ADOPT: Additive Optimal Transport Regression

Wookyeong Song
University of California, Davis

Hans-Georg Müller
University of California, Davis

Abstract

Regression models for responses Y taking values in general metric spaces $(\mathcal{M}, d)$, with Euclidean predictors $X \in \mathbb{R}^p$, has attracted growing interest in recent years. While additive regression is a powerful tool for enhancing interpretability and mitigating the curse of dimensionality in the presence of multivariate predictors, its direct extension is hindered by the absence of vector space operations in general metric spaces. We propose a novel framework for additive optimal transport regression, which incorporates additive structure through optimal geodesic transports. A key idea is to extend the notion of optimal transports in Wasserstein spaces to general geodesic metric spaces. This unified approach accommodates a wide range of responses, including probability distributions, symmetric positive definite (SPD) matrices with various metrics and spherical data. The practical utility of the method is illustrated with correlation matrices derived from resting state fMRI brain imaging data.

1 INTRODUCTION

We study regression where the response is a random object Y in a geodesic space $(\mathcal{M}, d)$ with general metric d and the predictor is Euclidean $X \in \mathbb{R}^p$. Such settings arise widely for modern non-Euclidean data (Petersen and Müller, 2019; Schötz, 2022), including diffusion tensor imaging in the space of symmetric positive definite matrices (Lin et al., 2023), mortality analysis in the space of one-dimensional probability distributions with the Wasserstein geometry (Chen et al., 2023), functional brain networks data (Fornito et al., 2016; Zhou and Müller, 2022), corpus linguistics data on Riemannian manifolds (Severn et al., 2022), functional data (Wang et al., 2016) and many others (Marron and Dryden, 2021; Song et al., 2026).

Local Fréchet regression extends kernel-based local linear regression to metric-valued responses (Petersen and Müller, 2019; Chen and Müller, 2022; Schötz, 2022). However, like standard nonparametric regression, it suffers from the curse of dimensionality when $p > 1$, which limits its practical use for multivariate predictors. Dimension reduction via single-index models has been proposed to address this issue (Bhattacharjee et al., 2025; Bhattacharjee and Müller, 2023; Ghosal et al., 2023; Hong et al., 2025; Zhang et al., 2024), but compressing predictor information into a single index may be too restrictive in some cases and then lead to biases. For the case of Euclidean responses, additive models have been shown to provide a competitive alternative that mitigates the curse of dimensionality while offering interpretability and flexibility (Linton and Nielsen, 1995; Yu et al., 2008).

The generalization of additive models for metric space-valued responses is challenging because vector space operations, such as addition or scalar multiplication are unavailable in general metric spaces. Existing work covers additive structures for non-Euclidean responses only in some specialized settings, such as Hilbert space-valued responses (Jeon and Park, 2020), distributional responses via extrinsic approaches (Han et al., 2020) and Lie group–valued data (Lin et al., 2023). We introduce a novel additive framework based on geodesic optimal transports that generalize optimal transports in Wasserstein spaces to scenarios that feature general geodesic metric spaces. To estimate the additive transport map, we develop a transport backfitting algorithm, which is an extension of the popular classical backfitting method (Breiman and Friedman, 1985; Hastie and Tibshirani, 1986).

Outline of the paper: In Section 2 and 3 we review related works and preliminaries regarding Fréchet means, local Fréchet regression and geodesic optimal transport. Section 4 presents the proposed additive optimal transport regression (ADOPT) model and as-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

sociated estimation procedure using a transport backfitting algorithm and theory on oracle convergence. Section 5 reports numerical experiments and Section 6 real data illustrations for fMRI correlation matrices.

2 RELATED WORK

Prevalence of non-Euclidean data in recent years has created demand for principled and interpretable statistical tools for their analysis. A promising approach is to treat the sample elements as points in a metric space $(\mathcal{M}, d)$ (Dubey et al., 2024; Wang et al., 2024). Unlike Euclidean settings, general metric spaces lack algebraic structure, motivating distance-based statistical methods. The selected metric can greatly affect statistically relevant properties of the metric space (Song and Müller, 2026). For example for the case of symmetric positive definite (SPD) matrices the log-Cholesky metric (Lin, 2019) has desirable properties such as avoiding the swelling effect and allowing for closed-form parallel transport along geodesics.

The regression analysis of metric-valued responses has been studied in specialized scenarios (Yuan et al., 2012; Cornea et al., 2017). A general approach is Fréchet regression (Petersen and Müller, 2019), based on the concept of conditional Fréchet means. Fréchet regression makes use of the vector space structure of predictors in $\mathbb{R}^p$ to find elements in the metric space that minimize a weighted squared distances, and various extensions have been developed (Song and Han, 2023).

For distributional responses with the Wasserstein metric, geodesics are represented by optimal transport maps, enabling rich statistical applications (Chewi et al., 2024). Transport-based regression models have been developed for distribution-on-distribution regression (Ghodrati and Panaretos, 2022) and autoregressive models (Zhu and Müller, 2023). Zhu and Müller (2025) proposed a geodesic optimal transport (GOT) regression model, extending a parametric multiple regression framework via transport maps. However, this model treats the transport maps as fixed covariates and restricts estimation to finite-dimensional multiplicative parameters in $\mathbb{R}^p$. However nonparametric additive extensions that directly estimate transport maps considering the intrinsic geometry of the space have remained unexplored. We address this gap by proposing an additive optimal transport regression.

3 PRELIMINARIES

3.1 Fréchet mean and Fréchet regression

Let $(\mathcal{M}, d)$ be a metric space with distance $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. Consider random objects $(\mathbf{X}, Y) \sim \mathcal{F}$, where $\mathcal{F}$ is their joint distribution, $\mathbf{X} = (X_1, \ldots, X_p) \in \prod_{j=1}^p \mathcal{X}_j$ with compact domain $\mathcal{X}_j \subset \mathbb{R}$, and $Y \in \mathcal{M}$.

The Fréchet mean (Fréchet, 1948) generalizes the Euclidean expectation to random objects Y situated in metric spaces, with population and sample versions

$$\mathbb{E}_{\oplus} Y = \underset{\upsilon \in \mathcal{M}}{\arg\min}\, \mathbb{E} d^2(Y, \upsilon), \tag{1}$$

$$\hat{\mathbb{E}}_{\oplus} Y = \underset{\upsilon \in \mathcal{M}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^n d^2(Y_i, \upsilon). \tag{2}$$

Since $\mathcal{M}$ lacks vector space structure, Euclidean regression models cannot be directly applied. The Fréchet regression function is an implementation of conditional Fréchet means, extending conditional expectation to metric-valued responses:

$$\mathbb{E}_{\oplus}[Y \mid \mathbf{X} = \mathbf{x}] = \underset{\upsilon \in \mathcal{M}}{\arg\min}\, \mathbb{E}\{d^2(Y, \upsilon) \mid \mathbf{X} = \mathbf{x}\}.$$

For scalar responses $Y \in \mathbb{R}$, Fréchet conditional means reduce to the standard conditional expectation.

Local Fréchet regression (Petersen and Müller, 2019) adapts local linear regression by replacing Euclidean distances with d and using kernel-based weights. We consider the univariate predictor $X \in \mathbb{R}$ with kernel K, bandwidth $h > 0$, and rescaled kernel $K_h(\cdot) = h^{-1}K(\cdot/h)$, where the extension to $X \in \mathbb{R}^p$ with $p > 1$ is straightforward. This leads to

$$\mathbb{E}_{\oplus L}[Y \mid X = x] = \underset{\upsilon \in \mathcal{M}}{\arg\min}\, \mathbb{E}\{w(x, h) d^2(Y, \upsilon)\},$$

with $w(x, h) = K_h(X - x)\{u_2 - u_1(X - x)\}/\sigma_0^2$, $u_j = \mathbb{E}\{K_h(X-x)(X-x)^j\}$, $j = 0, 1, 2$ and $\sigma_0^2 = u_0 u_2 - u_1^2$.

Given random samples $(X_i, Y_i) \in \mathbb{R} \times \mathcal{M}$, $i = 1, 2, \ldots, n$, the corresponding sample estimator is

$$\hat{\mathbb{E}}_{\oplus L}[Y \mid X = x] = \underset{\upsilon \in \mathcal{M}}{\arg\min}\, \frac{1}{n} \sum_{i=1}^n \hat{w}_i(x, h) d^2(Y_i, \upsilon),$$
$$\text{where } \hat{w}_i(x, h) = K_h(X_i - x)\{\hat{u}_2 - \hat{u}_1(X_i - x)\}/\hat{\sigma}_0^2, \tag{3}$$

with $\hat{u}_j = n^{-1} \sum_{i=1}^n K_h(X_i - x)(X_i - x)^j$ for $j = 0, 1, 2$ and $\hat{\sigma}_0^2 = \hat{u}_0 \hat{u}_2 - \hat{u}_1^2$. For various asymptotic results on the convergence of this estimator see Petersen and Müller (2019); Chen and Müller (2022).

3.2 Geodesic optimal transport

For a metric space $(\mathcal{M}, d)$, the length of a path $\gamma : [0, 1] \to \mathcal{M}$ is defined as

$$l(\gamma) = \sup_{0 = t_0 < t_1 < \ldots < t_n = T, n \in \mathbb{N}} \sum_{i=1}^n d(\gamma(t_{i-1}), \gamma(t_i)).$$

A path $\gamma_{\upsilon_1,\upsilon_2}$ is a geodesic from υ_1 to υ_2 if $\gamma_{\upsilon_1,\upsilon_2}(0) = \upsilon_1$, $\gamma_{\upsilon_1,\upsilon_2}(1) = \upsilon_2$ and $d\left(\gamma_{\upsilon_1,\upsilon_2}(t), \gamma_{\upsilon_1,\upsilon_2}(t')\right) \propto |t - t'|$ for all $t, t' \in [0,1]$. A metric space $(\mathcal{M}, d)$ is called a geodesic space if every pair of points $\upsilon_1, \upsilon_2 \in \mathcal{M}$ can be joined by a geodesic, and a unique geodesic space if this geodesic is unique (Burago et al., 2022; Lang, 2012). In Euclidean space, geodesics are straight lines that connect two points. Throughout we assume that $(\mathcal{M}, d)$ is a bounded, separable, and uniquely geodesic space. For geodesic optimal transports, we require the following ubiquity assumption (Zhu and Müller, 2025):

Assumption 1 For any random objects $\upsilon_1, \upsilon_2, \upsilon_3 \in \mathcal{M}$, where $\upsilon_1 \neq \upsilon_2$, there exists a map $\Gamma : \mathcal{M} \times \mathcal{M} \times \mathcal{M} \to \mathcal{M}$ and a unique element $\upsilon_4 \in \mathcal{M}$ such that

$$\Gamma(\upsilon_1, \gamma_{\upsilon_1,\upsilon_2}(r), \upsilon_3) = \gamma_{\upsilon_3,\upsilon_4}(r),$$

for any $r \in [0,1]$.

This ubiquity assumption ensures that for any geodesic $\gamma_{\upsilon_1,\upsilon_2}$ and any point $\upsilon_3 \in \mathcal{M}$, there exists a unique geodesic starting from υ_3 whose direction matches that of $\gamma_{\upsilon_1,\upsilon_2}$. In other words, any geodesic segment can be attached at any point in $\mathcal{M}$, producing a new geodesic with endpoint $\upsilon_4 \in \mathcal{M}$. In Euclidean space $\mathbb{R}^p$, the map is given by $\Gamma(\upsilon_1, \gamma_{\upsilon_1,\upsilon_2}(r), \upsilon_3) = \upsilon_3 + r(\upsilon_2 - \upsilon_1)$, for $r \in [0,1]$, and $\upsilon_4 = \upsilon_3 + (\upsilon_2 - \upsilon_1)$.

