Journal of Machine Learning Research 27 (2026) 1-49 Submitted 10/23; Revised 9/25; Published 3/26

Global Fréchet Manifold Learning for Random Objects, With Application to Low-Dimensional Wasserstein Representations of Distributional Data

Álvaro Gajardo AEGAJARDO@UCDAVIS.EDU
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Hans-Georg Müller HGMUELLER@UCDAVIS.EDU
University of California, Davis
Davis, CA 95616, USA

Editor: Aryeh Kontorovich

Abstract

We study manifold learning with multidimensional scaling for samples of metric space valued data. By adopting a global version of ISOMAP we obtain low-dimensional Euclidean representations. A key innovation is that we demonstrate that global Fréchet regression can be utilized for mapping the elements of a convex set in the Euclidean representation space back to the metric space where the objects reside. We refer to this approach as Fréchet manifold learning and showcase it with one-dimensional distributions as random objects, equipped with the Wasserstein metric, which is an important special case of our general approach. The resulting low-dimensional representations mimic the parametric representation in a parametric family of distributions but are entirely learned from the data without postulating any parametric model. These Wasserstein representations of distributional data can be viewed as an empirical parametrization of a sample of distributions. The utility of these representations rests on the map from the low-dimensional Euclidean representation space to the space of distributions, which is obtained with global Fréchet regression. We illustrate the proposed approach with distributional data for baby names, bike rentals and age pyramids and further demonstrate how it can be applied for a novel distributional regression method that features one-dimensional distributions as predictors.

Keywords: distributional data analysis, global Fréchet regression, MDS, metric space-valued data, nonlinear dimension reduction

1. Introduction

Advances in technology have enabled the acquisition of increasingly complex data such as functional data and next generation functional data (Wang et al. 2016), functional data on manifolds (Dai and Müller 2018), and random objects in general metric spaces such as networks, probability distributions, covariance matrices, point processes and many others (Dryden et al. 2009; Ginestet et al. 2017; Gervini and Khanal 2019; Dubey and Müller 2022; Gajardo and Müller 2022). Here random objects refer to metric-space valued random variables, following the definition in Müller (2016). Since such data typically lie in a high-dimensional or even infinite-dimensional ambient space, dimension reduction is often

mandatory for statistical analysis. The literature on dimension reduction in regression is vast and encompasses methodology ranging from single and multiple index modeling in regression analysis (Xia et al. 2002; Lee et al. 2013; Zhang et al. 2021; Ghosal et al. 2023a; Bhattacharjee and Müller 2023) to functional principal component analysis (Kleffe 1973; Ramsay and Silverman 2005; Hsing and Eubank 2015; Li and Song 2017), where the latter enables unsupervised lower-dimensional approximations for infinite-dimensional data.

More generally, manifold learning is a non-linear dimension reduction tool that has been shown to be useful for dimension reduction of infinite-dimensional functional data (Chen and Müller 2012), where data are assumed to lie on a low-dimensional manifold that is unknown. Manifold learning aims to identify an underlying manifold on which the data are located and if such a manifold exists provides nonlinear dimension reduction that facilitates subsequent statistical analysis and circumvents the curse of dimensionality. This leads to faster rates of convergence compared to fully nonparametric methods under general smoothness assumptions (Bickel and Li 2007; Lin and Yao 2019), which are known to suffer from the curse of dimensionality.

ISOMAP, which stands for isometric feature mapping (Tenenbaum et al. 2000), is a key method in manifold learning for the recovery of an underlying low-dimensional manifold on which data are assumed to lie. Variants include P-ISOMAP (Chen and Müller 2012) or Wassmap for distributional image data (Hamm et al. 2023), where the latter employs the 2-Wasserstein metric (Villani 2003) as the underlying geodesic distance. The ISOMAP approach approximates the geodesic distance in the manifold by finding shortest paths in local neighborhoods, followed by multidimensional scaling (MDS) (Borg and Groenen 2005), applied to the resulting pairwise distance matrix between objects and resulting in a low-dimensional Euclidean representation. These local neighborhoods are usually constructed based on all points within a fixed radius in the ambient space metric or alternatively via a K-nearest neighbor approach. Thus ISOMAP approximates the geodesic distance in the manifold by finding shortest paths on the weighted graph that is constructed across all data points, with edge weights corresponding to the distances between neighboring points.

In this paper, we adopt a global version of ISOMAP similarly as in Hamm et al. (2023), where the object space is assumed to be directly isometric in the ambient space metric to an unknown but low-dimensional Euclidean space. We apply the multidimensional scaling step directly and without finding shortest paths with Dijkstra's algorithm, which is not needed under the isometry assumption that underpins our approach. That is, we do not focus on general curved manifolds for which the geodesic estimation step along the locally weighted graph from ISOMAP would be required, since the isometry condition (in the ambient space metric) is satisfied by flat manifolds. While the flatness of the manifolds considered limits its scope, our framework covers important statistical models such as location-scale families in the special case of distributional data and has broad applicability for other non-Euclidean data, with random objects residing in more general metric spaces.

While various methods are available for non-linear dimension reduction for the exploration of complex data, they generally only include maps from the space of objects to their representation in low-dimensional Euclidean space but do not provide a construction for the inverse map from the Euclidean representation space to the object space. However, the latter is essential to assign objects to elements in the Euclidean representation space that do not directly represent an object, i.e., where there is no object that has been mapped to such

elements. To be able to make such assignments is especially relevant for regression models or principal component representations, as these, when performed in the lower-dimensional representation space, involve elements of this space that are not images of observed objects in the object space. This motivates the approach presented in this paper. Another motivation is that when the objects lie on an unknown manifold in a general metric space, interpolation is a difficult task as there are no vector operations available in the object space and interpolation is therefore performed in the representation space, where again the inverse map back into the object space is not available for the resulting interpolations. Our approach provides a solution to this quandary by constructing such inverse maps.

Probability distributions are a special case of substantial interest, for which we showcase the proposed methods in Sections 3-5. Distributional data analysis has recently become an increasingly popular area for statistics and data analysis with applications in many fields (Petersen and Müller 2016; Matabuena et al. 2021; Chen et al. 2023; Petersen et al. 2022; Pegoraro and Beraha 2022; Zhu and Müller 2024; Ghosal et al. 2023b; Gunsilius 2023). As an illustrative example, consider the simulation setting in Figure 3 below for a manifold in the 2-Wasserstein space that is generated from Gaussian variates with two independent sources of variation corresponding to a mean shift and a scale variation. This information is not used in the proposed approach. While it is possible to obtain a data-driven 2-dimensional Euclidean representation of both mean and variance parameters for each distribution, this would only entail the inverse map at these Euclidean representation points by plugging in the parameters, but the inverse map would remain unknown for other Euclidean 2-vectors, thus preventing interpolation in the object space. The proposed approach overcomes these limitations.

To obtain the inverse map at all Euclidean representation points is one of the main challenges that the proposed methodology addresses. To this end we develop novel methodology along with theoretical guarantees for the map from the Euclidean low-dimensional representation back to the object space, which then makes it possible to predict the object corresponding to any element in the low-dimensional Euclidean representation space. Moreover, the inverse map keeps track of the underlying geometry of the data generating mechanism for many statistical situations of interest, such as location-scale families when the objects consist of probability distributions. Identifying manifold structure for the case of probability distributions as random objects is of special interest as it enables an empirical parametrization of an observed sample of distributions of unknown nature. In contrast to the classical approach of postulating a parametric family of distributions from the start, with the exponential family as a prominent example (Efron 1978, 2018), we take the opposite approach by seeking a low-dimensional empirical parametrization for a given observed sample of random distributions that have smooth densities, thus deriving the family of distributions from the given data that consist of a sample of distributions.

For the case of functional data as random objects (Wang et al. 2016; Müller 2016) that fall on a low-dimensional manifold, Chen and Müller (2012) proposed a nonparametric regression kernel estimator for the inverse map under the assumption that the estimated ISOMAP representation converges to its target at a given rate. This inverse map required convexity of the underlying space, which may not hold for general objects lying in a metric space. Here we develop an alternative approach that harnesses Fréchet regression (Petersen and Müller 2019), which can be viewed as an implementation of conditional barycenters,

extending the notion of barycenter or Fréchet mean (Fréchet 1948) in arbitrary metric spaces. Specifically, we use Fréchet regression in the manifold learning problem as a tool to construct inverse maps from the low-dimensional Euclidean representation to the object space. For illustrative purposes and some of the theoretical developments, we focus on the case of one-dimensional distributional data, where one observes a sample of probability distributions on the real line. Along the way, we will discuss extensions to Fréchet manifold learning for more general metric spaces.

