Beyond Flatland: A Geometric Take on Matching Methods for Treatment Effect Estimation

Melanie F. Pradier MELANIEF@MICROSOFT.COM
Microsoft Research, Cambridge UK

Javier González GONZALEZ.JAVIER@MICROSOFT.COM
Microsoft Research, Cambridge UK

Editors: Biwei Huang and Mathias Drton

Abstract

Matching is a popular approach in causal inference to estimate treatment effects by pairing treated and control units that are most similar in terms of their covariate information. However, classic matching methods completely ignore the geometry of the data manifold, which is crucial to define a meaningful distance for matching, and struggle when covariates are noisy and high-dimensional. In this work, we propose *GeoMatching*, a matching method to estimate treatment effects that takes into account the intrinsic data geometry induced by existing causal mechanisms among the confounding variables. First, we learn a low-dimensional, latent Riemannian manifold that accounts for uncertainty and geometry of the original input data. Second, we estimate treatment effects via matching in the latent space based on the learned latent Riemannian metric. We provide theoretical insights and empirical results in synthetic and real-world scenarios, demonstrating that GeoMatching yields more effective treatment effect estimators, even as we increase input dimensionality, in the presence of outliers, or in semi-supervised scenarios.

Keywords: Causal inference, matching, Riemannian geometry, manifold.

1. Introduction

Causal inference based on observational data plays a pivotal role in unveiling cause-effect relationships, thereby guiding evidence-based practices across numerous sectors including medicine (Rosenbaum, 2012), public policy (Ben-Michael et al., 2023), epidemiology (Westreich et al., 2017), and econometrics (Sekhon and Grieve, 2012). The goal is to estimate the causal effect of a *treatment* T on a given *outcome* Y in the presence of (pre-treatment) input covariates or *confounders* X, as depicted in Figure 1A.

Confounders are responsible for the so-called confounding bias (Rubin, 2005), a fundamental issue in observational studies where there exists an imbalance/discrepancy in the distributions of treated and control units, due to the treatment depending on them instead of being randomly assigned to the observed population. Matching is a well-established approach in causal inference to alleviate confounding bias (Stuart, 2010), where the goal is to pair suitable control units to each treatment unit (and viceversa) based on the similarity of their covariate information.

Traditional matching methods suffer from two main limitations: first, they operate in the space of raw input covariates $\mathcal{X}$ which are noisy and high-dimensional, so it becomes harder to build exact matches (Luo and Zhu, 2017). Indeed, distances become increasingly meaningless for higher dimensions (Aggarwal et al., 2001; Abadie and Imbens, 2006). Second, observations are assumed to live in Euclidean spaces, overlooking the data manifold's geometry: this may lead to distorted

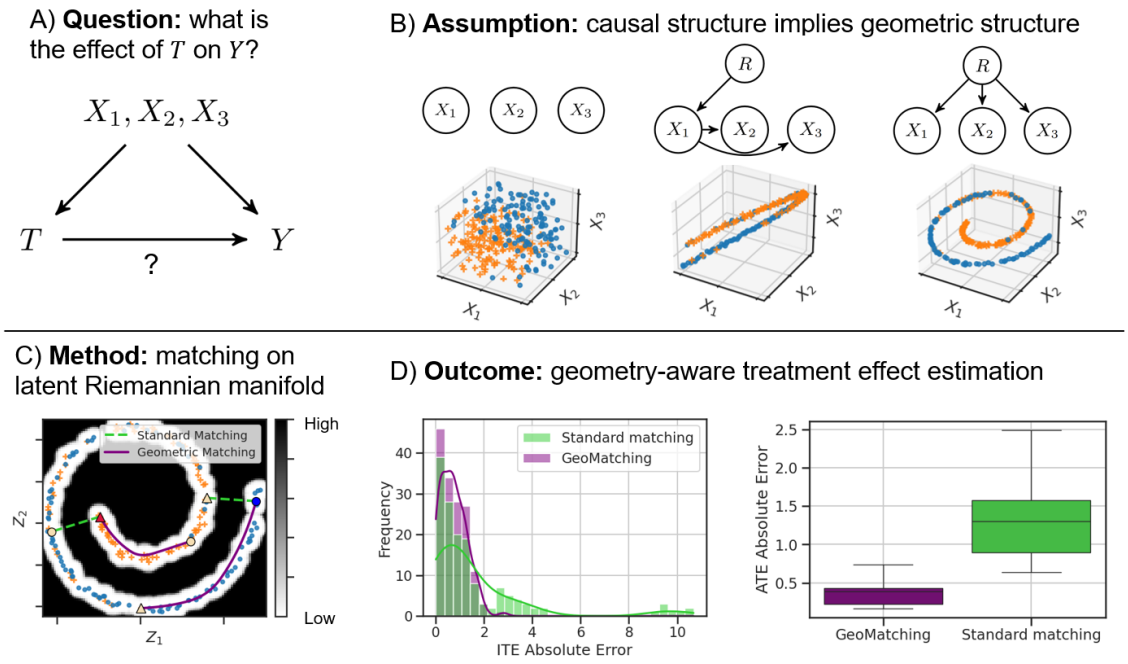


Figure 1: **Key idea of the paper**. We propose *GeoMatching*, a geometry-aware matching framework that pairs cases and controls along the manifold of confounders X. A) Assumed causal graph in this work; B) Different sub-causal graphs for covariates X often induce different geometric data structures; latent variable R represents the manifold structure; C) GeoMatching learns a latent Riemannian metric $\mathbf{G}$ (represented as colorbar) which captures the geometric structure and uncertainty of the data, enabling matched pairs to be faithful to the manifold structure. D) Geometry-aware matched samples translate into more accurate estimates of individual (ITE) and average treatment effects (ATE).

distances Tosi et al. (2014), which result in undesired matched units and imprecise estimates of treatment effects (Stuart, 2010).

In this work, we assume the confounders lie on a manifold of much smaller dimensionality, this is the so called *manifold hypothesis*. Such manifold structure arises from underlying causal mechanisms that induce statistical correlations among covariates in a constrained manner (Dominguez-Olmedo et al., 2023), as illustrated in Figure 1B. For example, a sport trainer might want to estimate the effect of a particular training strategy (treatment) on athletes performance (outcome). To estimate such effect using observational data, the trainer might need to adjust for a wide variety of covariates, including height, weight and physiological measurements (confounders). Here, two athletes should be more or less similar according to natural differences in their physiology's: ideally, distances should respect the manifold structure – here, anatomical constraints – such that confounding bias in treatment effect estimation can be reduced.

We propose *GeoMatching*, a framework to estimate treatment effects accounting for the underlying geometry and uncertainty of the data. Instead of working in the space of raw input characteristics, we learn a low-dim latent representation together with a latent *Riemannian metric* (a generalization of the Euclidean metric) such that raw distances along the original manifold embedded in the high-dim input space are preserved in the low-dim latent space (see Figure 1C). We can then leverage those *geometry-aware* distances for matching to estimate treatment effects more accurately, as shown in

Figure 1D.[1] The latent representation is a *Riemannian manifold*, where distance between two data points is given by minimum path length along the data manifold. To our knowledge, this work is the first one looking at geometry-aware methods (in a differential geometric sense) for causal inference.

Contributions The contributions of this work are the following: first, we introduce *GeoMatching*, a geometry-aware matching method that captures the intrinsic data structure of the confounders leveraging Riemannian geometry, accounting for data uncertainty with particular robustness to outliers. We describe assumptions under which we expect GeoMatching to improve over standard matching methods. Second, we empirically show that GeoMatching yields better results compared to other baselines on diverse synthetic and real-world datasets. We show its effectiveness as we increase the input space dimensionality, in the presence of outliers, or in semi-supervised scenarios.

2. Standard Matching and Problem Description

Notation. Let (X, T, Y) denote input (pre-treatment) covariates $X \in \mathcal{X}$, treatment variable $T \in \{0, 1\}$ and outcome variable $Y \in \mathcal{Y}$. We operate in the standard potential outcome framework introduced by Rubin (2005). That is, we assume that each observation has two potential outcomes $Y(0)$ and $Y(1)$ of which only one is observed, $Y = (1-T)Y(0) + TY(1)$, which is the outcome corresponding to the administered treatment T. Concretely, imagine that a single individual with pre-treatment covariates $X_\star \in \mathbb{R}^d$ is exposed to a treatment $T_\star = 1$, and we observe their outcome $Y_\star(1) \in \mathbb{R}$; this is called the *treated unit*. We also assume access to covariates X_j and outcomes $Y_j(0) \in \mathbb{R}$ of m individuals not exposed to treatment ($T_j = 0$) – the pool of available control units. Let $\mu_t(x) = \mathbb{E}[Y(t) \,|\, X = x]$ denote the expected outcome under treatment $T = t$. We make the following standard assumptions from causal inference (Pearl, 2009):

i) *ignorability*: $Y_i(1), Y_i(0) \perp T_i \,|\, X_i$ – there are no hidden confounders.
ii) *consistency*: $\forall t, T_i = t \to Y_i(t) = Y_i$ – $Y_i(t)$ is the observed outcome when $T_i = t$.
iii) *overlap*: $\forall x, 0 < P(T_i = 1 \,|\, X_i = x) < 1$ – there is a chance of receiving either the treatment or control for every unit.

Problem Formulation. The main goal of matching is to infer the effect of the treatment by comparing the treated outcome $Y_\star(1)$ to that of a *matched unit* $\hat{Y}_\star^{\text{match}}$ that plays the role of a hypothetical "twin" of the treated unit had they not been exposed to treatment. In other words, $\hat{Y}_\star^{\text{match}}$ aims to be as close as possible to their (unobserved) *potential* outcome $Y_\star(0)$. More formally, we want the matched unit to approximate the expected outcome without treatment, $\mu_0(x) = \mathbb{E}[Y(0) \,|\, X = x]$, i.e., $\mathbb{E}[\hat{Y}_\star^{\text{match}}] \approx \mu_0(X_\star)$. A matched unit is constructed as an average of the outcomes of a few control units. In the following and without loss of generality, we focus on the nearest neighbor matching, i.e., the case when a single control unit gets assigned to the treated unit of interest. Let $j^\dagger$ denote the index of the nearest neighbor (hypothetical "twin"). A matched unit can then be built as:

$$\hat{Y}_\star^{\text{match}} = \sum_{j:T_j=0} w_j Y_j, \quad \text{where} \quad \begin{cases} w_j = 1 & \text{if } j = j^\dagger \text{ (matched)} \\ 0 & \text{if } j \neq j^\dagger \text{ otherwise.} \end{cases} \tag{1}$$

Note that, if we allow w_j to take continuous values, i.e., $0 \leq w_j \leq 1$, and $\sum_j w_j = 1$, we recover the synthetic controls framework (Abadie, 2021; Curth et al., 2024).