Definition 1 (Geodesic optimal transport) *Under Assumption 1, for any random objects $\upsilon_1, \upsilon_2 \in \mathcal{M}$, the geodesic transport and inverse geodesic transport from υ_1 to υ_2, denoted as $T_{\upsilon_1,\upsilon_2} : \mathcal{M} \to \mathcal{M}$, $T^{-1}_{\upsilon_1,\upsilon_2} : \mathcal{M} \to \mathcal{M}$, respectively, are defined as*

$$T_{\upsilon_1,\upsilon_2}(\upsilon) = \Gamma(\upsilon_1, \upsilon_2, \upsilon), \;\; T^{-1}_{\upsilon_1,\upsilon_2}(\upsilon) = \Gamma(\upsilon_2, \upsilon_1, \upsilon),$$

for all $\upsilon \in \mathcal{M}$. Furthermore, the set of all geodesic transports maps is $\mathcal{T} = \{T_{\upsilon_1,\upsilon_2} : \upsilon_1, \upsilon_2 \in \mathcal{M}\}$.

We introduce an addition operation in the transport space $\mathcal{T}$ as function composition:

$$T_{\upsilon_1,\upsilon_2} \oplus T_{\upsilon_3,\upsilon_4} = T_{\upsilon_3,\upsilon_4} \circ T_{\upsilon_1,\upsilon_2},$$

for any $T_{\upsilon_1,\upsilon_2}, T_{\upsilon_3,\upsilon_4} \in \mathcal{T}$ (Zhu and Müller, 2023). For intuition, consider the Euclidean case $\mathcal{M} = \mathbb{R}^p$. For any $\upsilon \in \mathbb{R}^p$,

$$\begin{aligned}&\left[T_{\upsilon_1,\upsilon_2} \oplus T_{\upsilon_3,\upsilon_4}\right](\upsilon) = T_{\upsilon_3,\upsilon_4}(T_{\upsilon_1,\upsilon_2}(\upsilon)) \\ &= T_{\upsilon_3,\upsilon_4}\left(\upsilon + (\upsilon_2 - \upsilon_1)\right) = \upsilon + (\upsilon_2 - \upsilon_1) + (\upsilon_4 - \upsilon_3),\end{aligned}$$

so composing transports with different start and end points simply adds the corresponding displacement vectors. An analogous principle holds in Hilbert spaces and Riemannian manifolds $\mathcal{M}$ through parallel transport along geodesics.

Example 1 (Space of distributional data with the Wasserstein metric)

Let $(\mathcal{W}_2(\mathcal{D}), d)$ be the space of one-dimensional probability distributions on a compact domain $\mathcal{D} \subset \mathbb{R}$ with finite second moments with 2-Wasserstein metric d,

$$d^2(\upsilon_1, \upsilon_2) = \int_0^1 \left\{F_1^{-1}(s) - F_2^{-1}(s)\right\}^2 ds,$$

where $F_1^{-1}(\cdot)$ and $F_2^{-1}(\cdot)$ are the quantile functions of υ_1 and υ_2, respectively.

Let $\upsilon \in \mathcal{W}_2(\mathcal{D})$ be a random element of $(\mathcal{W}_2(\mathcal{D}), d)$ For any measurable function $l : \mathcal{D} \to \mathcal{D}$ and $\upsilon \in \mathcal{W}_2(\mathcal{D})$, let $l\#\upsilon$ denote the push-forward measure of υ, satisfying $l\#\upsilon(A) = \upsilon\left(\{x : l(x) \in A\}\right)$ for all $A \in \mathcal{B}(\mathcal{D})$, the Borel σ-algebra on $\mathcal{D}$. For any two distributions $\upsilon_1, \upsilon_2 \in \mathcal{W}_2(\mathcal{D})$, the optimal transport map is

$$OT_{\upsilon_1,\upsilon_2} = F_2^{-1} \circ F_1.$$

The geodesic $\gamma_{\upsilon_1,\upsilon_2}$ is given by McCann's interpolant (McCann, 1997),

$$\gamma_{\upsilon_1,\upsilon_2}(r) = (\mathrm{id} + r(OT_{\upsilon_1,\upsilon_2} - \mathrm{id}))\#\upsilon_1, \quad r \in [0,1],$$

where $\mathrm{id} : \mathcal{D} \to \mathcal{D}$ is the identity map. For notational convenience, we simply write $\mathcal{W}_2(\mathcal{D})$ as $\mathcal{W}_2$.

For any random objects, $\upsilon_1, \upsilon_2, \upsilon_3 \in \mathcal{W}_2$, where $\upsilon_1 \neq \upsilon_2$, assumption 1 is satisfied with

$$\begin{aligned}\Gamma(\upsilon_1, \gamma_{\upsilon_1,\upsilon_2}(r), \upsilon_3) &= (\mathrm{id} + r(OT_{\upsilon_1,\upsilon_2} - \mathrm{id}))\#\upsilon_3 \\ &= \gamma_{\upsilon_3,\upsilon_4}(r), \quad r \in [0,1],\end{aligned}$$

where $\upsilon_4 = T_{\upsilon_1,\upsilon_2}(\upsilon_3) = OT_{\upsilon_1,\upsilon_2}\#\upsilon_3$.

Example 2 (SPD matrices with the log-Cholesky metric)

Let $\mathcal{S}_m^+$ be the collection of $m \times m$ SPD matrics. For any $S \in \mathcal{S}_m^+$, there exists a unique lower triangular matrix L from the Cholesky decomposition, such that $S = LL^{\mathrm{T}}$, where L is composed of $\lfloor L \rfloor$, the strictly lower triangular part, and $D(L)$, the diagonal part. The log-Cholesky metric is defined as

$$\begin{aligned}&d^2(S_1, S_2) \\ &= \|\lfloor L_1 \rfloor - \lfloor L_2 \rfloor\|^2 + \|\log(D(L_1)) - \log(D(L_2))\|^2,\end{aligned}$$

where $S_1 = L_1 L_1^{\mathrm{T}}$, $S_2 = L_2 L_2^{\mathrm{T}}$, and $\|\cdot\|$ is the Frobenius norm. The geodesic γ_{S_1,S_2} is given by

$$\begin{aligned}&\gamma_{S_1,S_2}(r) = L(r)L(r)^{\mathrm{T}}, \\ &L(r) = \lfloor L_1 \rfloor + r(\lfloor L_2 \rfloor - \lfloor L_1 \rfloor) \\ &\quad + \exp\left[\log(D(L_1)) + r(\log(D(L_2)) - \log(D(L_1)))\right],\end{aligned}$$

for any $r \in [0,1]$. For any random objects, $S_1, S_2, S_3 \in \mathcal{S}_m^+$, where $S_1 \neq S_2$ and $S_3 = L_3 L_3^{\mathrm{T}}$, Assumption 1 is

satisfied for all $r \in [0,1]$ with

$$\Gamma(S_1, \gamma_{S_1,S_2}(r), S_3) = L_\Gamma(r) L_\Gamma(r)^{\mathrm{T}} = \gamma_{S_3,S_4}(r),$$
$$L_\Gamma(r) = \lfloor L_3 \rfloor + r(\lfloor L_2 \rfloor - \lfloor L_1 \rfloor) + \exp\left[\log(D(L_3)) + r(\log(D(L_2)) - \log(D(L_1)))\right],$$

$S_3 = L_3 L_3^{\mathrm{T}}$, $S_4 = L_\Gamma(1) L_\Gamma(1)^{\mathrm{T}}$.

Geodesic transports satisfying Assumption 1 can be constructed for several metrics on $\mathcal{S}_m^+$ including the Frobenius and power Frobenius metric and also for the space of graph Laplacians representing networks (Zhu and Müller, 2025).

4 ADDITIVE OPTIMAL TRANSPORT REGRESSION

4.1 Population Model

We propose an additive optimal transport regression model that allows to regress Y on the high dimensional predictors $\mathbf{X}$. We first consider a scalar response $Y \in \mathbb{R}$ with predictors $\mathbf{X} = (X_1, \ldots, X_p)$. The classical additive regression model is the conditional mean of response Y given $\mathbf{X} = \mathbf{x}$, defined as

$$m(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] = \beta_0 + g_1(x_1) + \cdots + g_p(x_p), \tag{4}$$

with the assumptions that $\mathbb{E}Y = \beta_0$, and $\mathbb{E}g_j(X_j) = 0$, $j = 1, \ldots, p$, for identifiability. Then, the classical additive regression model (4) is equivalent to

$$m(\mathbf{x}) = (g_1(x_1) - \mathbb{E}g_1(X_1)) + \cdots + (g_p(x_p) - \mathbb{E}g_p(X_p)) + \mathbb{E}Y. \tag{5}$$

In the case of scalar response $Y \in \mathbb{R}$ in (4), the jth additive term $g_j(X_j)$ is defined through conditional expectation of Y given $X_j = x_j$ is as follows:

$$m_j(x_j) = \mathbb{E}\left[Y \mid X_j = x_j\right] = \mathbb{E}\left[\mathbb{E}\left[Y \mid \mathbf{X}\right] | X_j = x_j\right] = \mathbb{E}Y + g_j(x_j) + \mathbb{E}\left[\sum_{k \neq j} g_k(X_k) \;\middle|\; X_j = x_j\right].$$

Then, the additive term $g_j(X_j)$ are given by

$$g_j(x_j) = \mathbb{E}\left[P_j | X_j = x_j\right],$$

where for $j = 1, \ldots, p$ the jth partial residual is

$$P_j = (Y - \mathbb{E}Y) - \sum_{k \neq j} \left(g_k(X_k) - \mathbb{E}g_k(X_k)\right). \tag{6}$$

A direct extension of model (5) may not be feasible for responses situated in general metric spaces $\mathcal{M}$, since basic vector space operations, such as addition, scalar multiplication and expectation may not exist. We generalize the expectation $\mathbb{E}$ to the Fréchet mean $\mathbb{E}_\oplus$ in (1). For convenience we reuse the notation for the jth additive term g_j,

$$g_j(x_j) \in \mathcal{M}, \quad x_j \in \mathcal{X}_j, \quad j = 1, \ldots, p.$$

The difference between the jth additive term and its Fréchet mean is replaced by the geodesic transport pushing $\mathbb{E}_\oplus g_j(X_j)$ to $g_j(x_j)$, denoted as $T_j(x_j) := T_{\mathbb{E}_\oplus g_j(X_j), g_j(x_j)} \in \mathcal{T}$. Analogously, we introduce the jth inverse geodesic transport $T_j^{-1}(x_j) := T^{-1}_{\mathbb{E}_\oplus g_j(X_j), g_j(x_j)} \in \mathcal{T}$.