The main contributions of this paper are as follows: First, we provide methodology with theoretical guarantees for the inverse map from the Euclidean representation to the object space when the general random objects lie in a manifold embedded in the ambient metric space, which is assumed to be isometric in the ambient metric to a subset of Euclidean space. Such a global isometry to a low-dimensional Euclidean space has also been recently considered in Hamm et al. (2023) when working with distributional data but without a construction of an inverse map, which is an essential step for representing distributional and other types of random objects. Under the global isometry assumption, we obtain a pairwise distance matrix from an initial MDS step that leads to an Euclidean distance matrix (EDM) (Gower 1985); if the dimension of the manifold is known, exact recovery of the Euclidean configuration that generates the EDM is possible up to a Euclidean rigid transformation (Young and Householder 1938). obtain rates of convergence in a realistic setting where the underlying dimension is not known and needs to be estimated from the data.

Second, we demonstrate that the proposed method allows for consistent prediction of the object corresponding for any point lying in the convex hull of the observed Euclidean MDS components and derive rates of convergence for this prediction. This is a highly desirable property for the inverse map as the MDS representation is only available at the sample level and therefore prediction of corresponding objects for any other elements the Euclidean space is challenging.

Third, when the objects correspond to distributional data, i.e., one-dimensional probability distributions in 2-Wasserstein space, we show through real data applications as well as numerical experiments that the proposed inverse map enables interpretation of the effect of each Euclidean MDS component on the distributional response objects.

Fourth, we provide convergence results with rates of convergence for the inverse map for the case where the distributional data are not fully observed but instead one has only a sample of observations generated by each of the distributions, as is usually the case in practice. For this we require increasing sample sizes across the distributions. This scenario is challenging, as the pairwise distances between the distributional objects are then only approximately known. There is a connection with the literature on perturbations of MDS, which have been studied previously for sensor network data in the presence of measurement error (Oh et al. 2010; Javanmard and Montanari 2013). For random objects that correspond to distributional data, we also provide an extension of our methodology to a regression framework, connecting either scalar or distributional responses with distributional predictors, based on a possibly multivariate multiple linear regression utilizing latent parametrizations of the predictors. We also derive convergence results for this regression approach.

The main approach and consistency results are presented in Section 2 for the general case of arbitrary random objects. In Section 3 we provide specifics for the case of distributional objects and also additional theory for this special case. In Section 4 we report simulation results and illustrate our methods with distributions of baby names over calendar years and distribution of bike rentals over the 24 hours of a day in Section 5. Auxiliary lemmas and proofs are in the appendix.

2. Fréchet Manifold Learning

2.1 Preliminaries

Our starting point is a m-dimensional random vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T \in \Theta \subset \mathbb{R}^m$, $m \geq 1$, with distribution F, where Θ is compact, and a totally bounded separable metric space (Ω, d) with metric d that consists of random objects $\nu \in \Omega$ with metric d. These two spaces are connected by a map $\psi : \Theta \to \Omega$, where $\mathcal{M} = \psi(\Theta) \subset \Omega$ is the image of ψ. The map ψ induces a probability distribution on $\mathcal{M}$ as it pushes forward the distribution F on Θ to a distribution F_ν on $\mathcal{M}$. We assume that the mean and covariance matrix of $\boldsymbol{\theta}$ with respect to F are well defined, $\boldsymbol{\mu} = E(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{\theta})$, with $\boldsymbol{\Sigma}$ being positive definite. For general metric spaces Ω and smooth maps ψ, $\mathcal{M}$ is a Riemannian manifold within the ambient space Ω.

Consider a sample of latent parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \overset{iid}{\sim} \boldsymbol{\theta}$ with $\nu_i = \psi(\boldsymbol{\theta}_i)$, $\nu_i \in \Omega$, with $(\boldsymbol{\theta}_1, \nu_1), \ldots, (\boldsymbol{\theta}_n, \nu_n) \overset{iid}{\sim} (\boldsymbol{\theta}, \nu)$. We assume that a sample of n random objects $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ is observed, while the underlying latent parameters $\boldsymbol{\theta}_i$ are unknown. Given the sample $\boldsymbol{\nu} \in \Omega^n$, we obtain low-dimensional global ISOMAP MDS representations of the objects ν_i and then aim at mapping back any element in the low-dimensional representation space. For this, we propose to utilize a suitable notion of regression function from Euclidean space to object space (Ω, d) as inverse map. For the special case where the random objects ν_i are distributions with a suitable metric in the space of distributions, an additional difficulty is that it is usually not realistic to assume that the probability distributions are known but instead one typically has samples of data that are generated by the distributions, where each distribution generates one sample. The theory in Section 4 covers this case that is important for real-world scenarios.

Our basic premise is that the random objects lie on a manifold in the ambient Ω space $\mathcal{M} = \{\nu_{\boldsymbol{\theta}_0} = \psi(\boldsymbol{\theta}_0) \in \Omega : \boldsymbol{\theta}_0 \in \Theta\}$. Given the sample $\boldsymbol{\nu} \in \Omega^n$, we adopt well-known dimension reduction methods to obtain a corresponding parametric representation of sample objects ν_i via MDS and apply global Fréchet regression to provide a suitable map back to object space, which then allows to obtain interpolated random objects at all Euclidean low-dimensional representations. The latter is a key difficulty as one often only has the pair of estimated MDS components and corresponding random objects. Interpolation of these data is challenging when (Ω, d) is not a linear vector space. Assuming for the moment that parameters $\boldsymbol{\theta}_i$ are actually observed in addition to the ν_i, one may view the pairs $(\boldsymbol{\theta}_i, \nu_i)$ as a form of scatterplot data with Euclidean vectors as predictors and objects lying in a metric space as responses. This motivates the application of the framework of global Fréchet regression (Petersen and Müller 2019) to obtain an inverse map back from the Euclidean representation space to the object space.

A key difficulty is that the predictors $\boldsymbol{\theta}_i$ are not observed but instead one has a proxy variable in the form of MDS components, which are well known not to be consistent for the true parameters $\boldsymbol{\theta}_i$ even if the underlying dimension m is known. However, we demonstrate that the linear structure embedded in global Fréchet regression allows to overcome this challenge, so that the estimated inverse map into object space indeed tracks the underlying geometry. When (Ω, d) is the 2-Wasserstein space of probability measures this covers location-scale models and also distributional data that lie on a geodesic in 2-Wasserstein space. In the context of manifold learning, it is often assumed that the target manifold is isometric (in the ambient metric) to a subset of Euclidean space (Javanmard and Montanari 2013; Hamm et al. 2023) that typically is of low dimension. The proposed framework does not pertain to general curved manifolds since the shortest path geodesic estimation that is a core step in ISOMAP is not considered. For our theoretical results, we adopt this framework and assume that the set $\mathcal{M}$ is isometric in the ambient metric to an invertible linear transformation of $\Theta \subset \mathbb{R}^m$ with respect to the ambient distance d as per the following condition, where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^m$.

(A1) There exists an invertible matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ such that for any $\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2 \in \Theta$ and corresponding random objects $\psi(\boldsymbol{\vartheta}_1) = \nu_{\boldsymbol{\vartheta}_1}, \psi(\boldsymbol{\vartheta}_2) = \nu_{\boldsymbol{\vartheta}_2} \in \mathcal{M}$ it holds that $d(\nu_{\boldsymbol{\vartheta}_1}, \nu_{\boldsymbol{\vartheta}_2}) = \|\mathbf{A}(\boldsymbol{\vartheta}_1 - \boldsymbol{\vartheta}_2)\|_2$.

We remark that this assumption is equivalent to $\mathcal{M}$ being isometric (in the ambient metric) to the subset of Euclidean space given by $\mathbf{A}\Theta = \{\mathbf{A}\boldsymbol{\vartheta} : \boldsymbol{\vartheta} \in \Theta\}$ and corresponds to the isometry condition for the map ψ when $\mathbf{A} = \mathbf{I}_m$ is the $m \times m$ identity matrix so that Assumption (A1) encompasses the isometry condition as a special case. Assumption (A1) holds for several important classes of statistical families of probability distributions in 2-Wasserstein space as we demonstrate in Section 3. For linearly transformed random parameter vectors $\tilde{\boldsymbol{\theta}} = \mathbf{A}\boldsymbol{\theta}$ the isometry condition in the ambient space directly holds and the covariance of the transformed vectors is $\boldsymbol{\Sigma}_0 = \mathrm{Cov}(\tilde{\boldsymbol{\theta}}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$.