1. Figure 1D shows ITE/ATE estimation results corresponding to the synthetic 3D-swissroll dataset, which are later expanded in Section 5.2. Appendix B describes the data generation process in detail.

To find the best covariate match $X_{j^\dagger}$ for treatment unit $X_\star$ we perform nearest neighbours search among all control units according to some distance metric $d(\cdot, \cdot)$, namely:

$$j^\dagger = \underset{j \in [m]}{\text{argmin}}\ d(X_\star, X_j). \tag{2}$$

The most popular choice is the squared Euclidean distance, i.e., $d(X_\star, X_j) = ||X_\star - X_j||^2$, which we refer to as *standard matching method*. However, if there is underlying structure in the confounders given by the natural causal mechanisms of the problem, *which is the best distance to use and under which assumptions?* This question motivates our contribution GeoMatching, which we describe next.

3. GeoMatching

3.1. Assumptions

Manifold Hypothesis The *manifold hypothesis* states that the high-dim confounders X lie near a low-dim (non-linear) manifold embedded in the original input space (Whiteley et al., 2024). Low-dimensional structure arises due to constraints from physical laws and nature, or in other words, geometry often results from the underlying causal mechanisms between different covariates (Dominguez-Olmedo et al., 2023), as illustrated in Figure 1B. The manifold hypothesis is satisfied under common classes of Structural Causal Models; data types where it is often fulfilled include images of 3D objects, phonemes in speech signals, or video streams (Fefferman et al., 2016).

Geometric Faithfulness We assume that the treatment effect varies continuously along the manifold of confounders. Continuity ensures that points close to each other in the latent manifold will have similar treatment effects. This property is central in GeoMatching, as it allows us to ensure that small changes in the input (manifold points) do not lead to large, abrupt changes in the treatment effect. Back to the motivational example in Section 1, we expect athletes with similar anatomical constraints (reflected by the manifold structure) to respond similarly to different training strategies. As another example, a chemist might want to assess the effect of adding a catalyst to a chemical reaction (treatment) on the chemical reaction rate (outcome) given experimental conditions such as pH, purity of reactants, pressure, temperature, etc (confounders). As we move along the space of possible experimental conditions, we expect a smooth variation of the treatment effect.

3.2. A Geometric Take on Treatment Effect Matching

The distance between a treated and control unit can be defined as the length of a curve. A Euclidean distance for example corresponds to the length of a straight line. In general, if confounders X lie on a manifold $\mathcal{M}$ (potentially curved space), we define the length of a smooth curve $\gamma : [0, 1] \to \mathcal{M}$ as:

$$\text{Length}(\gamma\,; \mathbf{G}) = \int_0^1 \sqrt{\gamma'(r)^T \mathbf{G}(\gamma(r))\gamma'(r)dr}, \tag{3}$$

where $\gamma'(r) = \frac{d}{dr}\gamma(r)$ denotes the velocity of the curve, and $\mathbf{G}$ denotes a *Riemannian metric*, which is a smooth function that assigns a symmetric positive definite matrix $\mathbf{G}(x)$ to different locations $x \in \mathcal{M}$. Intuitively, $\mathbf{G}$ encapsulates an infinitesimal notion of distance on manifold $\mathcal{M}$.

Standard matching relies on a Riemannian metric $\mathbf{G} = I$, resulting in the squared Euclidean distance:

$$d_{\mathcal{E}}(X_\star, X_j) = (X_\star - X_j)^T \mathbf{I}(X_\star - X_j). \tag{4}$$

Mahalanobis matching methods rely on a constant Riemannian metric $\mathbf{G} = \Sigma$,

$$d_{\mathcal{M}}(X_\star, X_j) = (X_\star - X_j)^T \mathbf{\Sigma}^{-1} (X_\star - X_j). \tag{5}$$

GeoMatching relies on a *Riemannian distance* $d_{\mathcal{R}}(X_\star, X_j)$, a generalization of the Euclidean and Mahalanobis distance that replaces the constant covariance matrix $\mathbf{I}$ or $\mathbf{\Sigma}$ by an input-dependent covariance matrix $\mathbf{G}(x)$, and integrates out the infinitesimal distance contributions along the shortest curve – also called *geodesic curve* – connecting the two confounders along the manifold.

$$d_{\mathcal{R}}(X_\star, X_j) = \min_{\gamma_j} \text{Length}\Big(\gamma_j \,; \mathbf{G}\Big). \tag{6}$$

See Appendix A for further details and precise mathematical definitions of key Riemannian concepts.

3.3. Algorithm Description

We can now introduce GeoMatching, a geometry-aware matching framework that leverages Riemannian distances to pair treated and control units.

Step 1: Learn latent Riemannian manifold. GeoMatching learns a topology-preserving low-dimensional representation $Z \in \mathcal{Z}$ and latent Riemannian metric $\mathbf{G}$ given confounders $X \in \mathcal{X}$. Considering a latent representation instead of the original input space makes it easier for metric learning, specially if the input space is high-dimensional, and enables to filter out irrelevant information unrelated to the geometry of the confounders. Several methods can be adopted for this stage. One option is to assume a linear projection using Principal Component Analysis (PCA)[2] and fit a parameterized Riemannian metric $\mathbf{G}$. An alternative is to train a probabilistic latent variable model to jointly learn Z and $\mathbb{E}[\mathbf{G}]$, propagating Jacobian information (Tosi et al., 2014). In the experimental Section 5, we resort to the former strategy for simplicity, and posit a parametric Local Inverse Variance (LIV) Riemannian metric (Arvanitidis et al., 2016) in the latent space $\mathcal{Z}$, which is defined as the inverse of a local diagonal covariance matrix with j-th entry defined as:

$$\mathbf{G}_{jj}(z) = \left(\sum_{i=1}^{N} w_i(z) \left(z_{ij} - z_j\right)^2 + \rho \right)^{-1}, \quad \text{where} \quad w_i(z) = \exp\left(-\frac{||z_i - z||^2}{2\sigma^2} \right). \tag{7}$$

Parameter σ controls how fast uncertainty increases as we move away from the manifold; parameter ρ is needed for numerical instability, influencing the magnitude values that the metric can take. LIV assumes a Euclidean metric locally, which is then regularized to have large volume measure in regions of the feature space far away from observations. The inductive bias underlying such metric is that shortest paths on the data manifold should remain close to the observed data, we want curves to avoid crossing low-density high-uncertainty regions. A key advantage of this metric is that geodesic computation is less costly, as off-diagonal terms in metric $\mathbf{G}$ are zero.

Step 2: Compute Riemannian distances. We compute geodesic or Riemannian distances in the latent representation between treated and control units:

$$d_{\mathcal{R}}(X_\star, X_j) = \min_{\gamma_j} \text{Length}\Big(\gamma_j \,; \mathbf{G}\Big), \tag{8}$$

2. Note that PCA is not guaranteed to preserve geometric structure in general, for which other algorithms such as Isomap or Multidimensional Scaling might be more suitable.

where $\gamma_j : [0,1] \to \mathcal{Z}$ is a curve connecting the latent representations of the confounders, i.e., $\gamma_j(0) = Z_\star$, $\gamma_j(1) = Z_j$. To compute geodesic distances, we instantiate a parameterized (discretized) geodesic curve[3] and optimize its parameters by minimizing the following second-order ordinary differential equation (Do Carmo and Flaherty Francis, 1992):

$$\gamma'' = -\frac{1}{2}\mathbf{G}^{-1}\left[\frac{\partial \operatorname{vec}\mathbf{G}}{\partial \gamma}\right]^T (\gamma' \otimes \gamma'), \tag{9}$$

where $\operatorname{vec}\mathbf{G}$ stacks the columns of $\mathbf{G}$ and $\otimes$ denotes the Kronecker product. Based on Picard-Lindelöf theorem (Tenenbaum and Pollard, 1985), geodesics are guaranteed to exist and are locally unique given a starting point and initial velocity γ'. See Appendix E for further details and discussion on computational cost of GeoMatching.

Step 3: Matching along Riemannian manifold. Given the Riemannian distances from Equation (8), we match treated and control units along the Riemannian manifold by minimizing:

$$j^\dagger = \underset{j\in[m]}{\arg\min}\, d_{\mathcal{R}}(X_\star, X_j). \tag{10}$$

Step 4: Estimate causal treatment effects. Given the matched units $\hat{Y}_\star^{\text{match}}$ defined in Equation (1), we can estimate the individual and average treatment effects as follows:

$$\widehat{\text{ITE}} = Y_\star - \hat{Y}_\star^{\text{match}}, \tag{11}$$

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i^{\text{match}}\right)^{T_i=1}\left(\hat{Y}_i^{\text{match}} - Y_i\right)^{T_i=0}. \tag{12}$$

4. Related Work

Matching in Latent Representations. Several previous works learn low-dim representations before matching to mitigate undesirable effects of high dimensions in causal inference. Clivio et al. (2022) leverage the representation of intermediate layers of a neural network that predicts the propensity score $p(T|X)$, showing that such representation is approximately a balancing score, i.e., $X \perp T|Z$. Luo and Zhu (2017) and Zhao et al. (2022) propose matching methods on linear projections of the data based on the sufficient dimensionality reduction property, i.e., $X \perp Y|Z$. Wang et al. (2021) and Li et al. (2016) propose matching algorithms for high-dim data: the former relies on variable selection for categorical data; the later conducts and aggregates matching on random projections of the data. Clivio et al. (2023) characterize the bias added to a weighting estimator when using a data representation and propose an objective to learn suitable representations for causal inference. In contrast to our work, none of these methods account for the geometry of confounders.