We further note that the actual responses in standard models (4) are $Y = m(\mathbf{x}) + \varepsilon$, however in metric spaces additive errors ε are not feasible. Contamination of responses by errors instead can be obtained through a random perturbation map $\mathcal{M} \to \mathcal{M}$ defined as $v' = \varepsilon(v)$, such that $\mathbb{E}_\oplus(v') = v$ and $\sigma^2 = \mathbb{E}d^2(v, v') > 0$ (Chen and Müller, 2022). Assuming that the perturbation map is independent of predictors $\mathbf{X}$, this leads to the proposed ADOPT model

$$Y = \varepsilon(m_\oplus(\mathbf{x})) = \left[\bigoplus_{j=1}^{p} T_j(x_j) \oplus \varepsilon\right](\mu_\oplus), \tag{7}$$

where $\mu_\oplus = \mathbb{E}_\oplus Y$ and $\bigoplus_{j=1}^p T_j = T_1 \oplus T_2 \oplus \cdots T_p$.

Generalizing partial residuals in the conditional expectation, the jth partial transport residual $P_{\oplus j} \in \mathcal{T}$, $j = 1, \ldots, p$, in the ADOPT model (7) is defined as

$$P_{\oplus j} = \left[T_{j-1}^{-1}(X_{j-1}) \oplus \cdots \oplus T_1^{-1}(X_1)\right] \oplus T_{\mu_\oplus, Y} \oplus \left[T_p^{-1}(X_p) \cdots \oplus T_{j+1}^{-1}(X_{j+1})\right]. \tag{8}$$

The left panel of Figure 1 shows the framework of the additive optimal transport regression and the right panel the partial transport residual $P_{\oplus 2}$ for a 3-dimensional predictor $\mathbf{X} \in \mathbb{R}^3$.

4.2 Estimation

Given i.i.d. pairs $(\mathbf{X}_i, Y_i) \in \Pi_{j=1}^p \mathcal{X}_j \times \mathcal{M}$, where $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})$, the sample ADOPT model is

$$Y_i = \left[\bigoplus_{j=1}^{p} T_j(X_{i,j}) \oplus \varepsilon_i\right](\mu_\oplus), \quad i = 1, \ldots, n, \tag{9}$$

where we substitute the sample Fréchet mean (2) for the Fréchet mean $\mu_\oplus$, To fit model (9), we apply standard backfitting (Breiman and Friedman, 1985; Hastie and Tibshirani, 1986), an iterative algorithm

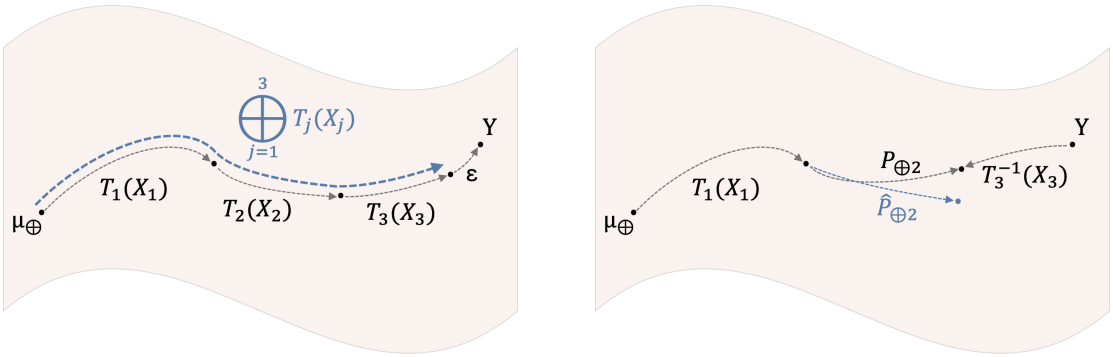

Figure 1: (Left panel) Additive optimal transport regression (ADOPT) framework with 3-dimensional predictors $X \in \mathbb{R}^3$ in (7). (Right panel) Population partial transport residual $P_{\oplus j}$ in (8) and its estimator $\hat{P}_{\oplus j}$ in (10) with $j = 2$.

with steps $t = 1, 2, \ldots$, alternating between 1) a backfitting step, where the partial residual P_j in (6) is smoothed with respect to X_j, i.e., $g_j(x_j) \leftarrow \mathbb{E}(P_j \mid X = x_j)$, and 2) mean centering for identifiability, i.e., $g_j(x_j) \leftarrow g_j(x_j) - \mathbb{E}g_j(x_j)$, for all $j = 1, \ldots, p$, until convergence. This alternating procedure is applied separately for each predictor, cycling through the predictor set. We initialize with the Fréchet mean estimate $\hat{\mu}_\oplus$ and the identity maps $\hat{T}_j^{(0)}(\cdot) = \mathrm{id}(\cdot) \in \mathcal{T}$, for $j = 1, 2, \ldots, p$.

At each iteration $t = 1, 2, \ldots$, we construct partial transport residuals

$$\hat{P}_{\oplus j}^{(t)} = \left[\left(\hat{T}_{j-1}^{(t)}\right)^{-1}(X_{j-1}) \oplus \cdots \oplus \left(\hat{T}_1^{(t)}\right)^{-1}(X_1)\right] \oplus T_{\hat{\mu}_\oplus, Y} \oplus \left[\left(\hat{T}_p^{(t-1)}\right)^{-1}(X_p) \cdots \oplus \left(\hat{T}_{j+1}^{(t-1)}\right)^{-1}(X_{j+1})\right], \tag{10}$$

with sample observations $\hat{P}_{\oplus i,j}^{(t)}$ obtained by replacing $(X_1, X_2, \ldots, X_p)$ with $(X_{i,1}, X_{i,2}, \ldots, X_{i,p})$, $i = 1, 2, \ldots, n$. The right panel of Figure 1 illustrates the estimated partial transport residuals.

The transport backfitting step utilizes local Fréchet regression where the responses are the partial transport residuals, given each univariate predictor $X_j = x_j$,

$$g_j^{(t)}(x_j) = \hat{\mathbb{E}}_{\oplus L}\left[\hat{P}_{\oplus j}^{(t)}(\hat{\mu}_\oplus) \mid X_j = x_j\right] = \underset{v \in \mathcal{M}}{\arg\min} \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j, h) d^2\left(\hat{P}_{\oplus i,j}^{(t)}(\hat{\mu}_\oplus), v\right).$$

The weights $\hat{w}_i$ are defined in (3) and the bandwidth h is selected by using 5-fold cross validation.

The mean centering step is generalized to a transport centering step,

$$\hat{T}_j^{(t)}(\cdot) = T_{\hat{\mathbb{E}}_\oplus g_j^{(t)}(X_j),\, g_j^{(t)}(\cdot)} \in \mathcal{T},$$

with sample Fréchet mean

$$\hat{\mathbb{E}}_\oplus g_j^{(t)}(X_j) = \underset{v \in \mathcal{M}}{\arg\min} \frac{1}{n}\sum_{i=1}^n d^2(g_j^{(t)}(X_{i,j}), v). \tag{11}$$

At iteration t, we apply the alternating steps cycling through predictors $X_j = x_j$, $j = 1, \ldots, p$ and subsequently update the fitted responses

$$\hat{Y}_i^{(t)} = \left[\bigoplus_{j=1}^p \hat{T}_j^{(t)}(X_{i,j})\right](\hat{\mu}_\oplus), \quad i = 1, 2, \ldots, n,$$

iterating until convergence when

$$\frac{1}{n}\sum_{i=1}^n d^2(\hat{Y}_i^{(t)}, \hat{Y}_i^{(t-1)}) < \epsilon,$$

for prespecified tolerance $\epsilon > 0$. The workflow is summarized in Algorithm 1.

Denote by $\hat{T}_j(\cdot)$ the estimates from Algorithm 1 when the other transports $T_1(\cdot)$, $\ldots$, $T_{j-1}(\cdot)$, $T_{j+1}(\cdot), \ldots$, $T_p(\cdot)$ are assumed to be known. Then one can obtain the following oracle convergence under mild assumptions (see the Supplement Section A for the proof and detailed assumptions).

Theorem 1 *Under Assumption 1 and the assumptions in Supplement Section A, if the bandwidth $h \to 0$ and $nh \to \infty$ as $n \to \infty$, it holds that*

$$d\left(T_j(x_j)(\nu), \hat{T}_j(x_j)(\nu)\right) = o_P(1),$$

for any $x_j \in \mathcal{X}_j$, and $\nu \in \mathcal{M}$.

5 SIMULATIONS

5.1 Distribution-valued responses with Wasserstein metric

We consider univariate distributional responses Y situated in the 2-Wasserstein space $\mathcal{W}_2$ with Euclidean

Algorithm 1 Transport Backfitting Algorithm

Require: $(X_i, Y_i) \in \mathbb{R}^p \times \mathcal{M}$, $i = 1, 2, \ldots, n$, and convergence tolerance $\epsilon > 0$.

Ensure: $\hat{T}_j(\cdot) \in \mathcal{T}$, $j = 1, 2, \ldots, p$.

1: Initialize $\hat{T}_j^{(0)}(\cdot) = \mathrm{id}$, $j = 1, \ldots, p$, and $\hat{Y}_i^{(0)} = \left[\bigoplus_{j=1}^{p} \hat{T}_j^{(0)}(X_{i,j})\right](\hat{\mu}_\oplus)$, $i = 1, 2, \ldots, n$, with sample Fréchet mean $\hat{\mu}_\oplus$.

2: **for** $t = 1, 2, \ldots$ until convergence **do**

3: **for** $j = 1, \ldots, p$ **do**

4: Transport backfitting via local Fréchet regression $\hat{\mathbb{E}}_{\oplus L}$ in (3),

$$g_j^{(t)}(x_j) = \hat{\mathbb{E}}_{\oplus L}\left[\hat{P}_{\oplus j}^{(t)}(\hat{\mu}_\oplus) \mid X_j = x_j\right],$$

with partial transport residual observations $\hat{P}_{\oplus j}^{(t)}$ in (10).

5: Transport Centering,

$$\hat{T}_j^{(t)}(\cdot) = T_{\hat{\mathbb{E}}_\oplus g_j^{(t)}(X_j),\, g_j^{(t)}(\cdot)} \in \mathcal{T},$$

with Fréchet mean $\hat{\mathbb{E}}_\oplus g_j^{(t)}(X_j)$ in (11).

6: **end for**

7: $\hat{Y}_i^{(t)} = \left[\bigoplus_{j=1}^{p} \hat{T}_j^{(t)}(X_{i,j})\right](\hat{\mu}_\oplus)$, $i = 1, 2, \ldots, n$.

8: Stopping rule: Stop if

$$\frac{1}{n}\sum_{i=1}^{n} d^2(\hat{Y}_i^{(t)}, \hat{Y}_i^{(t-1)}) < \epsilon.$$

9: **end for**

predictor vectors $\mathbf{X} \in [0,1]^3$, introduced in Example 1. In a slight abuse of notation, we use Y to denote the quantile function corresponding to a distribution. We generate distributional responses from the additive transport regression model in (9),

$$Y_i = \left[\bigoplus_{j=1}^{3} T_j(X_{i,j}) \oplus \varepsilon_i\right](U),$$

where $U = \mathrm{Unif}(0,1)$, $T_j(x_j) = T_{\mathbb{E}_\oplus g_j(X_j), g_j(x_j)}$, $x_j \in [0,1]$, $j = 1, 2, 3$, and $\varepsilon_i(f)(u) = f(u) + \frac{1}{2\pi}\xi_i \sin(2\pi f(u))$, for all quantile functions $f \in \mathcal{W}_2$, and $u \in [0,1]$, $\xi_i \sim \mathrm{Unif}(-1,1)$. We consider the following two cases:

- Case I (Beta distributions): Here $g_1(x_1) = \mathrm{Beta}(1 + 2x_1, 1)$, $g_2(x_2) = \mathrm{Beta}(1, 2 + 3x_2)$, and $g_3(x_3) = \mathrm{Beta}(\frac{1}{2} + \frac{1}{2}x_3, \frac{1}{2} + \frac{1}{2}x_3)$.
- Case II (Normal distributions): Here $g_1(x_1) = N(x_1, 1)$, $g_2(x_2) = N(x_2^2, 1)$, and $g_3(x_3) = N(e^{-x_3}, 1)$.