2.2 Global Fréchet Regression for Manifold Learning

To construct the map ψ from the object space to the Euclidean representation space we adopt classical MDS (Hamm et al. 2023), where we utilize the distance d in the object space Ω to obtain the distance matrix $\mathbf{D}_n$ with elements $[\mathbf{D}_n]_{ij} = d^2(\nu_i, \nu_j)$, $1 \le i, j \le n$. Then for a given dimension $r \in \{1, \ldots, n\}$, classical MDS aims at a lower-dimensional representation $\boldsymbol{\Psi}_r = (\boldsymbol{\eta}_1 \ldots \boldsymbol{\eta}_n) \in \mathbb{R}^{r\times n}$ such that the strain measure (Borg and Groenen 2005) is minimized,

$$\boldsymbol{\Psi}_r = \mathrm{argmin}_{\tilde{\boldsymbol{\Psi}}\in\mathbb{R}^{r\times n}} \ \|\tilde{\boldsymbol{\Psi}}^T\tilde{\boldsymbol{\Psi}} - \frac{1}{2}\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n\|_F, \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm and $\mathbf{J}_n = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$, where $\mathbf{I}_n$ is the identity matrix and $\mathbf{1}_n = (1, \ldots, 1)^T \in \mathbb{R}^n$.

It is well known that when the distance matrix $\mathbf{D}_n$ is Euclidean, namely when there exists a point configuration $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^{r^*}$ for some positive integer r^* such that $[\mathbf{D}_n]_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_2^2$, the rank of $\mathbf{D}_n$ is at most $r^* + 2$ independently of n (Gower 1985), and that

$\mathbf{D}_n$ is Euclidean if and only if the matrix $-(1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n$ is positive semi-definite (Young and Householder 1938). The optimization problem (1) does not admit a unique solution, since for any orthogonal matrix $\mathbf{R}_r \in \mathbb{R}^{r\times r}$, it holds that $\mathbf{R}_r\mathbf{\Psi}_r$ is another solution. If we knew the dimension r^* and then obtained MDS components via (1), a well known result is that the point configuration $\mathbf{z}_1, \ldots, \mathbf{z}_n$ can indeed be recovered up to an Euclidean rigid transformation (Young and Householder 1938), i.e., up to translation, rotation and reflections, as these transformations do not affect the pairwise distances (Dokmanic et al. 2015). In what follows, denote by $\lambda_1 \geq \cdots \geq \lambda_n$ the ordered eigenvalues of $-(1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n$.

Under the isometry condition (A1), it is readily seen that $\mathbf{D}_n$ is an Euclidean matrix with $r^* = m$, so that recovery of the underlying parametrization up to rigid transformations is in principle achievable. While in engineering problems such as sensor network location (Drineas et al. 2006) the embedded spatial dimension r^* is known in advance, the dimension is rarely known in manifold learning contexts, and therefore $r^* = m$ must be obtained in a data-adaptive way. Lemma 2 in the supplement shows that under mild assumptions on the covariance matrix $\mathbf{\Sigma}_0$, the estimate $\tilde{m}$ of m given by $\tilde{m} = \sup\{l = 1, \ldots, n \colon \lambda_l > 0\}$ converges to the true underlying dimension m, where we note that the eigenvalues λ_j are random.

The standard way of obtaining a solution of (1) is via spectral decomposition (Borg and Groenen 2005). Under (A1), we have $(-1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q} = (\boldsymbol{\nu}_1 \cdots \boldsymbol{\nu}_n) \in \mathbb{R}^{n\times n}$ contains eigenvectors $\boldsymbol{\nu}_i \in \mathbb{R}^n$ with corresponding ordered non-negative eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$. Then $\mathbf{\Psi}_r = \mathbf{\Lambda}_r^{1/2}\mathbf{Q}_r^T$, where $\mathbf{Q}_r = (\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_r) \in \mathbb{R}^{n\times r}$ and $\mathbf{\Lambda}_r = \operatorname{diag}(\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^{r\times r}$. Finally, the MDS components $\boldsymbol{\eta}_i$ are obtained from the columns of $\mathbf{\Psi}_r$ by setting $r = \tilde{m}$ as the estimated dimension.

The construction of the MDS representation $\boldsymbol{\eta}_i$ depends on the entire random sample $\nu_1, \ldots, \nu_n$ and is not defined for a generic point $\boldsymbol{\eta}$ in MDS space. This poses challenges for the construction of an inverse map from MDS back to object space, especially for points $\boldsymbol{\eta} \in \mathbb{R}^m$ that are not included among the MDS components $\boldsymbol{\eta}_i$. Some approaches were developed to overcome this problem in the framework of functional data by employing local constant fitting methods implemented with Nadaraya–Watson kernel type estimators to map back to function space, however these approaches rely on the vector space structure of L^2 (Chen and Müller 2012) and therefore cannot be generalized to cover more general object spaces.

For general object spaces, we therefore adopt here a different approach by utilizing the global Fréchet regression framework (Petersen and Müller 2019). Global Fréchet regression generalizes classical multivariate linear regression for the case where responses are situated in a metric space that satisfies certain entropy conditions. The global Fréchet regression function is defined as a weighted Fréchet mean,

$$\nu_{\oplus}(\boldsymbol{\theta}_0) = \operatorname*{arg\,min}_{w\in\Omega} E(s(\boldsymbol{\theta}, \boldsymbol{\theta}_0)d^2(\nu_{\boldsymbol{\theta}}, w)), \tag{2}$$

where $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^m$ and the global weights are given by

$$s(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = 1 + (\boldsymbol{\theta} - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{\theta}_0 - \boldsymbol{\mu}).$$

We note that this approach includes negative weights and is not subject to the curse of dimensionality, in contrast to local approaches. Here m is the reduced dimension and is of

modest size, typically effective dimension reduction means that $1 \leq m \leq 4$ and therefore the inversion of $\boldsymbol{\Sigma}$ is not expected to be problematic. Uniqueness of the minimizer in (2) and convergence of empirical estimates depend intrinsically on geometric and entropy properties of the metric space (Ω, d). For many metric spaces of interest, including those of one-dimensional probability distributions in 2-Wasserstein space and correlation matrices with various metrics, it has been shown that uniqueness holds and convergence rates for empirical estimators have been derived (Petersen and Müller 2019).

2.3 Consistency

We show here that one can consistently recover the global Fréchet regression function $\nu_{\oplus}(\boldsymbol{\theta}_0)$ at all unobserved parameter vectors $\boldsymbol{\theta}_i = \psi^{-1}(\nu_i)$ by employing the MDS representation components $\boldsymbol{\eta}_i$. Even if the true dimension m were known, at which there is an isometry in the ambient space metric between the MDS components and the random objects, the $\boldsymbol{\eta}_i$ differ from the $\boldsymbol{\theta}_i$ by an unknown and random Euclidean rigid transformation. If one had in hand the latent random parameter values $\boldsymbol{\theta}_i$, an oracle estimate $\hat{\nu}_{\oplus}(\cdot, \boldsymbol{\theta}_0)$ of $\nu_{\oplus}(\boldsymbol{\theta}_0)$ would be obtained by adopting the empirical version of global Fréchet regression (Petersen and Müller 2019), given by

$$\hat{\nu}_{\oplus}(\cdot, \boldsymbol{\theta}_0) = \underset{w \in \Omega}{\arg\min}\ n^{-1} \sum_{i=1}^{n} s_{in}(\boldsymbol{\theta}_0) d^2(\nu_i, w), \tag{3}$$

where the empirical weights s_{in} are defined as

$$s_{in}(\boldsymbol{\theta}_0) = 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}),$$

with $\bar{\boldsymbol{\theta}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T$.