Geometry-aware Machine Learning Models. Accounting for data geometry has been the object of study of several works in the machine learning community. Tosi et al. (2014) propose to learn a latent representation equipped with a Riemannian metric by assuming a Gaussian Process prior on the mapping function from latents to observations. Arvanitidis et al. (2021), Shao et al. (2017) and Chen et al. (2018) propose to exploit the geometry of the latent representation in deep generative

3. Considering a discretized curve for the geodesic scales more gracefully than other graph-based approaches, as the discretization of the curve is always one-dimensional independently of the dimension of the latent space.

models. These works conclude that treating the latent space as a curved Riemannian space instead of a Euclidean space yields several benefits for clustering, interpolation and prediction. A caveat of those methods is that the computation of geodesics is expensive and selection of hyperparameters for downstream tasks is often non trivial. More recently, Dominguez-Olmedo et al. (2023) show the benefits of accounting for the geometry in the generation of causally-grounded counterfactual explanations for ML classifiers.

Geometry-aware Causal Inference. While there has been extensive work on learning suitable latent representations for causal inference, very few have considered the geometry of the latent space for treatment effect estimation. Yan et al. (2024) is the only work we found that proposes to consider the geometry of confounders to improve TE estimation. They propose an optimal transport framework to promote the assignment of low weights for outliers, but they still assume Euclidean distances as opposed to a generic Riemannian distance along the data manifold. To our knowledge, our work is the first one to propose geometry-aware TE estimation based on Riemannian geometry.

5. Experiments and Results

We evaluate GeoMatching on synthetic and real-world datasets against other matching strategies in a synthetic swissroll toy scenario, a motion capture data, and two standard causality benchmark datasets. We show that GeoMatching mantains stability as we increase input dimensionality or in the presence of outliers, and benefits from semi-supervised scenarios.

5.1. Experimental Setup

In order to learn the latent Riemannian manifold (Step 1 in Section 3), we project the data using a deterministic dimensionality reduction technique (either PCA or Isomap) and posit a parametric smoothly changing kernel-based Riemannian metric defined by Equation (7). This choice is motivated twofold: i) the resulting pipeline is very simple and easy to implement, with effectively a single parameter σ for model selection, ii) we also explored learning the latent Riemannian metric using a Gaussian Process latent variable model, but found that it was less suitable in practice in terms of computational cost and quality of results. In terms of compute resources, we used 1 CPU for each run, with 10 GB and running time of $<$ 24h for each experiment.

Baselines. We evaluate the performance of GeoMatching according to two dimensions: i) which space gets considered for matching (the original input space $\mathcal{X}$, or a latent space $\mathcal{Z}$), and ii) which distance we rely on (either Euclidean, Mahalanobis, or Riemannian). We also include random matching to have a reference baseline on how difficult it is to match treated and control units on that particular dataset. In terms of latent representation $Z \in \mathcal{Z}$, we consider three alternatives: PCA, Isomap, and the latent representation of Neural Score Matching (NSM) from Clivio et al. (2022) which corresponds to the first layer of a neural network that predicts treatment assignment. Accounting for all combinations of matching space and considered distance, this gives us a total of 9 baselines and 3 ablations of GeoMatching.

Evaluation. For evaluation, we qualitatively inspect the matched units (geodesic curves) in the latent space, and quantitatively assess performance of treatment effect estimation via Average Treatment Effect (ATE) absolute error, defined as the absolute difference between the true and estimated ATEs. We also report Precision in Estimation of Heterogeneous Effect (PEHE) and distributions of

Individual Treatment Effect (ITE) in Appendix C. When plotting the latent representation of GeoMatching, we visualize the *magnification factor* – also called Riemannian volume – which measures the magnitude of the local distortion of space at location $p \in \mathcal{M}$, i.e., $V_{\mathbf{G}}(p) := \sqrt{\det \mathbf{G}(p)}$. Curves that traverse regions with high magnification factor tend to have larger lengths; it can also be seen as thr energy needed to traverse such space. Geodesics or shortest paths tend to avoid regions with high values of magnification factor, which reflect high uncertainty away from the observed data manifold.

Model Selection. We use single nearest neighbour matching strategy with replacement for all considered matching methods. We split our data 75/25 into train and test set: we use the train set to optimize parameter σ of the LIV Riemannian metric (which controls how fast the metric increases as we move away from the data manifold), by minimizing ATE absolute error in the train set. For the real-world datasets, we also select the dimensionality K of the latent representation based on performance on the train set, sweeping over $K = \{2, 3, 4, 5, 6\}$ for Lalonde dataset, and $K = \{2, 4, 6, 10\}$ for IHDP. For synthetic datasets, we fix the latent dimensionality $K = 2$. we report performance in both Train and Test set, averaged over 20 training random seeds for synthetic/semi-synthetic datasets, and 5 different training random seeds for real-world scenarios.

5.2. Synthetic Swissroll Dataset

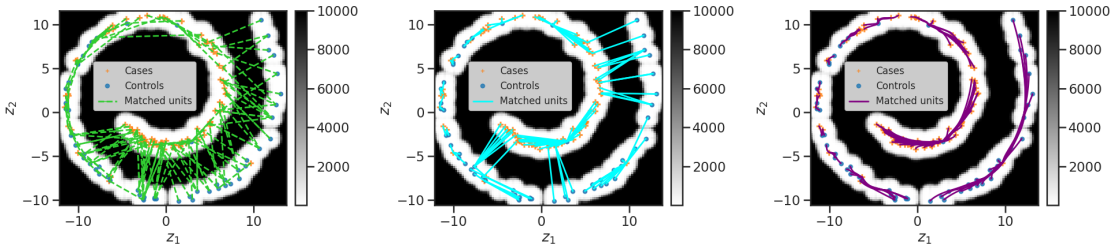
We adapt the classic swissroll dataset (see Appendix B for precise details on the data generation process) for the task of treatment effect estimation. We generate $N = 200$ observations where (pre-treatment) covariates X follow a 3D-swissroll manifold, treatment assignment variables T are Bernoulli draws from a smoothly varying sigmoid function along the swissroll manifold, and outcome variables Y are drawn from linear models. To assess the impact of input dimensionality on the different matching strategies, we embed the 3D-swissroll in a space of higher dimensionality $D \in \{3, 5, 10, 15, 25, 50, 100\}$, by adding additional noisy (irrelevant) input dimensions.

For this dataset, we project the input covariates into a 2D latent space. Both PCA and Isomap return similar latent representations, preserving the geometry of the swissroll. Here we report results for PCA. Figure 2 summarizes results on the swissroll dataset. When inspecting matched units (geodesic curves) in the latent space (Figures 2(a) to 2(c)), all baselines disregard the manifold geometry, pairing cases and controls across areas of high uncertainty. Instead, GeoMatching successfully couples cases and controls along the swissroll manifold by accounting for the confounders geometric structure and data uncertainty as we move away from the data manifold.

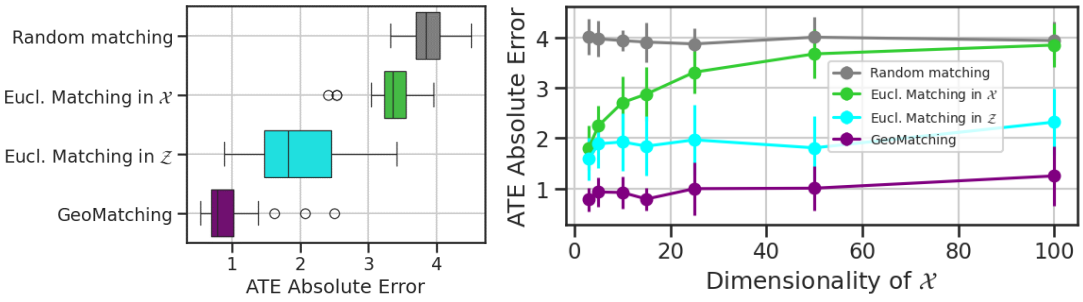
Looking at performance for TE estimation in Figures 2(d) and 2(e), GeoMatching yields the most accurate ATEs. As we increase the dimensionality D of input space $\mathcal{X}$, performance of euclidean matching in $\mathcal{X}$ degrades until reaching the same performance as random matching. Unsurprisingly, euclidean matching in $\mathcal{Z}$ remains competitive as D increases since it removes irrelevant dimensions (like GeoMatching) when projecting the data to a low-dim latent space.

5.3. Semi-synthetic Human Motion Capture

We use a semi-synthetic scenario of human motion capture data, which consists of real-world input covariates and synthetic treatment and outcome assignment variables. Such setup lies in-between an ideal/simplified synthetic scenario (like the 3D-swissroll example from previous section) and a fully realistic scenario where there is no access to ground truth ITEs (like the LaLonde dataset in Section 5.4). The goal of this experiment is to demonstrate i) effectiveness of GeoMatching handling



(a) Euclidean in $\mathcal{X}$ (b) Euclidean in $\mathcal{Z}$ (c) GeoMatching $\mathcal{Z}$



(d) TE Estimation Performance (e) Impact of input space dimensionality D

Figure 2: **Results for swissroll synthetic toy dataset**. (a-c) Matched pairs according to baselines and our method GeoMatching, (d-e) ATE Absolute Errors averaged over 20 random seeds.

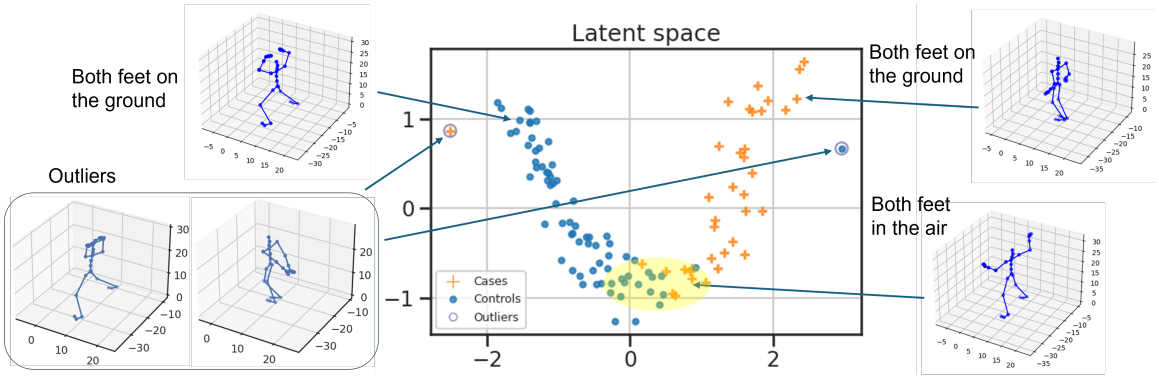
high-dim realistic covariates, and ii) robustness of GeoMatching against outlier confounders, in contrast to geometry-agnostic methods.