We generate 3-dimensional Euclidean predictors $\mathbf{X}_i = (X_{i,1}, X_{i,2}, X_{i,3}) = (\Phi(V_{i,1}), \Phi(V_{i,2}), \Phi(V_{i,3}))^T$, $i = 1, \ldots, n$, where Φ is the standard normal CDF and $\mathbf{V}_i = (V_{i,1}, V_{i,2}, V_{i,3})^T \sim N_3(\mathbf{0}, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.5 \\ 0.3 & 0.5 & 1 \end{pmatrix}$. For each case, we generate $n = 100$, and 300 pairs of $(\mathbf{X}_i, \mathbf{Y}_i)$, $i = 1, 2, \ldots, n$. We compare the finite-sample performance of the proposed ADOPT model with global Fréchet regression (GF) (Petersen and Müller, 2019), additive functional regression (Han et al., 2020) (ADR) and simply taking the constant Fréchet mean (FM), in terms of the apporximated mean integrated squared error (MISE),

$$\mathrm{MISE} \approx B^{-1}\sum_{b=1}^{B}\int_{[0,1]^3} d^2(\hat{Y}_{\mathbf{x}}^{(b)}, Y_{\mathbf{x}})d\mathbf{x}, \tag{12}$$

where d is 2-Wasserstein distance, $\hat{Y}_{\mathbf{x}}^{(b)}$ with $\mathbf{x} = (x_1, x_2, x_3)$ are estimates from the b^{th} Monte Carlo (MC) samples $\left\{(\mathbf{X}_i^{(b)}, \mathbf{Y}_i^{(b)}) : 1 \le i \le n\right\}$, and $Y_{\mathbf{x}}$ are the true distributional data without noise, $Y_{\mathbf{x}} = \left[\bigoplus_{j=1}^{p} T_j(x_j)\right](U)$. The results are in Table 1 and demonstrate that ADOPT has the smallest MISE among the comparison methods and as n increases the MISE of ADOPT decreases, while this is not the case for the competing methods.

Table 1: Mean integrated squared error (MISE) results ($\times 10^{-3}$) with standard errors in parentheses under distributional responses in the Wasserstein space for two cases with different sample sizes $n = 100$, and 300. We compare four methods: additive optimal transport regression (ADOPT), global Fréchet regression (Petersen and Müller, 2019) (GF), additive functional regression (Han et al., 2020) (ADR), and Fréchet mean (Fréchet, 1948) (FM) without predictor effect.

		Method			
Case	n	ADOPT	GF	ADR	FM
I	100	**0.317** (0.010)	0.591 (0.008)	0.761 (0.008)	11.019 (0.068)
	300	**0.181** (0.004)	0.499 (0.003)	0.586 (0.004)	11.119 (0.040)
II	100	**1.685** (0.057)	6.636 (0.057)	80.411 (0.702)	207.205 (1.681)
	300	**1.121** (0.041)	6.701 (0.037)	81.951 (0.455)	211.297 (1.080)

5.2 SPD matrices-valued response with log-Cholesky metric

We consider $m \times m$ SPD matrix valued responses $Y \in \mathcal{S}_m^+$ equipped with the log-Cholesky metric (Lin, 2019) introduced in Example 2, and Euclidean predictors $\mathbf{X} \in [0,1]^3$ and generate SPD responses from the additive transport regression model in (9),

$$Y_i = \left[\bigoplus_{j=1}^{3} T_j(X_{i,j}) \oplus \varepsilon_i \right] (\mathbf{I}_m) \in \mathcal{S}_m^+,$$

where $\mathbf{I}_m$ is $m \times m$ identity matrix, $T_j(x_j) = T_{\mathbb{E}_\oplus g_j(X_j), g_j(x_j)}$, $j = 1, \ldots, 3$, and for $S = LL^{\mathrm{T}} \in \mathcal{S}_m^+$ with Cholesky decomposition, the perturbation error map $\varepsilon_i(S) = \varepsilon_i(LL^{\mathrm{T}}) = e_i(L)e_i(L)^{\mathrm{T}}$, where $e_i(L)_{k,l} = L_{k,l} + \xi_{k,l}$, and $\xi_{k,l} \sim N(0, \sigma^2)$, $k > l$, $e_i(L)_{k,l} = L_{k,l}$, when $k = l$, and $e_i(L)_{k,l} = 0$ if $k < l$.

We generate the transport map $T_j(x_j)$ based on

$$g_j(x_j) = L_j(x_j)L_j(x_j)^{\mathrm{T}}, \quad x_j \in [0,1], \quad j = 1, 2, 3,$$

with random Cholesky factors $L_j(x_j)$ which are lower triangular matrices as follows:

- Case I (Linear in $\mathbf{x}$):

$$\begin{aligned} (L_1(x_1))_{k,l} &= \exp^{-|k-l|/2} (x_1 + 1)/2, \\ (L_2(x_2))_{k,l} &= \exp^{-|k-l|/3} (x_2 + 1)/4, \\ (L_3(x_3))_{k,l} &= \exp^{-|k-l|/4} (x_3 + 1)/8. \end{aligned}$$

- Case II (Non-linear in $\mathbf{x}$):

$$\begin{aligned} (L_1(x_1))_{k,l} &= \exp^{-|k-l|/2} \sin((x_1 + 1)\pi/4), \\ (L_2(x_2))_{k,l} &= \exp^{-|k-l|/4} \left(x_2^2 + 1\right)/2, \\ (L_3(x_3))_{k,l} &= \exp^{-|k-l|/4} \exp^{-x_3}, \end{aligned}$$

for $k \geq l$, and all other elements are 0 if $k < l$.

We then generate random covariates $\mathbf{X}_i = (X_{i,1}, X_{i,2}, X_{i,3})$, $i = 1, \ldots, n$, following the same procedure as in the distribution-valued simulations of Section 5.1. The responses are generated under each scenario with perturbations ε_i having standard deviation $\sigma = 0.01$. We sample $n = 100$ and 300 i.i.d. pairs of $(\mathbf{X}_i, Y_i) \in \mathbb{R}^3 \times \mathcal{S}_m^+$, with matrix sizes $m = 10$ and $m = 20$ ands compare the finite sample performance of the proposed ADOPT with the global Fréchet regression (GF) and the constant Fréchet mean (FM) in terms of mean integrated squared error (MISE), analogously defined as in (12). Table 2 displays the MISEs for $B = 200$ with standard error in parentheses. ADOPT has the lowest MISE and its MISE decreases as the sample size n increases, while this is not the case for the competing methods. As the size of SPD matrices increases, the MISE increases, but ADOPT always has substantially better performance.

Table 2: Mean integrated squared error (MISE) results ($\times 10^{-2}$) with standard errors in parentheses for $m \times m$ SPD matrices as responses in $\mathcal{S}_m^+$ with the log-Cholesky metric for two simulation cases with different sample sizes $n = 100$, and 300 and matrix dimensions $m = 10$, and 20, comparing three methods: additive optimal transport regression (ADOPT), global Fréchet regression (GF) and Fréchet mean (FM).

			Method		
Case	m	n	ADOPT	GF	FM
I	10	100	**2.769** (0.025)	9.495 (0.060)	135.344 (0.647)
		300	**1.655** (0.012)	9.714 (0.037)	136.122 (0.366)
	20	100	**5.712** (0.070)	14.110 (0.102)	192.688 (0.92)
		300	**3.788** (0.048)	14.254 (0.063)	193.796 (0.521)
II	10	100	**4.146** (0.088)	18.507 (0.090)	96.562 (0.490)
		300	**2.702** (0.036)	18.808 (0.052)	97.047 (0.303)
	20	100	**8.921** (0.167)	28.237 (0.14)	143.382 (0.706)
		300	**6.435** (0.100)	28.512 (0.078)	144.088 (0.437)

6 BRAIN CONNECTIVITY ANALYSIS

We use resting state functional Magnetic Resonance Imaging (rs-fMRI) data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). Brain signal analysis at the subject level relies on time series of Blood Oxygen Level Dependent (BOLD) signals, obtained for a set of regions of interest (ROIs). The coherence between pairwise ROIs is customarily quantified by Pearson correlation coefficients of the rs-fMRI time series, leading to a $l \times l$ correlation matrix for l ROIs (Badhwar et al., 2017; Perovnik et al., 2023). The BOLD signals for each subject were collected over a time interval from 0 to 270 seconds with $K = 136$ equispaced measurements taken every 2 seconds. We have 136×11 signal matrix S, and s_{kl} is the (k,l)th element of the

signal matrix S and an 11×11 connectivity matrix response Y was obtained using Pearson correlation,

$$(Y)_{l_1,l_2} = \frac{\sum_{k=1}^{K}(s_{kl_1} - \bar{s}_{l_1})(s_{kl_2} - \bar{s}_{l_2})}{\sqrt{\sum_{k=1}^{K}(s_{kl_1} - \bar{s}_{l_1})^2}\sqrt{\sum_{k=1}^{K}(s_{kl_2} - \bar{s}_{l_2})^2}},$$

with $\bar{s}_l = \frac{1}{K}\sum_{k=1}^{K} s_{kl}$, $l_1, l_2 = 1, 2, \ldots, 11$. For Alzheimer's disease (AD) studies, the CSF phosphorylated tau (p-Tau) concentration is another established biomarker, where elevated pTau levels are strongly associated with AD pathology (Karikari et al., 2022), with higher p-Tau typically indicating greater AD risk. Also, cerebrospinal fluid (CSF) amyloid-beta (Aβ) concentration is a widely used marker of amyloid pathology (Murphy and LeVine III, 2010; Hampel et al., 2021). Further details about ADNI data preprocessing, and the list of $l = 11$ ROIs used in the analysis (Andrews-Hanna et al., 2010) are described in Supplement Section B.1.

We analyze the 11×11 correlation matrices equipped with the log-Cholesky metric as responses using the ADOPT model in (9), based on $n = 929$ ADNI participants classified into six diagnostic stages: 248 cognitively normal (CN), 110 subjective memory complaint (SMC), 316 early mild cognitive impairment (EMCI), 12 mild cognitive impairment (MCI), and 171 late mild cognitive impairment (LMCI), and 72 Alzheimer's disease (AD).

We consider $p = 3$ predictors: X_1 is Aβ concentration ranging from 203 to 1700, with lower amyloid-beta signaling amyloid-burden; X_2 is the diagnostic stage coded ordinally from 0 to 5 in the order CN, SMC, EMCI, MCI, LMCI, AD; X_3 is the p-Tau concentration ranging from 8.00 to 92.08, with higher p-Tau typically indicating worse prognosis.

To compare brain coherence patterns across distinct risk profiles, we construct three covariate settings representing low, intermediate, and high risk groups. The low risk group corresponds to the upper 10% of Aβ values (1700), the lower 10% disease stage (CN = 0), and p-Tau values (12.8). The intermediate risk group is defined by median values of Aβ (995.8), disease stage (EMCI = 2), and p-Tau (21.94). The high risk group corresponds to the lower 10% of Aβ values (535), the upper 90% disease stage (LMCI = 4), and p-Tau values (42.01).