Since the $\boldsymbol{\theta}_i$ remain unobserved, (3) is unavailable. However, conditional on the observed sample ν_i, the MDS components $\boldsymbol{\eta}_i$ are available and in the proposed approach we substitute them for the unknown $\boldsymbol{\theta}_i$. This leads to the empirical estimates $\tilde{\nu}_{\oplus}(\boldsymbol{\eta})$ given by

$$\tilde{\nu}_{\oplus}(\boldsymbol{\eta}) = \underset{w \in \Omega}{\arg\min}\ n^{-1} \sum_{i=1}^{n} \tilde{s}_{in}(\boldsymbol{\eta}) d^2(\nu_i, w), \tag{4}$$

where $\boldsymbol{\eta} \in \mathbb{R}^r$, using an estimated dimension $r = \tilde{m}$. The modified global weights $\tilde{s}_{in}$ are

$$\tilde{s}_{in}(\boldsymbol{\eta}) = 1 + (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^T \tilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}),$$

with $\bar{\boldsymbol{\eta}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{\eta}_i$ and $\tilde{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^T$. The MDS components $\boldsymbol{\eta}_i$ can be shown to be naturally centered at $\mathbf{0}_n$ (Oh et al. 2010), where $\mathbf{0}_n = (0, \ldots, 0)^T \in \mathbb{R}^n$. Indeed, since $\mathbf{J}_n \mathbf{1}_n = \mathbf{0}_n$, and in view of the spectral decomposition $(-1/2)\mathbf{J}_n \mathbf{D}_n \mathbf{J}_n = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, we have $\boldsymbol{\Lambda}_r^{1/2} \mathbf{Q}_r^T \mathbf{1}_n = \mathbf{0}_r$. Denoting the canonical basis in $\mathbb{R}^n$ by $\mathbf{e}_i$ leads to

$$\bar{\boldsymbol{\eta}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{\Lambda}_r^{1/2} \mathbf{Q}_r^T \mathbf{e}_i = n^{-1} \boldsymbol{\Lambda}_r^{1/2} \mathbf{Q}_r^T \mathbf{1}_n = \mathbf{0}_r.$$

Existence and uniqueness of the Fréchet regression (3) depends on properties of the metric space (Ω, d). If these are satisfied, existence and uniqueness of the MDS based version (4) follows as we show in Theorem 1. This applies for the reconstructed MDS components $\boldsymbol{\eta} = \boldsymbol{\eta}_j$, $j = 1, \ldots, n$, as well as the convex hull of these MDS components. We demonstrate that taking $\boldsymbol{\eta} = \boldsymbol{\eta}_i$ in (4) leads to a consistent estimate of the global Fréchet regression function $\nu_{\oplus}(\boldsymbol{\theta}_0)$ evaluated at the true but unobserved parameter $\boldsymbol{\theta}_i$, $i = 1, \ldots, n$, as well as at any point lying in the convex hull of the set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$. Denote by $\|\cdot\|_1$ the 1-norm in Euclidean space and $\lambda_{01} \geq \cdots \geq \lambda_{0m}$ the ordered eigenvalues of the covariance matrix $\boldsymbol{\Sigma}_0$. We require the following mild regularity condition on $\boldsymbol{\Sigma}_0$.

(A2) The covariance matrix $\boldsymbol{\Sigma}_0$ possesses a positive eigengap $\delta_0 = \min_{j=1,\ldots,m}(\lambda_{0(j-1)} - \lambda_{0j}, \lambda_{0j} - \lambda_{0(j+1)}) > 0$ with $\lambda_{00} = \infty$ and $\lambda_{0(m+1)} = -\infty$.

Theorem 1 *Under (A1), (A2) and the regularity conditions* $(U0) - (U2)$ *in Petersen and Müller (2019), it holds that*

$$\max_{i=1,\ldots,n} d(\tilde{\nu}_{\oplus}(\boldsymbol{\eta}_i), \nu_{\oplus}(\boldsymbol{\theta}_i)) = O_p\left(n^{-1/(2(\alpha'-1))}\right),$$

for any $\alpha' > \alpha$ *with* α *as in (U2). Moreover, for any non-negative weights* $p_j \geq 0$, $j = 1, \ldots, n$, *such that* $\sum_{j=1}^n p_j = 1$, *for convex combinations* $\boldsymbol{\eta}_{\mathbf{p}} = \sum_{j=1}^n p_j \boldsymbol{\eta}_j$ *and* $\boldsymbol{\theta}_{\mathbf{p}} = \sum_{j=1}^n p_j \boldsymbol{\theta}_j$,

$$\sup_{\mathbf{p}=(p_1,\ldots,p_n):\, p_j \geq 0, \|\mathbf{p}\|_1 = 1} d(\tilde{\nu}_{\oplus}(\boldsymbol{\eta}_{\mathbf{p}}), \nu_{\oplus}(\boldsymbol{\theta}_{\mathbf{p}})) = O_p\left(n^{-1/(2(\alpha'-1))}\right).$$

For completeness, conditions (U0)-(U2) are included in the Appendix. Theorem 1 implies that $\sqrt{n}$-convergence rate can be arbitrarily closely achieved for recovering the inverse global Fréchet map ψ^{-1} over the observed MDS components whenever α as in (U2) can be taken as $\alpha = 2$, which is the case for various metric spaces of statistical interest and for distributional data with the 2-Wasserstein metric under a finite dimension condition in addition to $(U0) - (U2)$ or alternatively bypassing $(U0) - (U2)$ and using convergence properties directly in the Hilbert space of which quantile functions form a subset

We remark that a simple alternative approach to obtain an inverse map back to object space could be obtained via a naive nearest neighbor approach, where the predicted inverse map at an unobserved MDS point $\boldsymbol{\eta}$ is then given by the object ν_i for which its MDS representation $\boldsymbol{\eta}_i$ is closest to $\boldsymbol{\eta}$ in Euclidean norm. However, this approach has several drawbacks. A major weakness concerns interpretation, specifically understanding the effect that each individual MDS parameter has on the objects. For example, we could select one MDS component and fix the remaining MDS parameters at their mean value of 0 while varying the selected MDS component (i.e., moving along a line in MDS space) to study the effect on the predicted inverse map back to object space. Then the nearest neighbor approach will generally not be able to disentangle the effect that each MDS parameter has on the random objects, as there will be multiple confounding effects so that of the MDS component of interest cannot be isolated. See Appendix E for a simulation example that provides an illustrative baseline rather than a comprehensive comparison to demonstrate this issue for the case of distributional data.

3. Fréchet-Wasserstein Manifold Learning: Recovering Latent Parameter Families and Geodesics

We now focus on the important special case of one-dimensional distributional data, where a metric that has been shown to be very useful in statistical applications is the 2-Wasserstein distance (Bolstad et al. 2003; Bigot and Klein 2018; Álvarez Esteban et al. 2016), also popular due to its connection with optimal transport (Villani 2003). For two one-dimensional probability measures υ_1 and υ_2 with corresponding CDFs (cumulative distribution functions) F_{υ_1} and F_{υ_2}, this distance is

$$d^2_{\mathcal{W}_2}(\upsilon_1, \upsilon_2) = \int_0^1 (F^{-1}_{\upsilon_1}(t) - F^{-1}_{\upsilon_2}(t))^2 dt, \tag{5}$$

where quantile functions $F^{-1}_{\upsilon_j}(t) = \inf_{s\in\mathbb{R}}\{F_{\upsilon_j}(s) \geq t\}$, $t \in (0,1)$, are the non-decreasing and left-continuous inverses of F_{υ_j}, $j = 1, 2$. Let $(\mathcal{W}_2, d_{\mathcal{W}_2})$ denote the 2-Wasserstein space of absolutely continuous probability measures with finite second moments over $\mathbb{R}$, which is endowed with the 2-Wasserstein metric $d = d_{\mathcal{W}_2}$, and consider a totally bounded subset $\Omega \subset \mathcal{W}_2$. This metric leads to the corresponding MDS-based forward map from distribution space to Euclidean space as defined in Section 2.2, which we adopt in the following and which has been recently termed Wassmap (Hamm et al. 2023; Cloninger et al. 2025).

In the distributional metric space $(\Omega, d_{\mathcal{W}_2})$, assumption (A1) holds for translation manifolds (Hamm et al. 2023), i.e. when there exists an absolutely continuous probability measure $\mu_0 \in \mathcal{W}_2$ that serves as a template measure such that for all $\nu_{\vartheta_0} \in \mathcal{M}$, where $\vartheta_0 = \boldsymbol{\theta}_0^T \mathbf{e}_1 \in \Theta \subset \mathbb{R}$ and $m = 1$, it holds that $\nu_{\vartheta_0}(\cdot) = \mu_0(\cdot - \vartheta_0)$. Similarly, for dilation manifolds parametrized by a scaling parameter $\sigma > 0$, namely when the space $\Theta \subset \mathbb{R}^+$ is contained in the positive orthant and there exists an absolutely continuous template measure $\mu_0 \in \mathcal{W}_2$ such that $\nu_{\vartheta_0}(\cdot) = \mu_0(\cdot/\sigma)$, where $\vartheta_0 = \sigma$, following Lemma 3.7 in Hamm et al. (2023) one may take $A = \sqrt{\mathcal{M}_2(\mu_0)} > 0$ in (A1), where $\mathcal{M}_2(\mu_0) < \infty$ denotes the second moment of μ_0.