Human motion data has often been used to illustrate the impact of data geometry on downstream applications, as latent spaces are expected to be cylindrical or toroidal (doughnut-shaped) (Tosi et al., 2014; Urtasun et al., 2008). Specifically, we consider motion 16 from subject 22 from the CMU Motion Capture Database[4] which is a repetitive jumping jack motion. Each observation corresponds to a human pose as acquired by a marker-based motion capture system.

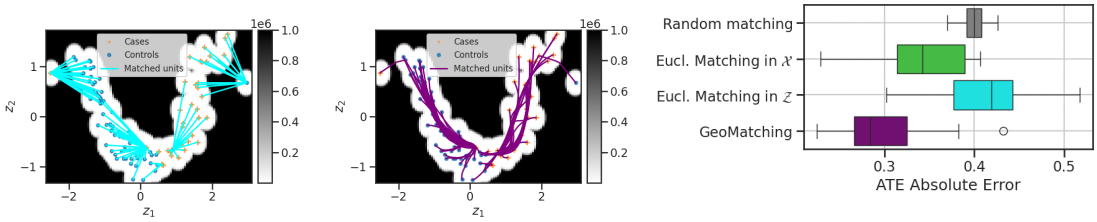
We adapt the CMU mocap data for the downstream task of TE estimation. To motivate this new setup, consider *Double Dutch*, a jump rope game that started in the streets and has now advanced to competitions, since 1974.[5]. To play, two people turn two long ropes in opposite directions while, multiple players jump simultaneously, entering the rope area one at a time. Let X be the motion capture of the current player jumping, T the strategy of entrance, either slow ($T = 0$) or fast ($T = 1$), and Y the probability of a successful entrance, when ropes do not trip up the new jumper's feet. To make the data more challenging, we intentionally perturb some time frames to simulate outlier confounders in the dataset (i.e., players jumping in a weird, unnatural manner, or motion capture sensors being defective). See Appendix B for further details about the data generation process.

Figure 3(a) shows a visualization of the data in a 2D-latent space using PCA projection, which preserves the geometric/periodic structure of the data. In this scenario, we expect GeoMatching's

4. http://mocap.cs.cmu.edu
5. https://youtu.be/iiEzf3J4iFk?feature=shared&t=49



(a) Data Description: CMU Motion Capture Data.



(b) Euclidean in $\mathcal{Z}$ (c) GeoMatching in $\mathcal{Z}$ (d) ATE Estimation Performance

Figure 3: **Results for semi-synthetic motion capture data**. (a) Visualization of the latent space and corresponding high-dim confounders after PCA projection. Outliers lie outside the intrinsic data manifold. (b-c) Matched units (geodesic curves) when pairing cases and controls according to Euclidean (baseline) or Riemannian metric (GeoMatching) in the latent space. (d) Average Treatment Effect Absolute Errors (avg across 20 random seeds).

advantage to be significant for those datapoints in the extremes of the manifold. Indeed, Figures 3(b) and 3(c) show that GeoMatching tends to match datapoints in the extremes to other datapoints inside the manifold, whereas Euclidean matching methods fail catastrophically, ignoring the manifold structure and matching several datapoints to the undesired outliers outside the manifold. All baselines are sensitive to the introduced outliers because they do not take into account data uncertainty; in contrast, GeoMatching is robust against the impact of outliers, which results in better ATE estimation, as shown in Figure 3(d).[6] These results align with our a priori expectations, as outliers tend to lie in low-density (high-volume) regions which are unlikely to be traversed by geodesics.

5.4. Real-world Causality Benchmark Datasets

We evaluate GeoMatching on two real-world datasets. We demonstrate that i) GeoMatching is competitive in more realistic scenarios, and ii) GeoMatching can benefit from semi-supervised settings, where we have access to several unlabelled covariate data, from which only a small subset has treatment and outcome information available. We consider the Infant Health Development Program Dataset and the Lalonde[7] dataset, see Appendix B for further details.

6. We note that Figure 3(d) looks less impressive than Figure 2(d) because ATE aggregates performance for all datapoints, and GeoMatching improves performance only for datapoints at the extremes of the manifold.
7. https://users.nber.org/~rdehejia/nswdata2.html

Table 1: **ATE absolute error for real-world datasets.**

(a) IHDP dataset.

	Original ($\mathcal{X}$)	PCA ($\mathcal{Z}$)	Isomap ($\mathcal{Z}$)	NSM ($\mathcal{Z}$)
Random	1.269 ± 0.053	NA	NA	NA
Euclidean	$\mathbf{1.126 \pm 0.035}$	1.137 ± 0.027	$\mathbf{1.107 \pm 0.012}$	1.166 ± 0.093
Mahalanobis	1.216 ± 0.063	1.136 ± 0.042	1.160 ± 0.051	1.207 ± 0.060
GeoMatching	1.131 ± 0.036	$\mathbf{1.118 \pm 0.029}$	1.112 ± 0.034	$\mathbf{1.157 \pm 0.094}$

(b) Lalonde dataset.

	Original ($\mathcal{X}$)	PCA ($\mathcal{Z}$)	Isomap ($\mathcal{Z}$)	NSM ($\mathcal{Z}$)
Random	610.1 ± 300.5	NA	NA	NA
Euclidean	541.8 ± 117.6	364.9 ± 219.9	1053.2 ± 921.3	811.5 ± 591.0
Mahalanobis	809.02 ± 158.8	843.6 ± 438.1	$\mathbf{766.5 \pm 468.4}$	948.8 ± 448.7
GeoMatching	$\mathbf{485.6 \pm 319.3}$	$\mathbf{363.2 \pm 122.6}$	809.6 ± 456.5	$\mathbf{561.4 \pm 396.3}$

Table 1 summarizes the results on real-world datasets. For PCA, GeoMatching improves upon all baselines. We also include results for the Isomap projection. In this case, Euclidean matching in $\mathcal{Z}$ is the best approach, slightly above GeoMatching. We hypothesize that this is due to a coarse grid when optimizing the σ parameter of the Riemannian metric (only 10 categorical values were explored in a grid search setup, see Appendix for details). In the case of the LaLonde dataset, using a Riemannian metric yield improvements in the Original covariate space, as well as in the PCA and NSM latent representations.

6. Conclusion

In this paper, we adopt a principled differential geometric approach for causal inference. We introduce GeoMatching, a novel matching method for treatment effect estimation that accounts for geometry structure and uncertainty of the confounders. The idea is to match cases and controls based on a latent Riemannian metric instead of a Euclidean metric, such that distances are more meaningful along the data manifold. We learn a non-parametric metric space by constructing a smoothly changing kernel-based deterministic metric that induces a Riemannian manifold. We empirically show that GeoMatching outbeats other baselines for treatment effect estimation, and remains compatitive for increasing input dimensionality, in the presence of outliers, and semi-supervised scenarios.

Limitations We note that the manifold learning stage, i.e., projecting the data in a way that preserves geometric structure, as well as the computational cost from computing geodesics can be improved. In general, neither PCA nor Isomap are probabilistic approaches, and neither of them are guaranteed to preserve geometric/topological structure. If the GeoMatching is performed in a latent representation that destroys geometric structure, treatment effects could be arbitrarily biased, potentially leading to negative societal impact. This is true for any matching method on latent representations. Computationally, the bottleneck is how to evaluate geodesics, which can be speed-up by minimizing a discretized length of the curve using graph information Chen et al. (2018), or using approximate Riemannian metrics Arvanitidis et al. (2022). We note that GeoMatching is a

framework rather than a closed-form procedure, allowing for enhancements in learning better latent representations or accelerating the computation of geodesics to be seamlessly integrated.

Broader Impact Geometry is an important inductive bias in ML and has been shown to yield benefits in diverse ML applications such as classification, interpolation, model training, and synthetic data generation. To the extend of our knowledge, this work is the first to look into the benefit of geometry-aware methods for causal inference. Estimation of treatment effects using differential geometry is a promising avenue for future research, given that causality often entails geometric manifold structures naturally (Dominguez-Olmedo et al., 2023). As future work, an interesting research direction would be to combine GeoMatching with causal discovery methods to inform the construction of the latent space. Instead of learning the entire manifold at once, one could learn all independent mechanisms separately and explicitly parameterize the manifold for computing geodesics, improving computational efficiency and accuracy. Another interesting research direction would be to account for geometry broadly in other causal inference methods, including meta-learners, synthetic controls or propensity-score weighting methods, leveraging Riemannian geometry instead of Euclidean spaces.

Acknowledgments

We thank Javier Zazo for continuous discussions and technical support. We also thank Aditya Nori for early discussions to the initial idea of this research.

References

Alberto Abadie. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425, June 2021. ISSN 0022-0515. doi: 10.1257/jel.20191450. URL https://pubs.aeaweb.org/doi/10.1257/jel.20191450.

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, 74(1):235–267, January 2006. ISSN 0012-9682, 1468-0262. doi: 10.1111/j.1468-0262.2006.00655.x. URL http://doi.wiley.com/10.1111/j.1468-0262.2006.00655.x.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Jan Van Den Bussche, and Victor Vianu, editors, *Database Theory — ICDT 2001*, volume 1973, pages 420–434. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-41456-8 978-3-540-44503-6. doi: 10.1007/3-540-44503-X_27. URL http://link.springer.com/10.1007/3-540-44503-X_27. Series Title: Lecture Notes in Computer Science.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. A Locally Adaptive Normal Distribution, September 2016. URL http://arxiv.org/abs/1606.02518. arXiv:1606.02518 [stat].

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent Space Oddity: on the Curvature of Deep Generative Models, December 2021. URL http://arxiv.org/abs/1710.11379. arXiv:1710.11379 [stat].

Georgios Arvanitidis, Bogdan M. Georgiev, and Bernhard Schölkopf. A prior-based approximate latent Riemannian metric. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 4634–4658. PMLR, May 2022. URL https://proceedings.mlr.press/v151/arvanitidis22a.html. ISSN: 2640-3498.