Figure 2 displays each transport map $T_j(x_j) = T_{\mathbb{E}_\oplus g(X_j), g(x_j)}$ as the difference of the Cholesky factors of the two correlation matrices, i.e., $L_{g(x_j)} - L_{\mathbb{E}_\oplus g(X_j)}$, where L_S is a lower triangular Cholesky factor of SPD matrices $S \in \mathcal{S}_l^+$, with ROIs labeling rows and columns of each Cholesky factor. The estimated transport maps $T_1(x_1)$ (first column), $T_2(x_2)$ (second column), $T_3(x_3)$ (third column), and their combined effect $[T_1(x_1) \oplus T_2(x_2) \oplus T_3(x_3)]$ (fourth column) are obtained from ADOPT. The rows correspond to the low, intermediate, and high risk groups, with covariates $\mathbf{x} = (x_1, x_2, x_3)$ defined in the previous paragraph. For the low risk group (first row), the transport maps generally indicate increasing correlations between ROIs, with the largest increase of 0.10 observed between pIPL and MPFC. For the intermediate group (second row), the transport maps are close to zero, implying that the Fréchet mean $\hat{\mu}_\oplus$ is similar to the estimated response $[T_1 \oplus T_2 \oplus T_3](\hat{\mu}_\oplus)$. In contrast, the high risk group (third row) shows decreasing correlations, with the strongest decline of 0.11 between pIPL and MPFC.

The additive structure of the model leads to enhanced interpretability, providing predictor-specific transport maps T_1, T_2, and T_3 corresponding to the individual effects of x_1 (Aβ), x_2 (disease stage), and x_3 (p-Tau), respectively. For the disease-stage transport map T_2, the strongest effect appears in the connection between the posterior cingulate cortex (PCC) and the posterior inferior parietal lobule (pIPL): approximately $+0.04$ in the low disease stage group versus -0.04 in the high disease stage group, consistent with well-known findings of reduced PCC–medial temporal lobe region (MTL, including pIPL) connectivity in Andrews-Hanna et al. (2010). For the p-tau transport map T_3, we find an overall decrease in connectivity in the high-risk group, consistent with literature showing that elevated p-tau is associated with broad reductions in functional connectivity (Luan et al., 2024).

We evaluate out-of-sample prediction performance using 5-fold cross-validation. Specifically, we partition the index set $1, \ldots, n$ into disjoint folds $\mathcal{I}_q$ ($q = 1, \ldots, 5$) with $n_q = |\mathcal{I}_q|$. For each fold, we define the Mean Prediction Error Reduction (MPER) as

$$\text{MPER} = \frac{1}{5}\sum_{q=1}^{5} \text{MPER}_q,$$

$$\text{MPER}_q = \frac{1}{n_q}\sum_{i \in \mathcal{I}_q} \left\{ d(\hat{\mu}_\oplus^{(-q)}, Y_i) - d(\hat{Y}_i^{(-q)}, Y_i) \right\},$$

where $\hat{Y}_i^{(-q)}$ is the predicted response from the regression model trained on the remaining folds, $\hat{\mu}_\oplus^{(-q)}$ is the Fréchet mean of the training responses (baseline without predictors) and d is the log-Cholesky metric. MPER measures the reduction in prediction error relative to the baseline; larger values indicate better predictive performance. We compare our method with global Fréchet regression (GF) (Petersen and Müller, 2019). Table 3 reports the average of MPER ($\times 10^{-4}$) for $B = 200$ Monte Carlo simulations for ADOPT and global Fréchet regression, demonstrating that ADOPT improves out-of-sample performance substantially in

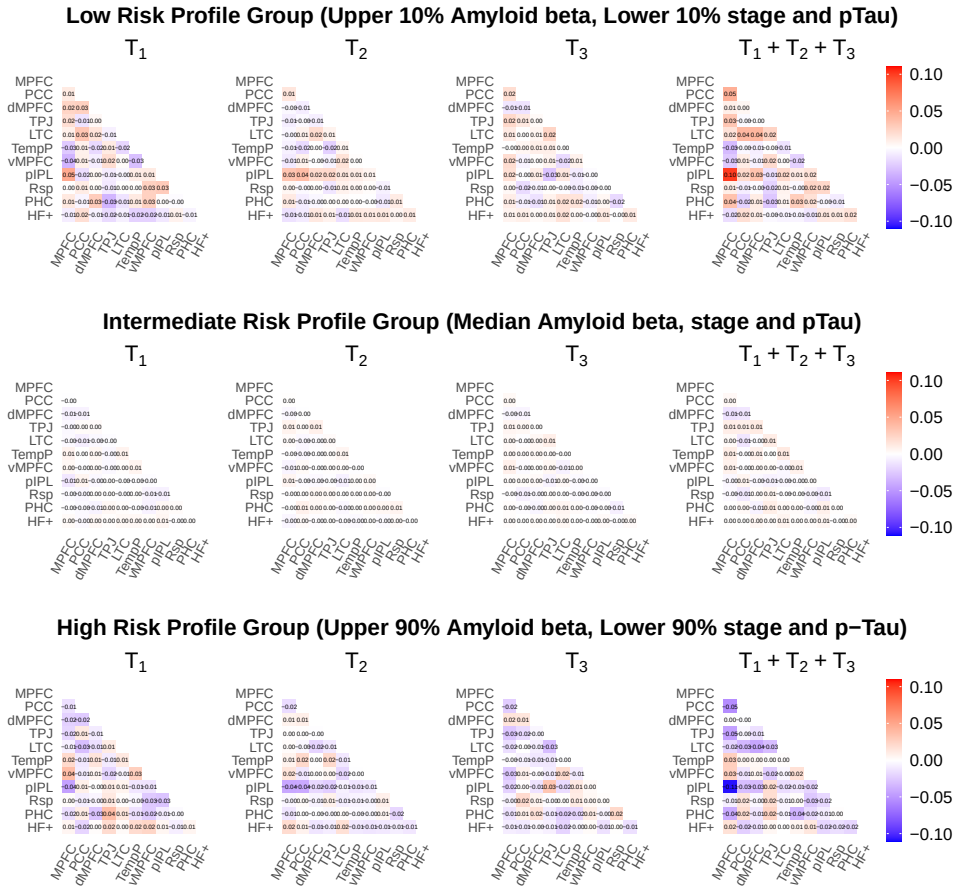

Figure 2: Estimated transport maps $T_1(x_1)$ (first column), $T_2(x_2)$ (second column), $T_3(x_3)$ (third column), and their combined map $[T_1(x_1) \oplus T_2(x_2) \oplus T_3(x_3)]$ (fourth column) represented as the difference of the lower triangular Cholesky factors, with regions of interest (ROIs) labeling rows and columns of each Cholesky factor. Rows correspond to low, intermediate, and high risk groups, with covariates $\mathbf{x} = (x_1, x_2, x_3)$ defined as follows: the low risk group corresponds to the upper 10% of Aβ values (1700), the lower 10% disease stage (CN = 0), and p-Tau values (12.8); the intermediate group is defined by the median Aβ value (995.8), median disease stage (EMCI = 2), and median p-Tau value (21.94); and the high risk group corresponds to the upper 90% of Aβ values (535), the lower 90% disease stage (LMCI = 4), and p-Tau values (42.01).

terms of MPER. Additional simulation results are provided in Supplement Section B.2.

Table 3: Mean Prediction Error Reduction (MPER, $\times 10^{-4}$) with standard errors in parentheses, based on ADNI brain correlation regression analysis with log-Cholesky metric. Higher MPER values indicate better out-of-sample prediction. We compare two methods: additive optimal transport regression (ADOPT), and global Fréchet regression (Petersen and Müller, 2019) (GF)

	ADOPT	**GF**
MPER	**2.625 (0.0554)**	1.713 (0.0518)

7 DISCUSSION

We develop ADOPT, an additive optimal transport regression approach for random object responses with multivariate continuous Euclidean predictors, thereby providing a flexible regression model for metric space-valued responses that addresses the curse of dimensionality and enhancing interpretability in real data analysis. The method is built on an extension of optimal transport ideas to general metric spaces and utilizes a transport map for each predictor.

We illustrate the proposed ADOPT with brain connectivity for resting state fMRI brain imaging data. Numerical experiments with distributional responses in Wasserstein space and SPD matrices with the log-Cholesky metric also demonstrate strong performance.

Acknowledgements

We acknowledge the helpful and valuable comments from anonymous reviewers, which led to numerous improvements. This research was supported in part by NSF grant DMS-2310450. This article is motivated by exploring the association between neuroimaging measures and cognitive function in patients with Alzheimer's disease (AD) through the analysis of the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (http://www.adni-info.org/). The ADNI is a large-scale multi-site neuroimaging study that has collected clinical, imaging, genetic and cognitive data at multiple time points from cognitive normal (CN) subjects, subjects with mild cognitive impairment (MCI), and AD patients.

References

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., and Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron*, 65(4):550–562.

Badhwar, A., Tam, A., Dansereau, C., Orban, P., Hoffstaedter, F., and Bellec, P. (2017). Resting-state network dysfunction in Alzheimer's disease: a systematic review and meta-analysis. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 8:73–85.

Bhattacharjee, S., Li, B., and Xue, L. (2025). Nonlinear global Fréchet regression for random objects via weak conditional expectation. *The Annals of Statistics*, 53(1):117–143.

Bhattacharjee, S. and Müller, H.-G. (2023). Single index Fréchet regression. *The Annals of Statistics*, 51(4):1770–1798.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Burago, D., Burago, Y., and Ivanov, S. (2022). *A Course in Metric Geometry*, volume 33. American Mathematical Society.

Chen, Y., Lin, Z., and Müller, H.-G. (2023). Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882.

Chen, Y. and Müller, H.-G. (2022). Uniform convergence of local Fréchet regression with applications to locating extrema and time warping for metric space valued trajectories. *The Annals of Statistics*, 50(3):1573–1592.

Chewi, S., Niles-Weed, J., and Rigollet, P. (2024). Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 3.

Cornea, E., Zhu, H., Kim, P., and Ibrahim, J. G. (2017). Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):463–482.

Dubey, P., Chen, Y., and Müller, H.-G. (2024). Metric statistics: Exploration and inference for random objects with distance profiles. *The Annals of Statistics*, 52(2):757–792.

Dubey, P. and Müller, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika*, 106(4):803–821.

Fornito, A., Zalesky, A., and Bullmore, E. (2016). *Fundamentals of Brain Network Analysis*. Academic press.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut Henri Poincaré*, 10(4):215–310.

Ghodrati, L. and Panaretos, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974.

Ghosal, A., Meiring, W., and Petersen, A. (2023). Fréchet single index models for object response regression. *Electronic Journal of Statistics*, 17(1):1074–1112.

Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(4):703–723.

Hampel, H., Hardy, J., Blennow, K., Chen, C., Perry, G., Kim, S. H., Villemagne, V. L., Aisen, P., Vendruscolo, M., Iwatsubo, T., et al. (2021). The amyloid-β pathway in Alzheimer's disease. *Molecular psychiatry*, 26(10):5481–5503.

Han, K., Müller, H.-G., and Park, B. U. (2020). Additive functional regression for densities as responses. *Journal of the American Statistical Association*, 115(530):997–1010.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 1(3):297–310.

Hong, H., Scealy, J. L., Wood, A. T., and Yang, Y. (2025). A robust extrinsic single-index model for spherical data. *arXiv preprint arXiv:2503.24003*.