Another important class of distributions that satisfy (A1) and encompases the two previous cases are location-scale families, for which there exists an absolutely continuous probability measure $\mu_0 \in \mathcal{W}_2$ with corresponding density function $f_0 > 0$ a.e. satisfying $\int_{\mathbb{R}} x f_0(x) dx = 0$ and $\int_{\mathbb{R}} x^2 f_0(x) dx = 1$, and a two-dimensional parameter $\boldsymbol{\theta} = (\mu, \sigma)^T \in \mathbb{R} \times \mathbb{R}^+$ such that $\nu_{\boldsymbol{\theta}} = \nu_{(\mu,\sigma)}$ has density

$$f_{\nu|\boldsymbol{\theta}}(\cdot, \mu, \sigma) = \frac{1}{\sigma} f_0\left(\frac{\cdot - \mu}{\sigma}\right).$$

Since the 2-Wasserstein distance between two one-dimensional probability measures $w_1, w_2 \in \mathcal{W}_2$ is the $L^2(0,1)$ distance between corresponding quantile functions (5), when w_1 and w_2 lie in the location-scale class, i.e. for some $\tilde{\boldsymbol{\theta}}_j = (\tilde{\mu}_j, \tilde{\sigma}_j) \in \Theta \subset \mathbb{R} \times \mathbb{R}^+$ it holds that $f_{w_j}(\cdot) = f_{\nu|\boldsymbol{\theta}}(\cdot, \tilde{\mu}_j, \tilde{\sigma}_j)$, $j = 1, 2$, the pairwise distance admits a simple well-known form (Panaretos and Zemel 2019)

$$d^2_{\mathcal{W}_2}(w_1, w_2) = (\tilde{\mu}_1 - \tilde{\mu}_2)^2 + (\tilde{\sigma}_1 - \tilde{\sigma}_2)^2.$$

Hence, considering the case $m = 2$ and a subclass where $(\mu, \sigma) \in \Theta = \Theta_1 \times \Theta_2$ with $\Theta_1 \subset \mathbb{R}$ and $\Theta_2 \subset \mathbb{R}^+$ are compact, (A1) is satisfied by taking the identity matrix $\mathbf{A} = \mathbf{I}_2$.

We remark here that location-scale families also cover the important statistical problem of time warping and deformation models (Bigot and Charlier 2011; Panaretos and Zemel 2019), where the random densities are assumed to be random warps or deformations with respect to some fixed template density. A simple warping model is $g(\cdot) = f_0(\cdot - \theta)$, where θ is a scalar random variable and f_0 is some template density function.

Following up on Theorem 1 in the distributional case, the entire 2-Wasserstein space $(\mathcal{W}_2, d_{\mathcal{W}_2})$ (Petersen and Müller 2019) is not totally bounded, which is a basic requirement to apply the convergence results, and therefore these conditions need to be checked for the subset $\Omega \subset \mathcal{W}_2$. This hinges on verifying assumption (U0), which entails existence and uniqueness of (2) and (3) for all $\|\boldsymbol{\theta}_0\|_2 \leq \mathrm{Diam}(\Theta) < \infty$; assumption (U2) is established analogously as in Proposition 1 in Petersen and Müller (2019), where the constant α can be taken as $\alpha = 2$. For assumption (U1), the assumptions listed in Proposition 1 in Petersen and Müller (2019) are insufficient and need to be complemented by assuming that the space Ω is finite-dimensional, which is the case for our purposes due to the finite-dimensional manifold representation. Alternatively, one can bypass assumption (U1) altogether by directly adopting the convergence results of Petersen et al. (2021).

Condition (U0) holds when $\Omega \subset \mathcal{W}_2$ is such that its corresponding quantile space $\mathcal{Q}_\Omega$ consisting of the left-continuous and non-decreasing generalized inverses of the corresponding CDFs is a closed and convex subset of $L^2(0,1)$. This is a consequence of the orthogonal projection theorem in the Hilbert space $L^2(0,1)$ and in view of the quantile representation of $d_{\mathcal{W}_2}$ (5), provided that the quantile space $\mathcal{Q}_\Omega$ consists of continuous functions. Indeed, if for any probability measure $\nu_0 \in \Omega$ its CDF F_{ν_0} is strictly increasing on the interval $\{s \in \mathbb{R} \colon 0 < F_{\nu_0}(s) < 1\}$, the quantile function space $\mathcal{Q}_\Omega$ is a subset of the space $\mathcal{C}(0,1)$ of continuous functions over $(0,1)$ (see e.g. Proposition A.7 in Bobkov and Ledoux (2019)).

(A3) Suppose that $(\Omega, d_{\mathcal{W}_2})$ is totally bounded in 2-Wasserstein space and the quantile space $\mathcal{Q}_\Omega \subset \mathcal{C}(0,1)$ is closed and convex when viewed as a subset of the Hilbert space $L^2(0,1)$.

We then have the following result on the recovery of the global Fréchet-Wasserstein inverse map over the convex hull of the observed MDS components.

Corollary 1 *Under (A1) – (A3), for any $\gamma > 0$ it holds that*

$$\max_{i=1,\ldots,n} d(\tilde{\nu}_\oplus(\boldsymbol{\eta}_i), \nu_\oplus(\boldsymbol{\theta}_i)) = O_p\left(n^{-1/(2+\gamma)}\right).$$

Moreover, for any non-negative weights $p_j \geq 0$, $j = 1, \ldots, n$, such that $\sum_{j=1}^n p_j = 1$, and denoting the convex combinations $\boldsymbol{\eta}_\mathbf{p} = \sum_{j=1}^n p_j \boldsymbol{\eta}_j$ and $\boldsymbol{\theta}_\mathbf{p} = \sum_{j=1}^n p_j \boldsymbol{\theta}_j$, it holds that

$$\sup_{\mathbf{p}=(p_1,\ldots,p_n)\colon p_j \geq 0, \|\mathbf{p}\|_1 = 1} d(\tilde{\nu}_\oplus(\boldsymbol{\eta}_\mathbf{p}), \nu_\oplus(\boldsymbol{\theta}_\mathbf{p})) = O_p\left(n^{-1/(2+\gamma)}\right).$$

The following examples showcase situations in which the global Fréchet regression function admits a closed form and recovers exactly the underlying distributions along with the latent generating family such as the location-scale family. Example 1 below concerns the case where the location and scale parameters depend linearly on two independent random variables in a way that the scale parameter is almost surely positive. We omit the detailed derivation as it follows from arguments of Petersen et al. (2021).

Example 1 *Suppose $m = 2$, $\Theta \subset \mathbb{R} \times \mathbb{R}^+$ and let $a_0, a_1, b_1, b_2 \in \mathbb{R}$ be constants such that $b_1 + b_2\boldsymbol{\theta}^T\mathbf{e}_2 > 0$ almost surely for all $\boldsymbol{\theta} = (\vartheta_1, \vartheta_2)^T$ in the support of $F_{\boldsymbol{\theta}}$. Consider the location-scale model,*

$$f_{\nu|\boldsymbol{\theta}}(\cdot, \boldsymbol{\theta}) = \frac{1}{b_1 + b_2\vartheta_2} f_0\left(\frac{\cdot - a_0 - a_1\vartheta_1}{b_0 + b_1\vartheta_2}\right),$$

where $f_0 > 0$ is a density function over $\mathbb{R}$ such that $\int_{\mathbb{R}} s f_0(s)ds = 0$ and $\int_{\mathbb{R}} s^2 f_0(s)ds = 1$, and suppose that $Cov(\vartheta_1, \vartheta_2) = 0$. For simplicity consider $\Omega = \mathcal{W}_2$. Then, the density $f_\oplus(\cdot, \boldsymbol{\theta}_0)$ of the global Fréchet regression function (2) satisfies

$$f_\oplus(\cdot, \boldsymbol{\theta}_0) = f_{\nu|\boldsymbol{\theta}}(\cdot, \boldsymbol{\theta}_0),$$

which means that it recovers the underlying latent location-scale structure.