Jonathan Bac, Evgeny M. Mirkes, Alexander N. Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021. URL https://www.mdpi.com/1099-4300/23/10/1368. Publisher: MDPI.

Eli Ben-Michael, Kosuke Imai, and Zhichao Jiang. Policy Learning with Asymmetric Counterfactual Utilities, November 2023. URL http://arxiv.org/abs/2206.10479. arXiv:2206.10479 [cs, stat].

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. URL https://ieeexplore.ieee.org/abstract/document/6472238/. Publisher: IEEE.

John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016. URL https://books.google.com/books?hl=en&lr=&id=JlSvDAAAQBAJ&oi=fnd&pg=PR13&dq=Butcher,+J.+C.+(2016).+Numerical+Methods+for+Ordinary+Di%EF%BF%BDerential+Equations.+John+Wiley+%26+Sons.&ots=Sv9fUxVfW6&sig=pdNwMjB9P-grY7u26U96Kb2ASjE.

Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick van der Smagt. Metrics for Deep Generative Models, February 2018. URL http://arxiv.org/abs/1711.01204. arXiv:1711.01204 [cs, stat].

Oscar Clivio, Fabian Falck, Brieuc Lehmann, George Deligiannidis, and Chris Holmes. Neural score matching for high-dimensional causal inference. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 7076–7110. PMLR, May 2022. URL https://proceedings.mlr.press/v151/clivio22a.html. ISSN: 2640-3498.

Oscar Clivio, Avi Feller, and Chris Holmes. Towards representation learning for general weighting problems in causal inference. *NeurIPS 2023 Workshop on Causal Representation Learning*, 2023.

Oscar Clivio, Avi Feller, and Christopher C. Holmes. Towards Representation Learning for Weighting Problems in Design-Based Causal Inference. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=KNgBCZXJkY.

Alicia Curth and Mihaela van der Schaar. Doing Great at Estimating CATE? On the Neglected Assumptions in Benchmark Comparisons of Treatment Effect Estimators, July 2021. URL http://arxiv.org/abs/2107.13346. arXiv:2107.13346 [cs, stat].

Alicia Curth, Hoifung Poon, Aditya V. Nori, and Javier González. Cautionary Tales on Synthetic Controls in Survival Analyses, February 2024. URL http://arxiv.org/abs/2312.00501. arXiv:2312.00501 [stat].

Rajeev H. Dehejia and Sadek Wahba. Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs, June 1998. URL https://www.nber.org/papers/w6586.

Manfredo Perdigao Do Carmo and J. Flaherty Francis. *Riemannian geometry*, volume 2. Springer, 1992. URL https://link.springer.com/book/9780817634902.

Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, Georgios Arvanitidis, and Bernhard Schölkopf. On Data Manifolds Entailed by Structural Causal Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8188–8201. PMLR, July 2023. URL https://proceedings.mlr.press/v202/dominguez-olmedo23a.html. ISSN: 2640-3498.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, February 2016. ISSN 0894-0347, 1088-6834. doi: 10.1090/jams/852. URL https://www.ams.org/jams/2016-29-04/S0894-0347-2016-00852-4/.

Jennifer L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, January 2011. ISSN 1061-8600. doi: 10.1198/jcgs.2010.08162. URL https://doi.org/10.1198/jcgs.2010.08162. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/jcgs.2010.08162.

Nicholas Krämer, Nathanael Bosch, Jonathan Schmidt, and Philipp Hennig. Probabilistic ODE solutions in millions of dimensions. In *International Conference on Machine Learning*, pages 11634–11649. PMLR, 2022. URL https://proceedings.mlr.press/v162/kramer22b.html.

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 265–274, New York, NY, USA, August 2017. Association for Computing Machinery. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098032. URL https://doi.org/10.1145/3097983.3098032.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986. URL https://www.jstor.org/stable/1806062. Publisher: JSTOR.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/hash/74934548253bcab8490ebd74afed7031-Abstract.html.

Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns. In *IJCAI*, pages 3768–3774, 2016. URL https://www.ijcai.org/Proceedings/16/Papers/530.pdf.

Gabriel Loaiza-Ganem, Brendan Leigh Ross, Rasa Hosseinzadeh, Anthony L. Caterini, and Jesse C. Cresswell. Deep Generative Models through the Lens of the Manifold Hypothesis: A Survey and New Connections, April 2024. URL http://arxiv.org/abs/2404.02954. arXiv:2404.02954 [cs, stat].

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal Effect Inference with Deep Latent-Variable Models, November 2017. URL http://arxiv.org/abs/1705.08821. arXiv:1705.08821 [cs, stat].

Wei Luo and Yeying Zhu. Matching Using Sufficient Dimension Reduction for Causal Inference, February 2017. URL http://arxiv.org/abs/1702.00444. arXiv:1702.00444 [stat].

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. *Advances in neural information processing systems*, 23, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/hash/8a1e808b55fde9455cb3d8857ed88389-Abstract.html.

Hariharan Narayanan and Partha Niyogi. On the Sample Complexity of Learning Smooth Cuts on a Manifold. In *COLT*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/020.pdf.

Judea Pearl. Causal inference in statistics: An overview. 2009. URL https://projecteuclid.org/journals/statistics-surveys/volume-3/issue-none/Causal-inference-in-statistics-An-overview/10.1214/09-SS057.short.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. URL https://books.google.com/books?hl=en&lr=&id=I0V2CwAAQBAJ&oi=fnd&pg=PR9&dq=pearl+2016&ots=9Bm0xu2Jnk&sig=Pgovvq2tRnLX4-w7ht6sgrAlDpY.

Paul R. Rosenbaum. Optimal Matching of an Optimally Chosen Subset in Observational Studies. *Journal of Computational and Graphical Statistics*, 21(1):57–71, 2012. ISSN 1061-8600. URL https://www.jstor.org/stable/23248823. Publisher: [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America].

Donald B Rubin. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214504000001880. URL http://www.tandfonline.com/doi/abs/10.1198/016214504000001880.

Donald B. Rubin. For objective causal inference, design trumps analysis. 2008. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/For-objective-causal-inference-design-trumps/10.1214/08-AOAS187.short.

Jasjeet Singh Sekhon and Richard D. Grieve. A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics*, 21(6):695–714, June 2012. ISSN 1099-1050. doi: 10.1002/hec.1748.

Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. The Riemannian Geometry of Deep Generative Models, November 2017. URL http://arxiv.org/abs/1711.08014. arXiv:1711.08014 [cs, stat].

Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 25(1):1–21, February 2010. ISSN 0883-4237. doi: 10.1214/09-STS313. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943670/.

Morris Tenenbaum and Harry Pollard. *Ordinary differential equations: an elementary textbook for students of mathematics, engineering, and the sciences*. Courier Corporation, 1985. URL https://books.google.com/books?hl=en&lr=&id=iU4zDAAAQBAJ&oi=fnd&pg=PP1&dq=Ordinary+Differential+Equations.+Tenenbaum+Pollard&ots=z0ycQhx8ul&sig=u2pmhCzqOKpjr3qwpH0uwriLUDg.

Alessandra Tosi, Søren Hauberg, Alfredo Vellido, and Neil D. Lawrence. Metrics for Probabilistic Geometries, November 2014. URL http://arxiv.org/abs/1411.7432. arXiv:1411.7432 [cs, stat].

Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popović, Trevor J. Darrell, and Neil D. Lawrence. Topologically-constrained latent variable models. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1080–1087, Helsinki, Finland, 2008. ACM Press. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390292. URL http://portal.acm.org/citation.cfm?doid=1390156.1390292.

Tianyu Wang, Marco Morucci, M. Usaid Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. FLAME: A Fast Large-scale Almost Matching Exactly Approach to Causal Inference, February 2021. URL http://arxiv.org/abs/1707.06315. arXiv:1707.06315 [cs, stat].

Daniel Westreich, Jessie K. Edwards, Catherine R. Lesko, Elizabeth Stuart, and Stephen R. Cole. Transportability of Trial Results Using Inverse Odds of Sampling Weights. *American Journal of Epidemiology*, 186(8):1010–1014, October 2017. ISSN 1476-6256. doi: 10.1093/aje/kwx164.

Nick Whiteley, Annie Gray, and Patrick Rubin-Delanchy. Statistical exploration of the Manifold Hypothesis, February 2024. URL http://arxiv.org/abs/2208.11665. arXiv:2208.11665 [cs, stat].

Yuguang Yan, Zeqin Yang, Weilin Chen, Ruichu Cai, Zhifeng Hao, and Michael Kwok-Po Ng. Exploiting Geometry for Treatment Effect Estimation via Optimal Transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16290–16298, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/29564. Issue: 15.

Haoran Zhao, Yinghao Zhang, Debo Cheng, Chen Li, and Zaiwen Feng. Matching Using Sufficient Dimension Reduction for Heterogeneity Causal Effect Estimation. In *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 9–16, December 2022. doi:

10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00036. URL https://ieeexplore.ieee.org/document/10074882.

Appendix A. Concepts of Riemannian Geometry

In this Section, we review key concepts of Riemannian geometry that will be useful for our work.

Definition 1 *(Manifold). A d-dimensional smooth manifold $\mathcal{M}$ is a topological space which locally resembles the Euclidean space $\mathbb{R}^d$ and has a smooth structure.*

Definition 2 *(Riemannian Manifold). A Riemannian manifold is a differentiable (smooth) manifold $\mathcal{M}$ provided with a Riemannian metric tensor $\mathbf{G}$.*

Definition 3 *(Riemannian metric). A Riemannian metric $\mathbf{G}$ on a manifold $\mathcal{M}$ is a smooth function that assigns a symmetric positive definite matrix to any point $z \in \mathcal{M}$.*

A Riemannian metric defines at each point z a smoothly varying inner product in the tangent space $T_z\mathcal{M}$. The inner product is defined as:

$$\langle z_1, z_2 \rangle_z = z_1^{\mathrm{T}} \mathbf{G}(z) z_2, \tag{13}$$

where $z_1, z_2 \in T_z\mathcal{M}$ and $z \in \mathcal{M}$. Intuitively, a Riemannian metric defines an infinitesimal notion of distance on the manifold $\mathcal{M}$. The length of a smooth curve $\gamma : [0, 1] \to \mathcal{M}$ is then defined as:

$$\mathrm{Length}(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^T \mathbf{G}(\gamma(t)) \dot{\gamma}(t) dt}, \tag{14}$$

where $\dot{\gamma}(t) = \frac{d}{dt}\gamma(t)$ denotes the velocity of the curve. The distance between two points on the manifold $\mathcal{M}$ is defined as the length of the shortest curve connecting them.