Jeon, J. M. and Park, B. U. (2020). Additive regression with Hilbertian responses. *The Annals of Statistics*, 48(5):2671–2697.

Karikari, T. K., Ashton, N. J., Brinkmalm, G., Brum, W. S., Benedet, A. L., Montoliu-Gaya, L., Lantero-Rodriguez, J., Pascoal, T. A., Suárez-Calvet, M., Rosa-Neto, P., et al. (2022). Blood phospho-tau in

alzheimer disease: analysis, interpretation, and clinical utility. *Nature Reviews Neurology*, 18(7):400–418.

Lang, Q. and Lu, F. (2023). Small noise analysis for tikhonov and rkhs regularizations. *arXiv preprint arXiv:2305.11055*.

Lang, S. (2012). *Differential and Riemannian manifolds*, volume 160. Springer Science & Business Media.

Lee, Y. and Sul, D. (2023). Depth-weighted means of noisy data: An application to estimating the average effect in heterogeneous panels. *Journal of Multivariate Analysis*, 196:105165.

Lin, Z. (2019). Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370.

Lin, Z., Müller, H.-G., and Park, B. U. (2023). Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika*, 110(2):361–379.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pages 93–100.

Luan, Y., Rubinski, A., Biel, D., Svaldi, D. O., Higgins, I. A., Shcherbinin, S., Pontecorvo, M., Franzmeier, N., Ewers, M., et al. (2024). Tau-network mapping of domain-specific cognitive impairment in alzheimer's disease. *NeuroImage: Clinical*, 44:103699.

Marron, J. S. and Dryden, I. L. (2021). *Object Oriented Data Analysis*. Chapman and Hall/CRC.

McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179.

Murphy, M. P. and LeVine III, H. (2010). Alzheimer's disease and the amyloid-β peptide. *Journal of Alzheimer's disease*, 19(1):311–323.

Perovnik, M., Rus, T., Schindlbeck, K. A., and Eidelberg, D. (2023). Functional brain networks in the evaluation of patients with neurodegenerative disorders. *Nature Reviews Neurology*, 19(2):73–90.

Petersen, A. and Müller, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47:691–719.

Schötz, C. (2022). Nonparametric regression in non-standard spaces. *Electronic Journal of Statistics*, 16(2):4679–4741.

Severn, K. E., Dryden, I. L., and Preston, S. P. (2022). Manifold valued data analysis of samples of networks, with applications in corpus linguistics. *The Annals of Applied Statistics*, 16(1):368–390.

Song, D. and Han, K. (2023). Errors-in-variables Fréchet regression with low-rank covariate approximation. *Advances in Neural Information Processing Systems*, 36:80575–80607.

Song, W. and Müller, H.-G. (2026). Inference for dispersion and curvature of random objects. *Journal of the American Statistical Association*, 121(553):729–740.

Song, W., Zhou, H., Zhou, Y., and Müller, H.-G. (2026). Non-Euclidean data analysis with metric statistics. *Harvard Data Science Review*, 8.

van der Vaart, A. and Wellner, J. A. (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Nature.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295.

Wang, X., Zhu, J., Pan, W., Zhu, J., and Zhang, H. (2024). Nonparametric statistical inference via metric distribution function in metric spaces. *Journal of the American Statistical Association*, 119(548):2772–2784.

Yu, K., Park, B. U., and Mammen, E. (2008). Smooth backfitting in generalized additive models. *The Annals of Statistics*, 36(1):228–260.

Yuan, Y., Zhu, H., Lin, W., and Marron, J. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):697–719.

Zhang, Q., Xue, L., and Li, B. (2024). Dimension reduction for Fréchet regression. *Journal of the American Statistical Association*, 119(548):2733–2747.

Zhou, Y. and Müller, H.-G. (2022). Network regression with graph Laplacians. *Journal of Machine Learning Research*, 23(320):1–41.

Zhu, C. and Müller, H.-G. (2023). Autoregressive optimal transport models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):1012–1033.

Zhu, C. and Müller, H.-G. (2025). Geodesic optimal transport regression. *Biometrika*, page asaf086.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Assumptions are introduced in Section 2 and Supplement Section A, the proposed

model and algorithm for estimation are provided in Section 4.

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We provide consistency results in Section 4, and complexity analysis in Supplement Section C.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Partial Yes] The resting-state fMRI data we used are fully accessible to qualified researchers through the ADNI website (https://adni.loni.usc.edu) upon approval of a standard data-use application (https://adni.loni.usc.edu/data-samples/adni-data). We cannot redistribute the raw data ourselves due to ADNI's data-use agreement, not meaning that the data are unavailable to other researchers.

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes] All theoretical assumptions are provided in Supplement Section A.

(b) Complete proofs of all theoretical results. [Yes] All theoretical proofs are provided in Supplement Section A.

(c) Clear explanations of any assumptions. [Yes] See Supplement Section A.

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Partial Yes] The resting-state fMRI data we used are fully accessible to qualified researchers through the ADNI website (https://adni.loni.usc.edu) upon approval of a standard data-use application (https://adni.loni.usc.edu/data-samples/adni-data). We cannot redistribute the raw data ourselves due to ADNI's data-use agreement, not meaning that the data are unavailable to other researchers.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] We use 60 CPUs for the simulation analysis.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials to "ADOPT: Additive Optimal Transport Regression"

A Technical Details

A.1 Assumptions

In this section, we state the assumptions required for Theorem 1 and introduce the necessary notation. For all $\nu \in \mathcal{M}$ and $T_j \in \mathcal{T}$, $j = 1, \ldots, p$, define $Z_j(x_j) = [\varepsilon \circ T_j(x_j)](\nu) \in \mathcal{M}$, $x_j \in \mathcal{X}_j$, where $\varepsilon \in \mathcal{T}$ is a random perturbation map (Chen and Müller, 2022). We define $T_j(x_j) = T_{\mathbb{E}_{\oplus} g_j(X_j), g_j(x_j)} \in \mathcal{T}$, and

$$
\begin{aligned}
M_{\oplus j}(\omega, x_j; \nu) &= \mathbb{E}\left[d^2\left(Z_j(X_j)(\nu), \omega\right) \mid X_j = x_j\right], \\
M^L_{\oplus j}(\omega, x_j; \nu) &= \mathbb{E}\left[w(x_j, h) d^2\left(Z_j(X_j)(\nu), \omega\right)\right], \\
\hat{M}^L_{\oplus j}(\omega, x_j; \nu) &= \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i(x_j, h) d^2\left(Z_{i,j}(X_{i,j})(\nu), \omega\right),
\end{aligned}
$$

where $Z_{i,j}(X_{i,j})(\nu) = [\varepsilon_i \circ T_j(X_{i,j})](\nu) \in \mathcal{M}$.

By definition we have

$$
\begin{aligned}
T_j(x_j)(\nu) &= \operatorname*{arg\,min}_{\omega \in \mathcal{M}} M_{\oplus j}(\omega, x_j; \nu), \\
T^L_j(x_j)(\nu) &= \operatorname*{arg\,min}_{\omega \in \mathcal{M}} M^L_{\oplus j}(\omega, x_j; \nu), \\
\hat{T}^L_j(x_j)(\nu) &= \operatorname*{arg\,min}_{\omega \in \mathcal{M}} \hat{M}^L_{\oplus j}(\omega, x_j; \nu).
\end{aligned}
$$

We need the following assumptions:

(A1) The kernel function K is a symmetric probability density function centered at zero. Furthermore, define $K_{kl} = \int_{\mathbb{R}} K^k(u) u^j du$, where $|K_{14}|$ and $|K_{26}|$ are both finite.

(A2) For all $j = 1, \ldots, p$, the marginal density f_j of X_j as well as the conditional densities $f_{|z_j}(x_j)$ of $X_j \mid Z_j(x_j)(\nu) = z_j$ and $f_{|x_j}(z_j)$ of $Z_j \mid X_j = x_j$ exist. The conditional densities $f_{|z_j}(x_j)$ are twice continuously differentiable for all $z_j \in \mathcal{M}$ and $\sup_{x_j, z_j} |f''_{|z_j}(x_j)| < \infty$. For any open set $U \subset \mathcal{M}$, $\int_U dF_{|x_j}(z_j)$ is continuous in $x_j \in \mathcal{X}_j$, where $F_{|x_j}(z_j)$ is the cumulative distribution function corresponding to $f_{|x_j}(z_j)$.

(A3) For all $\nu \in \mathcal{M}$, and $x_j \in \mathcal{X}_j$, $j = 1, \ldots, p$, objects $T_j(x_j)(\nu)$, $T^L_j(x_j)(\nu)$, and $\hat{T}^L_j(x_j)(\nu)$ exist and are unique, the latter almost surely. For all $\delta > 0$,

$$
\begin{aligned}
\inf_{d(\omega, T_j(x_j)(\nu)) < \delta} \left[M_{\oplus j}(\omega, x_j; \nu) - M_{\oplus j}(T_j(x_j)(\nu), x_j; \nu)\right] > 0, \\
\liminf_{h \to 0} \inf_{d(\omega, T^L_j(x_j)(\nu)) < \delta} \left[M^L_{\oplus j}(\omega, x_j; \nu) - M^L_{\oplus j}(T^L_j(x_j)(\nu), x_j; \nu)\right] > 0.
\end{aligned}
$$

Condition *(A1)* is standard kernel requirements typically assumed in local regression estimation. Condition *(A2)* is distributional assumptions on the predictors and responses for the convergence of Fréchet regression estimators (Petersen and Müller, 2019). Condition *(A3)* is a regularity condition commonly used to establish the consistency of M-estimators (van der Vaart and Wellner, 2023).

We need the following assumption on the geodesic optimal transport:

(A4) (Small Error Assumption) For perturbation maps $\{\varepsilon_{i,n}\}_{i=1}^n$ with sample size n, there exists a sequence $\delta_n > 0$ with $\delta_n \to 0$ as $n \to \infty$, such that

$$\mathbb{E}\left[d(\varepsilon_{1,n}(\nu), \nu)\right] \leq \delta_n,$$

for all $\nu \in \mathcal{M}$.

(A5) There exists a constant $C > 0$ such that

$$d(\Gamma(v_1, v_2, v_3), \Gamma(v_1', v_2', v_3')) \leq C\left\{d(v_1, v_1') + d(v_2, v_2') + d(v_3, v_3')\right\}$$

(A6) The transport maps $T_j(\cdot) \in \mathcal{T}$ are perturbation maps, i.e., for all $\nu \in \mathcal{M}$,

$$\nu = \underset{\omega \in \mathcal{M}}{\arg\min}\, \mathbb{E}\left\{d^2(T_j(X_j)(\nu), \omega)\right\},$$

and

$$\inf_{d(\omega,\nu)>\delta} \left[\mathbb{E}d^2(T_j(X_j)(\nu), w) - \mathbb{E}d^2(T_j(X_j)(\nu), \nu)\right] > 0,$$

for any $\delta > 0$.

The small error assumption (A4) appears at various instances in the classical nonparametric regression literature, including generalized regression models for longitudinal data (Hall et al., 2008), errors-in-variables regression (Lee and Sul, 2023), and Tikhonov/RKHS-regularized regression (Lang and Lu, 2023). (A4) is only required for Lemma 3, where we need the quantity $d(T \oplus \varepsilon, \varepsilon \oplus T)$ to be small, because the composition of the transport map with the error-perturbation map is not commutative and our transport backfitting algorithm iteratively applies partial transport maps. Controlling this non-commutativity error is therefore essential for establishing the theoretical guarantees. We note that (A4) is not needed in metric spaces where the $\oplus$ operations are commutative (e.g., SPD matrices under the log–Cholesky metric).