Example 2 *Suppose $m = 1$, $\Theta \subset \mathbb{R}$, and $\boldsymbol{\theta} = \vartheta \in \Theta$. Let $a_0, a_1 \in \mathbb{R}$ be constants such that $a_0 + a_1\vartheta > 0$ holds for all ϑ in the support of $F_{\boldsymbol{\theta}}$. Consider the exponential regression model $\nu_{\boldsymbol{\theta}} \sim Exp(m(\vartheta))$ with $m(\vartheta) = (a_0 + a_1\vartheta)^{-1}$ and suppose $\Omega = \mathcal{W}_2$. Thus the underlying latent family $\nu_{\boldsymbol{\theta}}$ has an exponential distribution with a scale that depends linearly on ϑ. Then, the global Fréchet regression function (2) corresponds to $Exp((a_0 + a_1\vartheta_0)^{-1})$, where $\boldsymbol{\theta}_0 = \vartheta_0 \in \Theta$, i.e., it recovers the underlying latent structure.*

The conclusion in Example 2 follows from $E(s(\boldsymbol{\theta}, \boldsymbol{\theta}_0)Q_{\boldsymbol{\theta}}(t)) = -\log(1-t)(a_0 + b_0\theta_0)$, $t \in (0, 1)$, which is a valid exponential quantile function and as the quantile function of (2) equals the L^2-orthogonal projection of $E(s(\boldsymbol{\theta}, \boldsymbol{\theta}_0)Q_{\boldsymbol{\theta}}(t))$ into quantile space. Here $Q_{\boldsymbol{\theta}}$ is the quantile function corresponding to $\nu_{\boldsymbol{\theta}}$.

Another statistical situation of interest corresponds to the case where probability measures ν_i, $i = 1, \ldots, n$, lie on a geodesic in 2-Wasserstein space between two absolutely continuous probability measures μ_0 and μ_1 with finite second moments (Fan and Müller 2025). For this, let $m = 1$ and $\boldsymbol{\theta}_i = \vartheta_i \in \Theta = [0, 1]$, $i = 1, \ldots, n$. Suppose that there exists a unique geodesic $\mu(t)$, $t \in [0, 1]$, such that $\mu(0) = \mu_0$, $\mu(1) = \mu_1$, and $\nu_i = \mu(\vartheta_i)$, where the ϑ_i may be regarded as random time locations. Theorem 1 in Fan and Müller (2025) shows that the global Fréchet regression function in (2) recovers the geodesic in the sense $\nu_\oplus(\boldsymbol{\theta}_0) = \mu(\boldsymbol{\theta}_0)$ is achieved, for all $\boldsymbol{\theta}_0 = \vartheta_0 \in \Theta$. By construction of the geodesic, one has $d_{\mathcal{W}_2}(\mu(\tilde{\vartheta}_1), \mu(\tilde{\vartheta}_2)) = |\tilde{\vartheta}_1 - \tilde{\vartheta}_2| d_{\mathcal{W}_2}(\mu_0, \mu_1)$ for any $\tilde{\vartheta}_1, \tilde{\vartheta}_2 \in \Theta$ so that (A1) is satisfied. We assume that the geodesic $\mu(t)$, $t \in [0, 1]$, lies in the distribution space Ω.

Corollary 2 *Under (A1) – (A3), suppose that for two distinct probability measures $\mu_0, \mu_1 \in \Omega$ there exists a unique geodesic $\mu(t)$, $t \in [0, 1]$, contained in Ω such that $\mu(0) = \mu_0$, $\mu(1) = \mu_1$, and the sample of probability measures $\nu_i = \mu(\vartheta_i)$, $i = 1, \ldots, n$, with $\boldsymbol{\theta}_i = \vartheta_i \in \Theta = [0, 1]$ is observed, where almost surely all ϑ_i are distinct. Then, for any $\gamma > 0$,*

$$\sup_{\mathbf{p}=(p_1,\ldots,p_n):\, p_j \geq 0, \|\mathbf{p}\|_1 = 1} d_{\mathcal{W}_2}(\tilde{\nu}_\oplus(\boldsymbol{\eta}_{\mathbf{p}}), \mu(\vartheta_{\mathbf{p}})) = O_p\left(n^{-1/(2+\gamma)}\right),$$

where $\vartheta_{\mathbf{p}} = \sum_{j=1}^n p_j\vartheta_j$ and $\boldsymbol{\eta}_{\mathbf{p}} = \sum_{j=1}^n p_j\boldsymbol{\eta}_j$. If furthermore the distribution of $\boldsymbol{\theta} = \vartheta$ has a density f_ϑ that is continuous and strictly positive over Θ with CDF $F_\vartheta(s) = \int_0^s f_\vartheta(u)du$,

taking any $\rho_1 > 0, \rho \in (0,1)$ *and the sequence* $\epsilon_n = 1 - F_\vartheta^{-1}(1-n^{-\rho}) = o(1)$ *as* $n \to \infty$*, the event*

$$[\epsilon_n, 1-\epsilon_n] \subseteq [\vartheta_{(1)}, \vartheta_{(n)}] = \Big\{ \vartheta_{\mathbf{p}} = \sum_{j=1}^{n} p_j \vartheta_j \colon p_1, \ldots, p_n \geq 0, \|\mathbf{p}\|_1 = 1 \Big\},$$

occurs with probability at least $1 - n^{-\rho_1}$ *as* $n \to \infty$*. Here* $\vartheta_{(1)}$ *and* $\vartheta_{(n)}$ *are the first and last order statistics from the sample* $\vartheta_1, \ldots, \vartheta_n$*.*

Corollary 2 shows that the geodesic path $\mu(t)$, $t \in (\epsilon_n, 1-\epsilon_n)$, can be uniformly recovered in the 2-Wasserstein metric by employing the MDS components, where the interval $(\epsilon_n, 1-\epsilon_n)$ asymptotically grows to the entire time window $[0,1]$, and the uniform recovery over $(\epsilon_n, 1-\epsilon_n)$ occurs with a probability that converges to 1 at any polynomial rate. We remark that here the convex hull is equivalent to a standard coverage interval (Wilks 1948) employing the first and last order statistics.

4. Fréchet-Wasserstein Manifold Learning Under Uncertainty

The previous developments require knowledge of the exact true pairwise distances $[\mathbf{D}_n]_{ij} = d^2_{\mathcal{W}_2}(\nu_i, \nu_j)$, $i,j = 1, \ldots, n$, which would be feasible if for example the density functions f_i corresponding to ν_i were fully observed. However, this may not be the case in practice. Instead, it is often more common to observe an increasing sample of scalars $Y_{i1}, \ldots, Y_{in_i} \overset{iid}{\sim} f_i$ with $n_i \to \infty$ (Petersen and Müller 2016), which is then employed to estimate the required pairwise distances $[\mathbf{D}_n]_{ij}$ via a suitable $[\hat{\mathbf{D}}_n]_{ij}$. This introduces further estimation errors that need to be accounted for in the final empirical estimates for (2). Writing $[\hat{\mathbf{D}}_n]_{ij} = [\mathbf{D}_n]_{ij} + \epsilon_{ij}$, where $\epsilon_{ij} = [\hat{\mathbf{D}}_n]_{ij} - [\mathbf{D}_n]_{ij}$ is the estimation error, reveals a connection with the recent literature on perturbation of MDS components (Oh et al. 2010; Javanmard and Montanari 2013) where a characteristic of the situation we consider here is that the estimation error can be controlled and shown to be diminishing as the number of observations per density $n_i = n_i(n)$ diverges with sample size n.

Suppose that conditional on $\boldsymbol{\theta}_i$, $i = 1, \ldots, n$, an independent random mechanism generates a sample $Y_{i1}, \ldots, Y_{in_i} \overset{iid}{\sim} \nu_i$, where $n_i \geq N(n) \to \infty$ as $n \to \infty$ and ν_i or equivalently f_i remains unobserved. We can then employ these samples to obtain estimates of $[\mathbf{D}_n]_{ij}$; this can be achieved using empirical quantile functions. The estimated pairwise distance is given by $[\hat{\mathbf{D}}_n]_{ij} = \int_0^1 (\hat{Q}_i(t) - \hat{Q}_j(t))^2 dt$, where $\hat{Q}_i(t) = \inf_{s \in \mathbb{R}} \{\hat{F}_i(s) \geq t\}$, $t \in (0,1)$, is the non-decreasing and left-continuous generalized inverse of the empirical CDF $\hat{F}_i$ given by $\hat{F}_i(y) = n_i^{-1} \sum_{j=1}^{n_i} 1_{\{Y_{ij} \leq y\}}$, $y \in \mathbb{R}$. The previous increasing sample framework is then necessary to obtain consistent estimates of the unobserved pairwise distances. We formalize this requirement in the following condition; see also Petersen and Müller (2016).