Definition 4 *(Geodesic). A geodesic curve between two points x_1 and x_2 is a length-minimising curve connecting the two points:*

$$\gamma_{\mathbf{G}} = \underset{\gamma}{\mathrm{argmin}}\ \mathrm{Length}(\gamma), \qquad \gamma(0) = x_1, \gamma(1) = x_2. \tag{15}$$

We call *Riemannian* or *geodesic distance* the length of the geodesic curve, according to the underlying Riemannian metric $\mathbf{G}$.

Definition 5 *(Riemannian distance). The Riemannian distance $d_{\mathbf{G}}(r, s)$ between two points $r, s \in \mathcal{M}$ on a Riemannian manifold $(\mathcal{M}, \mathbb{G})$ is defined as the infimum of the length of all smooth curves $\gamma : [0, 1] \to \mathcal{M}$ connecting r and s,*

$$d_{\mathbf{G}}(r, s) = \inf\{\mathrm{Length}(\gamma) \,|\, \gamma(0) = \mathrm{r},\ \gamma(1) = \mathrm{s}.\} \tag{16}$$

Definition 6 *(Magnification factor) The magnification factor – also called Riemannian volume – measures the magnitude of the local distortion of space at location $p \in \mathcal{M}$, i.e., $V_{\mathbf{G}}(p) := \sqrt{\det \mathbf{G}(p)}$.*

Curves that traverse regions with high magnification factor will tend to have larger lengths (it can also be seen as how much energy is needed to traverse such space). Geodesics or shortest paths tend to avoid regions with high values of magnification factor (which reflect high uncertainty away from the observed data manifold).

Appendix B. Further Information about Datasets

Synthetic Data Generation for the Swissroll Dataset. In Section 5.2, we generate a three-dimensional swissroll in a space of higher dimensionality $D \in \{3, 5, 10, 15, 25, 50, 100\}$. The data generation process for a specific datapoint (X, T, Y) is as follows:

$$
\begin{aligned}
R &\sim \mathcal{U}([1.5\pi, 4.5\pi]) && (17)\\
X_1 &= R\cos(R) && (18)\\
X_2 &= R\sin(R) && (19)\\
X_3 &\sim \mathcal{U}([0, 8]) && (20)\\
X_4, \ldots, X_D &\sim \mathcal{N}\left(0, \sigma_x \mathbf{I_{D-3}}\right) && (21)\\
T &\sim \text{Bernoulli}\left(\text{sigmoid}(3\pi)R\right) && (22)\\
\mu_0(x) &= -\frac{1}{2}x + \frac{3}{2}\pi && (23)\\
\mu_1(x) &= -\frac{3}{2}x + \frac{9}{2}\pi && (24)\\
Y(0) &\sim \mathcal{N}(\mu_0(R - \frac{3}{2}\pi), \sigma_y \mathbf{I}) && (25)\\
Y(1) &\sim \mathcal{N}(\mu_1(R - \frac{3}{2}\pi), \sigma_y \mathbf{I}) && (26)\\
Y &= (1-T)Y(0) + TY(1), && (27)
\end{aligned}
$$

where R is the intrinsic manifold where the data lives, σ_x is the standard deviation for the additional uninformative input dimensions, and σ_y is the noise standard deviation for outcome Y. Note that the results we report in the main text assume a 2D latent representation, we are essentially matching in the (X_1, X_2) space, filtering out the rest of input covariates. While for this specific toy, we could have mapped the original space $\mathcal{X}$ into a 1D Euclidean space, in which case the Riemannian distance would be equivalent to the Euclidean distance, this is not always feasible. Thus, we deliberately project the original confounders into a 2D latent space to illustrate the more generic situation in which the data manifold cannot be mapped to a Euclidean space.

Semi-synthetic Data Generation for the CMU Mocap Dataset. We consider real-world covariates from the CMU Motion Capture Database[8] (motion 16 from subject 22), and we synthetically generate the treatment and outcome variables. We chose human motion data because it has been used in previous works to illustrate the impact of geometry on downstream applications, as latent spaces are expected to be cylindrical or toroidal/doughnut-shaped (Urtasun et al., 2008; Tosi et al., 2014).

Regarding the covariates, for each random seed we inject two outliers by randomly selecting a motion capture frame of each side of the data manifold (corresponding to frames with feet close to the ground, one treated and one control unit). We then obtain a perturbed frame by linear interpolation of these two covariates, in order to get a realistic motion capture outside of the main data manifold.

8. http://mocap.cs.cmu.edu

The data generation process for the generation of (T, Y) is as follows:

$$
\begin{aligned}
R &\sim \mathcal{U}([0, 6\pi]) && (28)\\
T &\sim \text{Bernoulli}\left(\text{sigmoid}(z_1)\right) && (29)\\
\mu_0(x) &= 0.4\sin(x + \frac{1}{2}) + \frac{1}{2} && (30)\\
\mu_1(x) &= 0.4\sin(x + \frac{1}{2}) + \frac{1}{2} && (31)\\
Y(0) &\sim \mathcal{N}(\mu_0(R), \sigma_y \mathbf{I}) && (32)\\
Y(1) &\sim \mathcal{N}(\mu_1(R), \sigma_y \mathbf{I}) && (33)\\
Y &= (1-T)Y(0) + TY(1), && (34)
\end{aligned}
$$

IHDP Dataset. The Infant Health and Development Program (IHDP) is a randomized controlled study conducted between 1985 and 1988, designed to evaluate the effect of home visit from specialist doctors on the cognitive test scores of premature infants. (Hill, 2011). We chose this dataset because it is commonly used as a benchmark dataset in the causality community, as it allow us to assess performance against ground truth values Hill (2011); Curth and van der Schaar (2021). We use the synthetic version of IHDP introduced by Louizos et al. (2017), as it allow us to assess performance against ground truth values. We preprocess the data according to the recommendations in Curth and van der Schaar (2021).

LaLonde Dataset. The LaLonde dataset is a dataset from econometrics which looks at the impact of training program policies in employee's salary (LaLonde, 1986; Dehejia and Wahba, 1998). The treatment is whether the participant attended a job training program, the outcome is the earning in 1978, and pre-treatment covariates include 6 covariates capturing demographic information. The dataset contains interventional (RCT) and observational data. Replicating the experimental protocol from Kuang et al. (2017), we consider treated units from the RCT, and controls from observational data. Thanks to the complete RCT data, we can estimate a ground truth value for the ATE. Our goal is then to estimate the ATE using controls from observational data, which requires handling confounding bias/correcting for distribution imbalance between cases and controls.

Appendix C. Further Empirical Results

C.1. Additional Results for Datasets considered in Main Manuscript

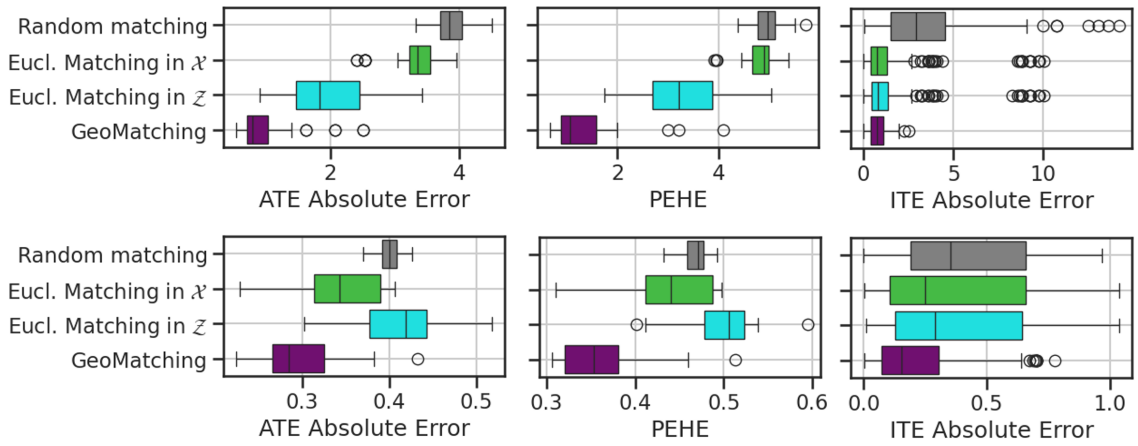


Figure 4: **Additional Results for (Semi)-Synthetic Datasets**. (1st row) Swissroll data; (2nd row) Motion Capture data. 1st column is the same as in the main text to facilitate comparison; 2nd column corresponds to Precision in Estimation of Heterogeneous Treatment Effects (PEHE) averaged over all seeds, and 3rd column shows Individual Treatment Effect (ITE) Absolute Error for a single seed.

Table 2: **Precision in Estimation of Heterogeneous Effect (PEHE).**

(a) IHDP dataset.