Condition *(A5)* holds for random objects in complete, non-positively curved metric spaces (Hadamard spaces) (Zhu and Müller, 2025). Examples include the space of univariate distributions with the Wasserstein metric and the space of symmetric positive definite (SPD) matrices under the log-Cholesky, log-Euclidean, Frobenius, or power-Frobenius metrics, all of which are Hadamard spaces. Condition *(A6)* is also satisfied in these settings. For distributional-valued response in Wasserstein space, let F_{X_j}, $F_{\oplus j}$, and F_ν be distribution functions, and $F_{X_j}^{-1}$, $F_{\oplus j}^{-1}$, and F_ν^{-1} their corresponding quantile functions of $g_j(x_j)$, $\mathbb{E}_\oplus g_j(X_j)$, and $\nu \in \mathcal{M}$. Then the transport map is represented as a quantile function $T_j(X_j)(\nu) = F_{X_j}^{-1} \circ F_{\oplus j} \circ F_\nu^{-1}$, and its Fréchet mean $F_{\oplus j}^{-1} \circ F_{\oplus j} \circ F_\nu^{-1} = F_\nu^{-1}$, so *(A6)* is satisfied. For SPD matrices with the log-Cholesky metric, the Cholesky decomposition yields a vector representation in Euclidean space, which directly implies that Condition *(A6)* holds. The same argument extends to other metrics on SPD matrices, including the Frobenius, and power-Frobenius metrics. The inequality in Condition *(A6)* serves as a regularity condition analogous to *(A3)*, ensuring the consistency of M-estimators and the uniqueness of the population Fréchet mean.

A.2 Proof of Theorem 1

We need to show that for any $\nu \in \mathcal{M}$ and $x_j \in \mathcal{X}_j$, if the bandwidth $h \to 0$ and $nh \to \infty$, it holds that

$$d(T_j(x_j)(\nu), \hat{T}_j(x_j)(\nu)) = o_p(1),$$

where $\hat{T}_j(\nu)$ is the estimates for the proposed transport backfitting algorithm.

For fixed $\nu \in \mathcal{M}$ and $x_j \in \mathcal{X}_j$ for fixed j, we have

$$d\left(T_j(x_j)(\nu), \hat{T}_j(x_j)(\nu)\right) \leq d\left(T_j(x_j)(\nu), \hat{T}_j^L(x_j)(\nu)\right) + d\left(\hat{T}_j^L(x_j)(\nu), \hat{T}_j(x_j)(\nu)\right) \tag{S.1}$$

By Lemma 1 and the proof Theorem 3 in the Supplement of Petersen and Müller (2019), under conditions *(A1)-(A3)*, we observe the bias term

$$d\left(T_j(x_j)(\nu), T_j^L(x_j)(\nu)\right) = o(1), \tag{S.2}$$

as the bandwidth $h \to 0$. Also, by Lemma 2 in the Supplement of Petersen and Müller (2019), under conditions *(A1)* and *(A3)*, we observe the stochastic term

$$d\left(T_j^L(x_j)(\nu), \hat{T}_j^L(x_j)(\nu)\right) = o_p(1) \tag{S.3}$$

as $h \to 0$ and $nh \to \infty$. Combining (S.2) and (S.3), we have the first term of (S.1)

$$d\left(T_j(x_j)(\nu), \hat{T}_j^L(x_j)(\nu)\right) = o_p(1) \tag{S.4}$$

Thus, it is enough to show that the second term of (S.1)

$$d\left(\hat{T}_j^L(x_j)(\nu), \hat{T}_j(x_j)(\nu)\right) = o_p(1). \tag{S.5}$$

By the transport backfitting algorithm we can represent,

$$\hat{T}_j(x_j)(\nu) = T_{\hat{\mathbb{E}}_\oplus \hat{g}_j(X_j), \hat{g}_j(x_j)}(\nu),$$

where

$$\hat{g}_j(x_j) = \underset{\omega \in \mathcal{M}}{\arg\min} \frac{1}{n} \sum_{i=1}^n \hat{w}_i(x_j, h) d^2(P_{\oplus i,j}(\nu), \omega),$$
$$\hat{\mathbb{E}}_\oplus \hat{g}_j(X_j) = \underset{\omega \in \mathcal{M}}{\arg\min} \frac{1}{n} \sum_{i=1}^n d^2(\hat{g}_j(X_{i,j}), \omega).$$

By condition *(A6)* and the property of transport map, we have $\hat{T}_j^L(x_j)(\nu) = T_{\nu, \hat{T}_j^L(x_j)(\nu)}(\nu) = T_{\mathbb{E}_\oplus T_j(X_j)(\nu), \hat{T}_j^L(x_j)(\nu)}(\nu)$. The equation (S.5) is

$$\begin{aligned} d\left(\hat{T}_j^L(x_j)(\nu), \hat{T}_j(x_j)(\nu)\right) &= d\left(T_{\mathbb{E}_\oplus T_j(X_j)(\nu), \hat{T}_j^L(x_j)(\nu)}(\nu),\ T_{\hat{\mathbb{E}}_\oplus \hat{g}_j(X_j), \hat{g}_j(x_j)}(\nu)\right) \\ &\le C\left\{d\left(\mathbb{E}_\oplus T_j(X_j)(\nu), \hat{\mathbb{E}}_\oplus \hat{g}_j(X_j)\right) + d\left(\hat{T}_j^L(x_j)(\nu), \hat{g}_j(x_j)\right)\right\}, \end{aligned}$$

for some constant $C > 0$, and the last inequality came from condition *(A5)*.

Lemma 1 *Under (A1), (A3), (A4), and (A5), if the bandwidth $h \to 0$ and $nh \to \infty$ as $n \to \infty$, it holds that*

$$d\left(\hat{T}_j^L(x_j)(\nu), \hat{g}_j(x_j)\right) = o_P(1),$$

for any $x_j \in \mathcal{X}_j$, and $\nu \in \mathcal{M}$.

Lemma 2 *Under (A1)-(A6), if the bandwidth $h \to 0$ and $nh \to \infty$ as $n \to \infty$, it holds that*

$$d\left(\mathbb{E}_\oplus T_j(X_j)(\nu), \hat{\mathbb{E}}_\oplus \hat{g}_j(X_j)\right) = o_P(1),$$

for any $x_j \in \mathcal{X}_j$, and $\nu \in \mathcal{M}$.

Combining Lemma 1 and Lemma 2, we have $d\left(\hat{T}_j^L(x_j)(\nu), \hat{g}_j(x_j)\right) = o_p(1)$ and $d\left(\mathbb{E}_\oplus T_j(X_j)(\nu), \hat{\mathbb{E}}_\oplus \hat{g}_j(X_j)\right) = o_p(1)$, leading to the desired results in (S.5),

$$d\left(\hat{T}_j^L(x_j)(\nu), \hat{T}_j(x_j)(\nu)\right) = o_p(1).$$

Then, the equation (S.4) and (S.5) verifies our goal of (S.1),

$$d\left(T_j(x_j)(\nu), \hat{T}_j(x_j)(\nu)\right) = o_p(1).$$

A.3 Proof of Lemmas

A.3.1 Proof of Lemma 1

By Corollary 3.2.3 in van der Vaart and Wellner (2023), for the proof of the consistency of M-estimators $d\left(\hat{T}_j^L(x_j)(\nu), \hat{g}_j(x_j)\right) = o_p(1)$, it is sufficient to show that the uniform consistency of loss functions,

$$\sup_{\omega\in\mathcal{M}} \left| \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) d^2\left(Z_{i,j}(X_{i,j})(\nu),\omega\right) - \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) d^2(P_{\oplus i,j}(\nu),\omega) \right| = o_p(1).$$

Combining Theorem 1.3.6 and 1.5.4 of van der Vaart and Wellner (2023), we need to show that

(1) (Pointwise consistency) $\frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) d^2\left(Z_{i,j}(X_{i,j})(\nu),\omega\right) - \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) d^2(P_{\oplus i,j}(\nu),\omega) = o_p(1)$ for all $\omega \in \mathcal{M}$, and

(2) (Asmptotically equicontinuous in probability) For all $\epsilon, \eta > 0$, there exists $\delta > 0$ such that

$$\limsup_n P\left(\sup_{d(\omega_1,\omega_2)<\delta} |A_n(\omega_1) - A_n(\omega_2)| > \epsilon \right) < \eta,$$

where $A_n(\omega) = \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) d^2\left(Z_{i,j}(X_{i,j})(\nu),\omega\right) - \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) d^2(P_{\oplus i,j}(\nu),\omega)$.

Begin with (1),

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) \left\{ d^2\left(Z_{i,j}(X_{i,j})(\nu),\omega\right) - d^2(P_{\oplus i,j}(\nu),\omega) \right\} \right| \\
&= \left| \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) \Big\{ d^2\left([\varepsilon_{i,n}\circ T_j(X_{i,j})](\nu),\omega\right) \right. \\
&\qquad \left. - d^2([T_{j+1}^{-1}(X_{i,j+1})\circ\cdots\circ T_p^{-1}(X_{i,p})\circ\varepsilon_{i,n}\circ T_p(X_{i,p})\circ\cdots\circ T_{j+1}(X_{i,j+1})\circ T_j(X_{i,j})](\nu),\omega) \Big\} \right| \\
&\le 2\mathrm{diam}(\mathcal{M}) \\
&\times \frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) d\left([\varepsilon_{i,n}\circ T_j(X_{i,j})](\nu), [T_{j+1}^{-1}(X_{i,j+1})\circ\cdots\circ T_p^{-1}(X_{i,p})\circ\varepsilon_{i,n}\circ T_p(X_{i,p})\circ\cdots\circ T_j(X_{i,j})](\nu)\right),
\end{aligned}$$

where the last inequality is from $d^2(a,\omega) - d^2(b,w) \le |d(a,w) - d(b,w)|(d(a,w) + d(b,w)| \le 2\mathrm{diam}(\mathcal{M}) d(a,b)$, for any $a, b \in \mathcal{M}$, and fixed $\omega \in M$.

Lemma 3 *Under conditions (A4) and (A5), it holds that*

$$\frac{1}{n}\sum_{i=1}^n d\left([T\circ\varepsilon_{i,n}](\xi), [\varepsilon_{i,n}\circ T](\xi)\right) = o_p(1),$$

for any $T \in \mathcal{T}$, and $\xi \in \mathcal{M}$.

By Lemma 3 with $T = T_{j+1}^{-1}(X_{i,j+1})\circ\cdots\circ T_p^{-1}(X_{i,p}) \in \mathcal{T}$ and $\xi = [T_p(X_{i,p})\circ\cdots\circ T_j(X_{i,j})](\nu) \in \mathcal{M}$, we observe that

$$\frac{1}{n}\sum_{i=1}^n d\left([\varepsilon_{i,n}\circ T_j(X_{i,j})](\nu), [T_{j+1}^{-1}(X_{i,j+1})\circ\cdots\circ T_p^{-1}(X_{i,p})\circ\varepsilon_{i,n}\circ T_p(X_{i,p})\circ\cdots\circ T_j(X_{i,j})](\nu)\right) = o_p(1).$$

Also, by Lemma 2 in Supplement of Petersen and Müller (2019), $\frac{1}{n}\sum_{i=1}^n \hat{w}_i(x_j,h) = O_p(1)$, with $\hat{w}_i(x,h) > 0$ for all x. Since $\mathcal{M}$ is bounded, we have

$$\left| \frac{1}{n}\sum_{i=1}^n \hat{w}(x_j,h) \left\{ d^2\left(Z_{i,j}(X_{i,j})(\nu),\omega\right) - d^2(P_{\oplus i,j}(\nu),\omega) \right\} \right| = o_p(1),$$

which is the pointwise consistency in (1).