(A4) Suppose that conditionally on $\boldsymbol{\theta}_i$, $Y_{i1}, \ldots, Y_{in_i} \overset{iid}{\sim} f_i$, where $n_i \geq N(n)$ holds for all $i = 1, \ldots, n$, and the positive-integer sequence $N(n)$ satisfies $N(n) \to \infty$ as $n \to \infty$.

Let $\hat{\mathbf{D}}_n$ be the $n \times n$ matrix with entries $[\hat{\mathbf{D}}_n]_{ij}$, $i,j = 1, \ldots, n$. The estimated MDS components $\hat{\boldsymbol{\eta}}_i$ are given by $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\Lambda}}_{\hat{m}}^{1/2} \hat{\mathbf{Q}}_{\hat{m}}^T \mathbf{e}_i$, where the $\mathbf{e}_i$ are the canonical basis of $\mathbb{R}^n$ and

$\hat{m}$ is a suitable estimate of the underlying dimension m (see Theorem 2 and Corollary 3 below). Here $\hat{\mathbf{\Lambda}}_{\hat{m}}$ and $\hat{\mathbf{Q}}_{\hat{m}}$ come from the spectral decomposition $(-1/2)\mathbf{J}_n\hat{\mathbf{D}}_n\mathbf{J}_n = \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^T$, where $\hat{\mathbf{Q}} = (\hat{\boldsymbol{\nu}}_1 \cdots \hat{\boldsymbol{\nu}}_n) \in \mathbb{R}^{n\times n}$ contains the eigenvectors $\hat{\boldsymbol{\nu}}_i \in \mathbb{R}^n$ with corresponding ordered eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$, $\hat{\mathbf{\Lambda}} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \in \mathbb{R}^{n\times n}$, $\hat{\mathbf{Q}}_{\hat{m}} = (\hat{\boldsymbol{\nu}}_1 \cdots \hat{\boldsymbol{\nu}}_{\hat{m}}) \in \mathbb{R}^{n\times \hat{m}}$, and $\hat{\mathbf{\Lambda}}_{\hat{m}} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{m}}) \in \mathbb{R}^{\hat{m}\times\hat{m}}$. Here $\hat{\boldsymbol{\nu}}_l^T\boldsymbol{\nu}_l \geq 0$ are assumed aligned, $l = 1, \ldots, \hat{m}$. It is easy to show that there is no loss of generality with this eigenvector alignment condition as the sign of $\hat{\boldsymbol{\nu}}_l$ does not alter the empirical global weights $\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j)$, which are defined by

$$\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) = 1 + (\hat{\boldsymbol{\eta}}_i - n^{-1}\sum_{k=1}^n \hat{\boldsymbol{\eta}}_k)^T \tilde{\hat{\mathbf{\Sigma}}}^{-1}(\hat{\boldsymbol{\eta}}_j - n^{-1}\sum_{k=1}^n \hat{\boldsymbol{\eta}}_k),$$

where $1 \leq i, j \leq n$, and $\tilde{\hat{\mathbf{\Sigma}}}$ is the sample covariance of the $\hat{\boldsymbol{\eta}}_i$. The estimated global Fréchet regression function then becomes

$$\hat{\tilde{\nu}}_\oplus(\boldsymbol{\eta}) = \operatorname*{arg\,min}_{w\in\Omega}\ n^{-1}\sum_{i=1}^n \tilde{\hat{s}}_{in}(\boldsymbol{\eta}) d^2_{\mathcal{W}_2}(\hat{\nu}_i, w), \tag{6}$$

where $\hat{\nu}_i$ is the unique probability measure with quantile function $\hat{Q}_i$ (see e.g. Proposition A.2 in Bobkov and Ledoux (2019)).

Since uniform convergence across all pairwise 2-Wasserstein distances is necessary, we require an additional regularity condition on the quantile space $\mathcal{Q}_\Omega$. Denote by $J_2(\nu_0) = \int_{\mathbb{R}} F_{\nu_0}(x)(1 - F_{\nu_0}(x))/f_{\nu_0}(x)dx$ the J_2 functional (Bobkov and Ledoux 2019), where f_{ν_0} is the density function of $\nu_0 \in \Omega$ and $F_{\nu_0}(s) = \int_{-\infty}^s f_{\nu_0}(u)du$, $s \in \mathbb{R}$, is its CDF.

(A5) Suppose that $(\Omega, d_{\mathcal{W}_2})$ is totally bounded in 2-Wasserstein space with $\sup_{\nu_0\in\Omega} J_2(\nu_0) < \infty$, and the quantile space $\mathcal{Q}_\Omega \subset \mathcal{C}(0,1)$ is closed and convex when viewed as a subset of the Hilbert space $L^2(0,1)$.

Here we remark that the condition $J_2(\nu_0) < \infty$ for $\nu_0 \in \Omega$ in (A5) implies ν_0 is supported on an interval $\mathcal{I}$ in $\mathbb{R}$ with an almost surely positive density f_{ν_0} on its support (Bobkov and Ledoux 2019; Bigot et al. 2018); here the support of ν_0 is the smallest closed subset of $\mathbb{R}$ with ν_0-measure 1 (Bobkov and Ledoux 2019). This implies F_{ν_0} is strictly increasing over $\mathcal{I}$ and hence the quantile function $F^{-1}_{\nu_0}$ is continuous in $(0,1)$. Thus $\mathcal{Q}_\Omega$ is a subset of the space of continuous functions $\mathcal{C}(0,1)$. The quantile space of bi-Lipschitz functions in $[0,1]$ considered in Gajardo and Müller (2022) satisfies the regularity condition (A5) and has the property that Ω is totally bounded since it is compact. Denote by Q_ν the (random) quantile function corresponding to the random probability measure ν.

Theorem 2 *Under (A1), (A2), (A4), and (A5), suppose that there exists $\rho \in (0, 1/3)$ such that $nN^{-\rho} = o(1)$ as $n \to \infty$. Let $\epsilon > 0$ and consider $\hat{m} = \sup\{l = 1, \ldots, n\colon n^{-1}\hat{\lambda}_l \geq nN^{-\rho}\epsilon\}$. If $E(\|Q_\nu\|^2_{L^2(0,1)}) < \infty$, then*

$$n^{-1}\sum_{j=1}^n d_{\mathcal{W}_2}(\hat{\tilde{\nu}}_\oplus(\hat{\boldsymbol{\eta}}_j), \tilde{\nu}_\oplus(\boldsymbol{\eta}_j)) = O_p(n^{3/2}N(n)^{-1/2}),$$

and for any $\gamma > 0$,

$$n^{-1}\sum_{j=1}^{n} d_{\mathcal{W}_2}(\hat{\tilde{\nu}}_{\oplus}(\hat{\boldsymbol{\eta}}_j), \nu_{\oplus}(\boldsymbol{\theta}_j)) = O_p(n^{-1/(2+\gamma)} + n^{3/2}N(n)^{-1/2}).$$

Moreover, for any non-negative weights $p_j \geq 0$, $j = 1, \ldots, n$, *such that* $\sum_{j=1}^{n} p_j = 1$, *and denoting the convex combinations* $\hat{\boldsymbol{\eta}}_{\mathbf{p}} = \sum_{j=1}^{n} p_j \hat{\boldsymbol{\eta}}_j$ *and* $\boldsymbol{\theta}_{\mathbf{p}} = \sum_{j=1}^{n} p_j \boldsymbol{\theta}_j$, *it holds that*

$$\sup_{\mathbf{p}=(p_1,\ldots,p_n)\colon p_j \geq 0, \|\mathbf{p}\|_1 = 1} d_{\mathcal{W}_2}(\hat{\tilde{\nu}}_{\oplus}(\hat{\boldsymbol{\eta}}_{\mathbf{p}}), \nu_{\oplus}(\boldsymbol{\theta}_{\mathbf{p}})) = O_p\left(n^{-1/(2+\gamma)} + n^{3/2}N(n)^{-1/2}\right).$$

If the distributions in Ω have support included in some compact interval (Petersen and Müller 2016), faster convergence rates are available. Various applications where data may be viewed as density functions naturally fall in this category. For example, for age pyramids in demography histograms are available and corresponding density functions are obtained via smoothing methods, where the support is often considered to be $[0, \mathcal{T}]$ for some $\mathcal{T}$ such as $\mathcal{T} = 100$ years (Cazelles et al. 2018). Bike rental pickups over the course of a day at a station can be viewed as event arrival times generated by a temporal point process (Gajardo and Müller 2022), where the support is $[0, 24]$ hours. Another example is house price distributions (Chen et al. 2023).