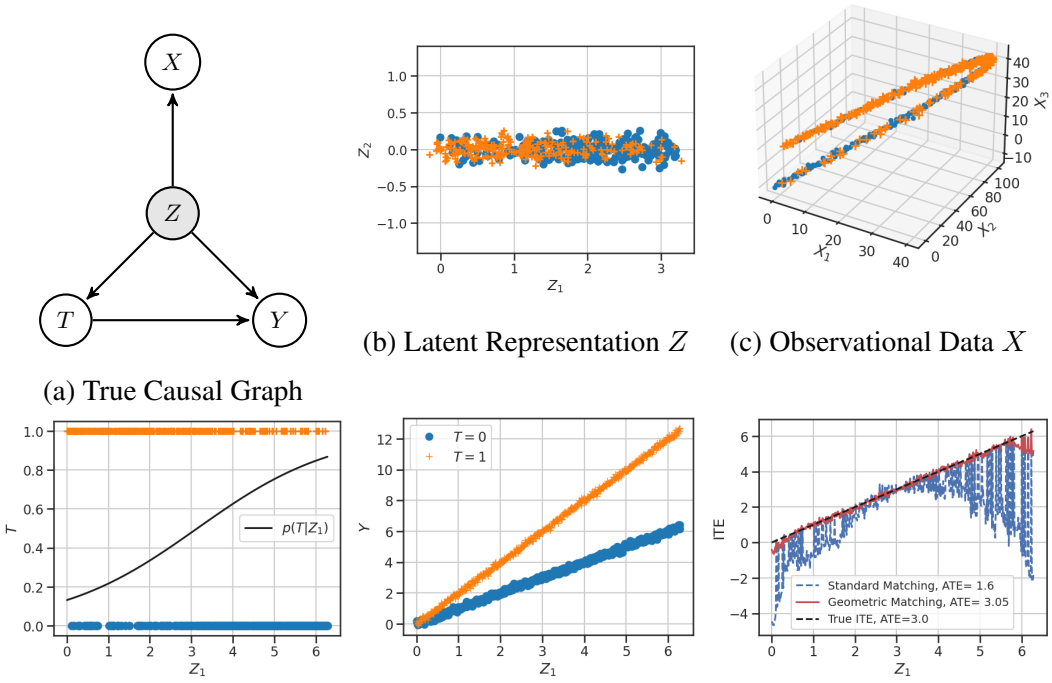
	Original ($\mathcal{X}$)	PCA ($\mathcal{Z}$)	Isomap ($\mathcal{Z}$)	NSM ($\mathcal{Z}$)
Random	1.613 ± 0.083	NA	NA	NA
Euclidean	1.452 ± 0.016	$\mathbf{1.399 \pm 0.024}$	1.422 ± 0.034	1.448 ± 0.113
Mahalanobis	1.543 ± 0.048	1.433 ± 0.035	1.409 ± 0.043	1.506 ± 0.087
GeoMatching	$\mathbf{1.433 \pm 0.023}$	1.416 ± 0.056	$\mathbf{1.390 \pm 0.042}$	$\mathbf{1.438 \pm 0.107}$

Table 3: For Lalonde dataset, no true ITE available, so we cannot compute PEHE.

C.2. Results for U-shaped Strip Synthetic Toy

We generate an additional synthetic toy consisting of a U-shaped Strip as follows:

$$
\begin{aligned}
Z &\sim \mathcal{U}([0, 2\pi]) && (35)\\
X \,|\, Z &\sim \mathcal{N}(f(Z), \sigma_x^2 I) && (36)\\
T \,|\, Z &\sim \mathcal{N}(\text{sigmoid}(Z_1; k, z_0), \sigma_x^2 I) && (37)\\
Y \,|\, T, Z &\sim T\mathcal{N}(2Z_1; \sigma_Y^2) + (1-T)\mathcal{N}(Z_1; \sigma_Y^2) && (38)\\
ITE \,|\, Z &\sim \mathcal{N}(Z_1; 2\sigma_Y^2) && (39)
\end{aligned}
$$



(a) True Causal Graph (b) Latent Representation Z (c) Observational Data X

(d) Treatment Assignment T (e) Treatment Response Y (f) ATE Estimation

Figure 5: **Results for U-shaped Strip Synthetic Toy.** This proof-of-concept toy shows that accounting for data geometry improves estimation of treatment effects.

Increasing dimensionality of observational space for U-shape synthetic toy Similar to Figure 2(e) in the main text, we embed the U-shape strip in a high-dimensional space by adding additional uninformative input dimensions. Standard matching degrades as we increase the input dimensionality, whereas GeoMatching performance remains stable.

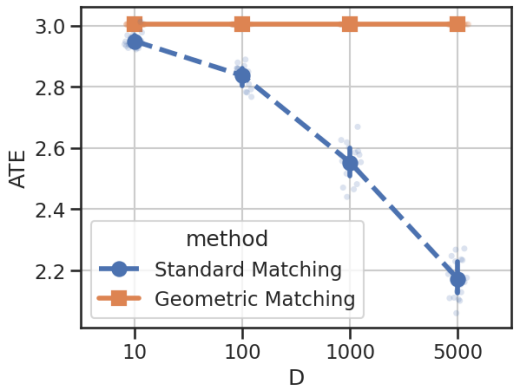


Figure 6: **Impact of dimensionality D on TE estimation.** As D increases, matching all covariates perfectly becomes impossible.

C.3. Sensitivity Analysis w.r.t Hyperparameters

Figure 7 depicts variation of TE estimation performance as we vary the two hyperparameters considered in this paper: the dimensionality of the latent representation and the curvature σ of the Riemannian manifold. GeoMatching is sensitive to the selected hyperparameters (manifold learning is a non-trivial task in general). We selected hyperparameters based on ATE performance in a separate validation set. Worth noting is that, even if the grid search we ran was quite coarse (due to limited compute), we were still able to observe improvements empirically. We hypothesize that results would be much better with a finer grid, or using Bayesian optimization to dynamically optimize parameter σ, which controls the curvature of the Riemannian manifold.

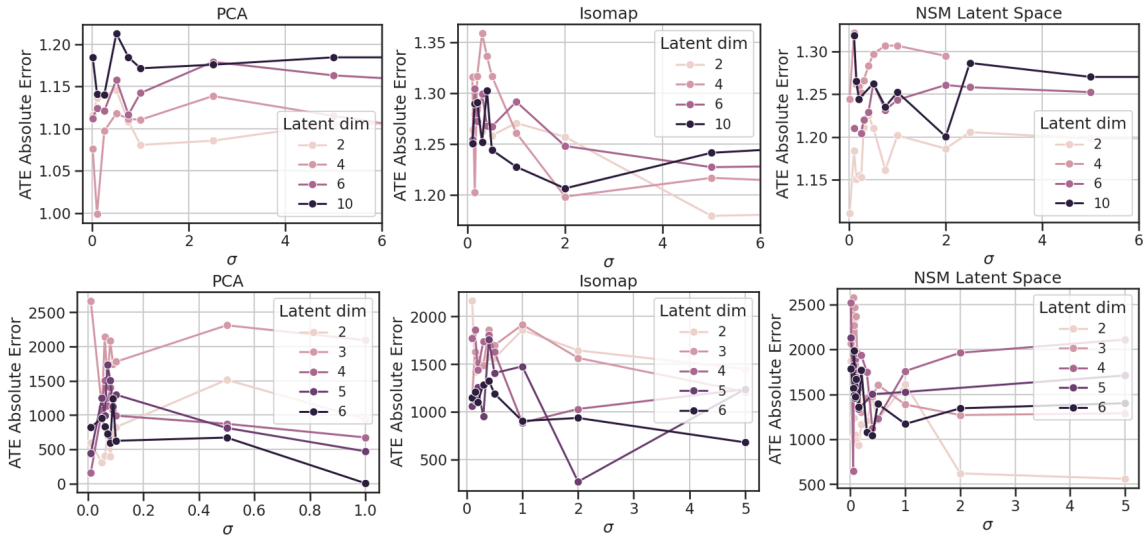


Figure 7: **Sensitivity Analysis of GeoMatching w.r.t different Hyperparameters**. We consider variability for a single random seed in ATE Absolute Error performance w.r.t different latent representations (PCA, Isomap, and NSM latent space), latent dimensionality, and σ, the parameter that controls the curvature of the Riemannian manifold.

Appendix D. Bias Decomposition

In this section, we conduct a theoretical analysis based on the decomposition of the estimation bias for GeoMatching. In the following, we assume that the *manifold hypothesis* and *geometric faithfulness* assumptions stated in Section 3.1 hold, resulting in a *well-specified* representation of covariates. The geometric faithfulness assumption ensures that points close to each other in the latent manifold will have similar treatment effects. By decomposing the estimation bias, we show that the Riemannian metric is optimal (minimizes extrapolation bias) when the mapping function from covariate to outcome space is an isometry.

Let us consider covariates $X_\star \in \mathbb{R}^d$ of a treated individual ($T_\star = 1$). Let $\mu_t(x) = \mathbb{E}[Y(t) \,|\, X = x]$ denote the expected outcome under treatment $T = t$. Our goal is to estimate the individual treatment effect $\text{ITE}(X_\star) = |\mu_1(X_\star) - \mu_0(X_\star)|$, where $\mu_0(X_\star)$ is unobserved. Given a fixed distance metric $d(\cdot, \cdot)$ in covariate space, we can solve $j^\dagger = \underset{j \in [m]}{\text{argmin}}\, d(X_\star, X_j)$ among a pool of m controls, and build an ITE matching estimator:

$$\widehat{\text{ITE}}(X_\star; d) = |\mu_1(X_\star) - \mu_0(X_{j^\dagger(d)})| = |\underbrace{\mu_1(X_\star) - \mu_0(X_\star)}_{\text{ITE}(X_\star)} + \underbrace{\mu_0(X_\star) - \mu_0(X_{j^\dagger(d)})}_{\text{Bias}(X_\star; d)}|, \tag{40}$$

where $\text{Bias}(X_\star; d)$ is the *extrapolation bias* or cost to pay from using covariate $X_{j^\dagger(d)}$ instead of $X_\star$.

When matched units are imperfect, matching estimators need to extrapolate, incurring in a non-zero extrapolation bias:

$$d(X_\star, X_{j^\dagger(d)}) \geq 0 \quad \Rightarrow \quad \text{Bias}(X_\star; d) = |\mu_0(X_\star) - \mu_0(X_{j^\dagger(d)})| \geq 0. \tag{41}$$

For example, if the expected outcome without treatment is linear $\mu_0(x) = \beta x$, we incur a bias $d(X_\star, X_{j^\dagger(d)})\beta$ directly proportional to the distance to the match in covariate space. We say a distance is optimal if it minimizes the bias of the matching estimator:

$$d^\dagger = \underset{d}{\text{argmin}}\, \text{Bias}(X_\star; d). \tag{42}$$

GeoMatching pairs treated and control units along the Riemannian manifold by minimizing:

$$j^\star = \underset{j \in [m]}{\text{argmin}}\, d_{\mathcal{R}}(X_\star, X_j). \tag{43}$$

Ideally, we would like to match treated and control units such that the extrapolation bias gets minimized:

$$j^\star = \underset{j \in [m]}{\text{argmin}}\, |\mu_0(X_\star) - \mu_0(X_j)|. \tag{44}$$

Thus, GeoMatching is optimal, i.e., minimizes the extrapolation bias, when the expected outcome function μ_0 is an isometry, i.e., a distance preserving transformation from latent to outcome space.

