e. $\bar{\theta}^{t<t^*}$) a weight of zero, only using the post-treatment observations.

We round out this section with an appealing information-theoretic point of view. This manifold RD estimate is fundamentally analogous to the existing methods. The main difference is that this procedure accounts for the topological information of our embedding space. Existing methods are quick to discard this information and assume it away. Principled manifold learning approaches combined with the flexible probabilistic advantages of

Bayesian methods can therefore yield estimates that respect both prior and topological information, resulting in more flexible and data-driven inference.

5. Limitations and Future Direction

Future work could address the use of priors for the construction of the mean projection vectors in the pre- and post-treatment periods, respectively.

Appendix A.

All code and documentation can be found here. [Insert Link]

References

Zach Branson, Maxime Rischard, Luke Bornn, and Luke W. Miratrix. A nonparametric bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, 202:14–30, September 2019. ISSN 0378-3758. doi: 10.1016/j.jspi.2019.01.003. URL `http://dx.doi.org/10.1016/j.jspi.2019.01.003`.

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`.