Moving on to (2),

$$\begin{aligned}|A_n(\omega_1) - A_n(\omega_2)| &= \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(x_j, h)\left\{d^2(Z_{i,j}(X_{i,j})(\nu), \omega_1) - d^2(Z_{i,j}(X_{i,j})(\nu), \omega_2)\right\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i(x_j, h)\left\{d^2(P_{\oplus i,j}(\nu), \omega_1) - d^2(P_{\oplus i,j}(\nu), \omega_2)\right\} \\ &\leq 4\mathrm{diam}(\mathcal{M}) d(\omega_1, \omega_2)\frac{1}{n}\sum_{i=1}^{n}|\hat{w}_i(x_j, h)|.\end{aligned}$$

Then, $\frac{1}{n}\sum_{i=1}^{n}|\hat{w}_i(x_j, h)| = O_p(1)$ leads to $|A_n(\omega_1) - A_n(\omega_2)| = O_p(d(\omega_1, \omega_2))$, which verifies (2).

A.3.2 Proof of Lemma 2

We observe

$$d\left(\mathbb{E}_{\oplus}T_j(X_j)(\nu), \hat{\mathbb{E}}_{\oplus}\hat{g}_j(X_j)\right) \leq d\left(\mathbb{E}_{\oplus}T_j(X_j)(\nu), \hat{\mathbb{E}}_{\oplus}T_j(X_j)(\nu)\right) + d\left(\hat{\mathbb{E}}_{\oplus}T_j(X_j)(\nu), \hat{\mathbb{E}}_{\oplus}\hat{g}_j(X_j)\right).$$

By Dubey and Müller (2019), under condition *(A6)*, the consistency of sample Fréchet mean holds, i.e.,

$$d\left(\mathbb{E}_{\oplus}T_j(X_j)(\nu), \hat{\mathbb{E}}_{\oplus}T_j(X_j)(\nu)\right) = o_p(1). \tag{S.6}$$

Following proof of Lemma 1, for the proof of $d\left(\hat{\mathbb{E}}_{\oplus}T_j(X_j)(\nu), \hat{\mathbb{E}}_{\oplus}\hat{g}_j(X_j)\right) = o_p(1)$, it is sufficient to show that

$$\sup_{\omega\in M}\left|\frac{1}{n}\sum_{i=1}^{n} d^2(\hat{g}_j(X_{i,j}), \omega) - \frac{1}{n}\sum_{i=1}^{n} d^2(T_j(X_{i,j}), \omega)\right| = o_p(1).$$

We observe

$$\begin{aligned}\sup_{\omega\in M}\left|\frac{1}{n}\sum_{i=1}^{n} d^2(\hat{g}_j(X_{i,j}), \omega) - \frac{1}{n}\sum_{i=1}^{n} d^2(T_j(X_{i,j}), \omega)\right| &\leq 2\mathrm{diam}(\mathcal{M})\frac{1}{n}\sum_{i=1}^{n} d(\hat{g}_j(X_{i,j}), T_j(X_{i,j})) \\ &\leq 2\mathrm{diam}(\mathcal{M})\frac{1}{n}\sum_{i=1}^{n}\left\{d(\hat{g}_j(X_{i,j}), \hat{T}_j^L(X_{i,j})) + d(\hat{T}_j^L(X_{i,j}), T_j(X_{i,j}))\right\}.\end{aligned}$$

By Lemma 1, the first term $\frac{1}{n}\sum_{i=1}^{n} d(\hat{g}_j(X_{i,j}), \hat{T}_j^L(X_{i,j})) = o_p(1)$ and by equation (S.4), the second term $\frac{1}{n}\sum_{i=1}^{n} d(\hat{T}_j^L(X_{i,j}), T_j(X_{i,j})) = o_p(1)$. Thus, we have

$$d\left(\hat{\mathbb{E}}_{\oplus}T_j(X_j)(\nu), \hat{\mathbb{E}}_{\oplus}\hat{g}_j(X_j)\right) = o_p(1). \tag{S.7}$$

Combining (S.6) and (S.7), we verify

$$d\left(\mathbb{E}_{\oplus}T_j(X_j)(\nu), \hat{\mathbb{E}}_{\oplus}\hat{g}_j(X_j)\right) = o_p(1).$$

A.3.3 Proof of Lemma 3

For fixed $T \in \mathcal{T}$, and $\xi \in \mathcal{M}$, we observe

$$\begin{aligned}d\left([T \circ \varepsilon_{i,n}](\xi), [\varepsilon_{i,n} \circ T](\xi)\right) &\leq d\left([T \circ \varepsilon_{i,n}](\xi), T(\xi)\right) + d\left(T(\xi), [\varepsilon_{i,n} \circ T](\xi)\right) \\ &\leq Cd\left(\varepsilon_{i,n}(\xi), \xi\right) + d\left(T(\xi), [\varepsilon_{i,n} \circ T](\xi)\right),\end{aligned}$$

for constant $C > 0$. The first inequality is from triangle inequality, and second inequality is from condition *(A5)*.

Then,

$$\frac{1}{n}\sum_{i=1}^{n} d\left([T\circ\varepsilon_{i,n}](\xi),[\varepsilon_{i,n}\circ T](\xi)\right) \le C\frac{1}{n}\sum_{i=1}^{n} d\left(\varepsilon_{i,n}(\xi),\xi\right)+\frac{1}{n}\sum_{i=1}^{n} d\left(T(\xi),[\varepsilon_{i,n}\circ T](\xi)\right)$$
$$\to C\mathbb{E}[d(\varepsilon_{1,n}(\xi),\xi)]+\mathbb{E}[d(\varepsilon_{1,n}(T(\xi)),T(\xi))] \quad \text{by WLLN.}$$

By condition *(A4)*, we have

$$\frac{1}{n}\sum_{i=1}^{n} d\left([T\circ\varepsilon_{i,n}](\xi),[\varepsilon_{i,n}\circ T](\xi)\right)=o_p(1),$$

which is the desired results.

B Further details on Brain connectivity for fMRI data analysis

B.1 Data Preprocessing

Brain signal analysis at the subject level relies on time series of Blood Oxygen Level Dependent (BOLD) signals, into a set of regions of interest (ROI). The coherence between pairwise ROIs is usually measured by Pearson correlation coefficients of the fMRI time series, leading to $l\times l$ correlation matrix for l ROIs. Alzheimer's disease has been found to be associated with anomalies in the functional integration of ROIs (Badhwar et al., 2017; Perovnik et al., 2023). The list of $l=11$ ROIs used in the analysis is provided in Table S.1. Preprocessing was conducted in MATLAB using the Statistical Parametric Mapping (SPM12, `www.fil.ion.ucl.ac.uk/spm`) and the rs-fMRI Data Analysis Toolkit V1.8 (REST1.8, `https://rfmri.org/REST`); further details are available in Zhou and Müller (2022).

Table S.1: Brain region of interest (ROI) used in the analysis along with its ROI label (Andrews-Hanna et al., 2010).

Region Full Name	ROI Label
Anterior medial prefrontal cortex	MPFC
Posterior cingulate cortex	PCC
Dorsal medial prefrontal cortex	dMFPC
Temporal parietal junction	TPJ
Lateral temporal cortex	LTC
Temporal pole	TempP
Ventral medial prefrontal cortex	vMFPC
Posterior inferior parietal lobule	pIPL
Retrosplenial cortex	Rsp
Parahippocampal cortex	PHC
Hippocampal formation	HF+

B.2 Additional Simulations

In this subsection, we provide out-of-sample prediction performance comparison results for a stratified analysis where we use stage-specific Fréchet means. Subjects were grouped into three diagnostic stages:

- **Group 1** (358 subjects): Cognitively Normal (CN), Subjective Memory Complaint (SMC),
- **Group 2** (499 subjects): Early Mild, Mild, and Late Mild Cognitive Impairment (EMCI, MCI, LMCI),

- **Group 3** (72 subjects): Alzheimer's Disease (AD).

We evaluate out-of-sample prediction performance using 5-fold cross-validation and partition the index set $1, \ldots, n$ into disjoint folds $\mathcal{I}_q$ $(q = 1, \ldots, 5)$ with $n_q = |\mathcal{I}_q|$. For each fold, we define the refined Mean Prediction Error Reduction (rMPER) as

$$\text{rMPER} = \frac{1}{5}\sum_{q=1}^{5} \text{rMPER}_q,$$
$$\text{rMPER}_q = \frac{1}{n_q}\sum_{i\in\mathcal{I}_q} \left\{ d(\hat{\mu}_{i,\oplus}^{(-q)}, Y_i) - d(\hat{Y}_i^{(-q)}, Y_i) \right\},$$

where $\hat{Y}_i^{(-q)}$ is the predicted response from the regression model trained on the remaining folds, $\hat{\mu}_{i,\oplus}^{(-q)}$ is the disease-profile Fréchet mean corresponding to the ith subject's group (Group 1, 2, or 3) in the training set of fold q, and d is the log-Cholesky metric. rMPER measures the reduction in prediction error relative to the disease-profile baseline; positive rMPER values indicate improved prediction relative to the disease-profile baseline. We compare our method against global Fréchet regression (GF) (Petersen and Müller, 2019). Table S.2 reports the average of rMPER ($\times 10^{-4}$) obtained for $B = 200$ Monte Carlo simulations for ADOPT and global Fréchet regression. ADOPT achieves a larger rMPER, indicating better out-of-sample predictive performance than global Fréchet regression.

Table S.2: Mean Prediction Error Reduction (MPER $\times 10^{-4}$) with standard errors in parentheses, based on ADNI brain correlation regression analysis with log-Cholesky metric. Higher MPER values indicate better out-of-sample prediction. We compare two methods: additive optimal transport regression (ADOPT), and global Fréchet regression (Petersen and Müller, 2019) (GF)

	ADOPT	**GF**
MPER	**3.020 (0.0526)**	2.108 (0.0518)

To check the robustness of the ADOPT model, we repeat the following procedure 1000 times: Each time 1) fitting the ADOPT model on a random 80% subsample and 2) identifying the strongest signal among 55 ROI-to-ROI correlations. The PIPL–MPFC connection appears the strongest signal in 545 of 1000 runs, consistent with the full-data result.

C Computational Complexity Analysis

The computational complexity of ADOPT depends on the underlying metric space $(\mathcal{M}, d)$. Let n be the number of observations, and p be the dimension of predictors. For each cycle of transport backfitting iteration, roughly speaking, the overall complexity is $O(n^2p+np^2)$, which simplifies to $O(n^2p)$ in the common setting where $p << n$. For specific object spaces $(\mathcal{M}, d)$:

- Univariate distributions with 2-Wasserstein metric: When probability distributions are represented by quantile functions with q discretization points, the complexity becomes $O(n^2pq)$.
- Symmetric positive definite (SPD) matrices with Log-Cholesky metric: For $m \times m$ SPD matrices, operations scale with the number of lower-triangular components. In this case, the complexity is $O(n^2pm^2)$.