We note that the provided analysis elucidates which condition needs to be fulfilled for the extrapolation bias to be minimized, and holds for any distance metric d, including Riemannian distance (but it is not specific to the Riemannian distance). What is specific to GeoMatching are the two key underlying assumptions stated in Section 3.1.

The analysis in this section studies the error due to *the choice of distance metric* given a fixed representation of covariates, i.e., which distance metric minimizes extrapolation bias assuming a

well-specified representation. In contrast to our case, Clivio et al. (2024) study the error due to the *choice of the representation* given a fixed distance metric. A future research direction would be to provide a full theoretical characterization of GeoMatching by leveraging the theoretical results of Clivio et al. (2024).

Appendix E. Computational Impact of GeoMatching

Solving an ODE, to compute a single geodesic distance, scales linearly with the intrinsic dimensionality of the latent space (Krämer et al., 2022). We need to solve one ODE to compute each pairwise distance. The current implementation of GeoMatching (which we will make publicly available), relies on naïve nearest neighbour search, which scales linearly with the latent dimensionality, and quadratically with the number of datapoints.

We leverage the Stochman public library for computation of geodesics. We compute geodesics using CPUs and sequentially, which is unsurprisingly not very fast (each experiment would take around one night to run, we launched a Wandb grid sweep in a cluster of 1000 nodes in parallel).

We note that our work does not focus on computational efficiency, but on a conceptual and methodological improvement of matching methods. Making GeoMatching computationally efficient is a promising research direction, which can be done in several regards: i) computing geodesics faster by constraining the space of Riemannian metrics to learn (Arvanitidis et al., 2022), ii) leveraging extensions to nearest neighbor search such as Locality Sensitive Hashing to scale sub-linearly instead of quadratically, iii) exploiting hardware improvements, using GPUs and parallel computation.

Appendix F. Frequently Asked Questions

Can we apply GeoMatching to out of sample data?

Computing TEs for an unknown observation (very different from any sample seen so far in the train set) is problematic for any matching method, as it violates the overlap assumption (Rubin, 2005). All reported results correspond to in-distribution performance; out-of-distribution is out of scope.

Does causal structure always imply geometric structure?

Some geometrical structures could be implied by some causal structure but not all causal structure would reflect on the structure of the data. We note that geometric structure alone might not be enough to recover causal structure; this relates to identifiability of the causal graph. The key assumption of this work is that causal mechanisms most often constraint observations to be sparse/non i.i.d. (e.g., given a set of patient covariates, we would never get a realistic patient record if sampling from each covariate marginally) and smooth (e.g., there exists a smooth variation in patients vitals/measurements).

What is the role of the manifold hypothesis in Causal Inference?

The manifold hypothesis states that the high-dimensional covariates X lie near a low-dimensional (non-linear) manifold embedded in the original input space. Such hypothesis is not commonly discussed nor exploited in the causality field, despite being widely accepted within the machine learning community. This work aims to bring awareness to and exploit the manifold hypothesis in the context of causal inference, bridging the gap between both communities.

There is a plethora of arguments supporting the existence of low-dim structure in high-dim data (Loaiza-Ganem et al., 2024), which could directly benefit causal inference. First, the manifold hypothesis implies two properties that are commonly observed in real-world data: data sparsity in input space, and a notion of smoothness in observations. Second, theoretical works have shown that learning complexity scales exponentially with intrinsic dimensionality, and yet, modern algorithms are still successful at learning low-dim representations (Bengio et al., 2013), providing strong implicit justification for such geometric structure underneath. Finally, various works have directly estimated the intrinsic dimensionality to be orders of magnitude smaller than their input dimension for diverse datasets, including images or physics data (Levina and Bickel, 2004; Bac et al., 2021).

How does GeoMatching behave when the dimensionality of the manifold rises?

The difficulty of learning a manifold (i.e., number of samples required) scales exponentially with the intrinsic dimensionality of the manifold, polynomially on the curvature and linearly on the intrinsic volume of the manifold (Narayanan and Mitter, 2010). Solving an ODE, needed to compute a single geodesic distance, scales linearly with the intrinsic dimensionality (Krämer et al., 2022).

Thus, as the intrinsic dimensionality of the manifold rises, the most challenging step is manifold learning. While undesirable, we note that performance for other tasks such as classification also depend exponentially on the manifold dimensionality (Narayanan and Niyogi, 2009). The triumph of modern deep learning to solve these tasks in a wide range of datasets provides strong implicit justification for the manifold hypothesis (Loaiza-Ganem et al., 2024).

How can model selection be used to choose from different manifold learning algorithms?

Model selection can be done if we have access to treatment effects for a small subset of datapoints. This is the case for our real-world experiments, as described in Section 5.1. Most often, ground truth ATE/ITE are not available, so further assumptions are needed. Model selection can then be done according to a different criterion, e.g., based on the quality of the learned representation, reconstruction loss, etc. In general, model selection for causal inference methods remains challenging and an active research area in the community.

Is ATE computed on a test set?

Yes, results are reported in a separate split. We separate the dataset randomly in two splits: the first one is to select hyperparameters, and the second split is to report final performance results.

How to avoid that the learned representation loses outcome information?

This is a very interesting remark, in fact we are currently working on a future extension of GeoMatching that incorporates outcome information, lying in-between regression methods that predict model outcomes and methods that learn a subspace related to outcomes. This is an emerging research area, it is non-trivial and out-of-scope of the current submission. GeoMatching falls in the bucket of methods that learn a latent representation solely based on pre-treatment covariates (vast majority of matching methods out there). Matching without looking at the outcome is desirable to avoid undesired biases and inadvertently induce selection bias, redirecting on the importance of separating the "design" and "analysis" phases of a non-randomized study (Rubin, 2008). This is a fundamental assumption within the potential outcome framework in causal inference (Pearl et al., 2016).

Not looking at the outcome incurs the risk of learning a latent space that is uninformative/irrelevant w.r.t treatment response, which would not be very useful for TE estimation. Losing valuable outcome information is a potential problem for any matching method that only relies on pre-treatment covariates, including Euclidean, Mahalanobis, and Propensity Score Matching.

Nothing in all these matching methods as well as GeoMatching explicitly prevents losing outcome information in the latent representation. Yet, all these methods are still useful because they rely on another fundamental, reasonable assumption: that patients similar in covariate space will tend to respond similarly, and that observed variability in covariate space is related/predictive of treatment response. Thus, by preserving information/structure present in the pre-treatment covariates, these methods indirectly preserve outcome information.

How to solve the ODE system for computation of geodesics?

We leverage the Stochman public library[9] for computation of geodesics. For each geodesic, we solve an ODE numerically using the standard Runge-Kutta method (Butcher, 2016). We highlight that GeoMatching can be implemented with any choice of ODE solver underneath, benefiting from any advancement in the research front of ODE solvers.

9. Stochman Library: https://github.com/MachineLearningLifeScience/stochman(https://github.com/MachineLearningLifeScience/stochman.

How does the number of neighbors used for matching influence TE estimation?

As the number of considered neighbors for matching increases, the TE estimation will incur in higher bias (towards average treatment response) but lower variance (less sensitivity to measurement noise in the outcome). These effects apply to any matching method, regardless of the considered distance (Stuart, 2010). In our work, we focus on the impact of improving the considered space (latent representation) and distance to perform matching.

If the data lies on a noisy manifold, are there methods to learn the denoised manifold? Can GeoMatching be applied in such situations?

All experiments in the paper involve a noisy manifold to some extent. For example, in the synthetic 3D-swissroll from Section 5.2, we make the manifold noisy by adding $(D-3)$ dimensions of pure Gaussian noise. In the rest of scenarios, real covariates are unlikely to lie on a perfect manifold, but rather on a noisy one. The proposed strategy of PCA projection + fitting a parameterized Riemannian metric is empirically effective to recover the intrinsic denoised manifold. An alternative is to model the manifold noise explicitly using a probabilistic latent variable model like a GPLVM to learn a stochastic Riemannian manifold, but this approach is computationally more expensive and was harder to train in our experience. Of course, one can investigate methods to reconstruct the manifold that has different levels of tolerance and plug them into our pipeline, but this is beyond the scope of this work.

What happens when the covariates are not lying on a lower dimensional manifold? Are there ways to detect it?

If the covariates are not lying on a lower dimensional manifold, GeoMatching is equivalent to standard matching with Euclidean distance and should perform similarly. There exist several methods to estimate the intrinsic dimensionality of a manifold Levina and Bickel (2004); Bac et al. (2021). Interestingly, such works have estimated that the intrinsic dimensionality of high-dimensional data is generally orders of magnitude smaller than their input dimension.

If knowledge of the causal graph among the covariates is provided, can this be used to inform the construction of the latent space? If yes, can one combine GeoMatching with some causal discovery algorithm?

Indeed, GeoMatching could be combined with causal discovery algorithms. In our introduction, we discuss how the causal mechanism across variables creates the manifold structure we use. Knowing this graph can help constrain manifold reconstruction. Instead of learning the entire manifold at once, one could learn all independent mechanisms separately and explicitly parameterize the manifold for computing geodesics, improving computational efficiency and accuracy. A causal discovery algorithm could integrate these distances with respect to the equivalence family of discovered graphs.

Is it possible to have an isometry between the latent manifold and the outcome space if the latent manifold has a dimension larger than one?

If the latent manifold has higher dimensionality than the output space, it is generally not possible to have an isometry. This is because reducing the dimensionality typically involves some loss of information, which means that the distances between points cannot be perfectly preserved. The latent

manifold can nonetheless live in any space of higher dimensionality, which gives some degree of freedom to our proposed procedure.

In the 3D-swissroll example in Section 5.2, the 1D latent manifold exists in a 2D space. Because there exists an isometry between the latent manifold and the outcome space, GeoMatching is optimal. In this example, finding the latent manifold is easy, and GeoMatching provides a mechanism to predict the output function by computing Riemannian distances.

In general, the goal is to approximate the isometry by preserving as much of the relevant geometric structure as possible. Our proposed approach leverages the intuition that projecting the original datapoints into a distance-preserving manifold of lower dimensionality removes the extra covariates that would make predicting the output function directly more difficult. Given the projected data, we then compute Riemannian distances, which generalizes other distances such as Euclidean or Mahalanobis, resulting in similar performance if the learned projection is "flat", and better performance if the learned projection is a curved manifold. With GeoMatching, a user does not need to worry about the shape of the manifold. We show this intuition holds in our other examples.