Corollary 3 below provides a version of Theorem 2 for the case where densities have bounded support. Then a weaker growth rate condition on the lower bound $N(n)$ suffices for consistent recovery of the global Fréchet regression function uniformly across the convex hull of the latent parameters. The dimension estimate $\hat{m}$ is obtained by lower-thresholding the empirical and perturbed eigenvalues $n^{-1}\hat{\lambda}_l$ by a decaying lower bound that depends on how densely sampled the probability distributions ν_i are, as reflected in the growth rate of $N(n)$.

Corollary 3 *Under (A1), (A2), (A4), and (A5), suppose that there exists* $\mathcal{T} > 0$ *such that for any* $\nu_0 \in \Omega$ *its corresponding density* f_{ν_0} *has support contained in* $[0, \mathcal{T}]$. *Let* $\epsilon > 0$, $\rho \in (0, 1/2)$, *and consider* $\hat{m} = \sup\{l = 1, \ldots, n\colon n^{-1}\hat{\lambda}_l \geq N(n)^{-\rho}\epsilon\}$. *Then*

$$n^{-1}\sum_{j=1}^{n} d_{\mathcal{W}_2}(\hat{\tilde{\nu}}_{\oplus}(\hat{\boldsymbol{\eta}}_j), \tilde{\nu}_{\oplus}(\boldsymbol{\eta}_j)) = O_p(N(n)^{-1/2}),$$

and for any $\gamma > 0$,

$$n^{-1}\sum_{j=1}^{n} d_{\mathcal{W}_2}(\hat{\tilde{\nu}}_{\oplus}(\hat{\boldsymbol{\eta}}_j), \nu_{\oplus}(\boldsymbol{\theta}_j)) = O_p(n^{-1/(2+\gamma)} + N(n)^{-1/2}).$$

Moreover, for any non-negative weights $p_j \geq 0$, $j = 1, \ldots, n$, *such that* $\sum_{j=1}^{n} p_j = 1$, *and convex combinations* $\hat{\boldsymbol{\eta}}_{\mathbf{p}} = \sum_{j=1}^{n} p_j \hat{\boldsymbol{\eta}}_j$ *and* $\boldsymbol{\theta}_{\mathbf{p}} = \sum_{j=1}^{n} p_j \boldsymbol{\theta}_j$, *it holds that*

$$\sup_{\mathbf{p}=(p_1,\ldots,p_n)^T\colon p_j \geq 0, \|\mathbf{p}\|_1 = 1} d_{\mathcal{W}_2}(\hat{\tilde{\nu}}_{\oplus}(\hat{\boldsymbol{\eta}}_{\mathbf{p}}), \nu_{\oplus}(\boldsymbol{\theta}_{\mathbf{p}})) = O_p\left(n^{-1/(2+\gamma)} + N(n)^{-1/2}\right).$$

5. Fréchet-Wasserstein Manifold Learning for Distributional Regression

5.1 Scalar on Distribution Regression

First consider the case of a scalar response $Y \in \mathbb{R}$ and a distributional predictor ν as defined in Section 2.1. Suppose that $(\nu_1, Y_1), \ldots, (\nu_n, Y_n) \overset{iid}{\sim} (\nu, Y)$ and assume (A1) holds with $\tilde{\boldsymbol{\theta}}$ the parametrization of ν that achieves the isometry condition on $\mathcal{M}$ as defined in Section 2.1 with $d = d_{\mathcal{W}_2}$. This implies $d_{\mathcal{W}_2}(\nu_i, \nu_j) = \|\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}_j\|_2$. A natural regression model for the response Y on ν can be directly obtained by formulating a linear regression relationship for Y on the underlying parameter $\tilde{\boldsymbol{\theta}}$. For this purpose, suppose that $Y_i = \beta_0 + \boldsymbol{\beta}^T \tilde{\boldsymbol{\theta}}_i + \epsilon_i$ with ϵ_i a measurement error independent of $\tilde{\boldsymbol{\theta}}_i$ and satisfying $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2 < \infty$. Assume for simplicity the distributions ν_i are fully observed.

Define the matrix $\mathbf{M}_n \in \mathbb{R}^{n \times m}$ whose ith row is given by $\boldsymbol{\eta}_i^T$, $i = 1, \ldots, n$, and the design matrix $\mathbf{X}_n = (\mathbf{1}_n \ \mathbf{M}_n) \in \mathbb{R}^{n \times (m+1)}$. Let $P_i = \mathbf{e}_i^T \mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n$ be the ith prediction, $i = 1, \ldots, n$. Here the $\boldsymbol{\eta}_i$ are the MDS components obtained from the distributional objects ν_i with the 2-Wasserstein metric as in Section 2.2. The next result shows that the regression function $E(Y|\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_0)$ can be consistently recovered at sample points $\tilde{\boldsymbol{\theta}}_j$ by employing a linear regression fit on the scatterplot $(\boldsymbol{\eta}_i, Y_i)_{i=1,\ldots,n}$, and also that the squared prediction error across the sample diminishes as $n \to \infty$. Let $\tilde{m}$ be defined as in Section 2.2 and consider the estimated version $\hat{P}_i$ of P_i obtained after replacing m by $\tilde{m}$.

Theorem 3 *Suppose that (A1) and (A2) hold. Consider any fixed $j = 1, \ldots, n$, then*

$$\hat{P}_j = E(Y|\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_j) + O_p(n^{-1/2}),$$

where the bound is uniform in j. Moreover,

$$S_n = n^{-1} \sum_{i=1}^{n} (\hat{P}_i - E(Y|\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_i))^2 = O_p(n^{-1}).$$

5.2 Distribution on Distribution Regression

The regression problem between two univariate probability distributions on the real line has received great interest in the literature; see for example Chen et al. (2023) where a method based on projection to tangent spaces in Wasserstein space followed by a linear functional regression in the L^2 tangent spaces has been developed. Here we propose a simpler straightforward distributional linear regression model that is based on the low-dimensional representation of the response and predictor distributional objects.

Consider a (random) probability distribution ν parametrized by $\boldsymbol{\theta}$ as described in Section 2.1, which serves as predictor. For the response, we consider a second (random) distribution ν' that admits a low-dimensional representation $\boldsymbol{\theta}' \in \mathbb{R}^m$ as follows. Let $\Theta' \subset \mathbb{R}^m$ be compact. A first random mechanism generates an Euclidean vector $\boldsymbol{\theta}' \in \Theta'$ and then, conditional on $\boldsymbol{\theta}' = \boldsymbol{\theta}'_0$, ν' corresponds to an Ω-valued object $\nu' = \nu'_{\boldsymbol{\theta}'_0}$ with conditional distribution $F_{\nu'|\boldsymbol{\theta}'}(\cdot, \boldsymbol{\theta}'_0)$, which is assumed to exist. Consider a sample of latent parameters $\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_n \overset{iid}{\sim} \boldsymbol{\theta}'$ which generate corresponding individual random objects $\nu'_i \in \Omega$, $i = 1, \ldots, n$, so that $(\boldsymbol{\theta}'_1, \nu'_1), \ldots, (\boldsymbol{\theta}'_n, \nu'_n) \overset{iid}{\sim} (\boldsymbol{\theta}', \nu')$. This encapsulates the low-dimensional representation of ν' in terms of $\boldsymbol{\theta}'$. A natural regression approach is then to

connect the low-dimensional representations $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$. For this, suppose that the scatterplot $(\nu_1, \nu_1'), \ldots, (\nu_n, \nu_n') \overset{iid}{\sim} (\nu, \nu')$ of probability distribution objects is available, where the ν_i and ν_i' are assumed fully observed for simplicity; an extension to the case where one has data that are sampled from these distributions is feasible and proceeds in analogy to the development in Section 4. For the predictor distribution ν we require the isometry condition (A1), where we choose $d = d_{\mathcal{W}_2}$ and analogously for the response distribution ν' as per the following condition.

(A1') There exists an invertible matrix $\mathbf{A}' \in \mathbb{R}^{m \times m}$ such that for any $\boldsymbol{\vartheta}_1', \boldsymbol{\vartheta}_2' \in \Theta'$ and corresponding probability measures $\nu'_{\boldsymbol{\vartheta}_1'}, \nu'_{\boldsymbol{\vartheta}_2'} \in \mathcal{M}' = \{\nu'_{\boldsymbol{\theta}_0'} \in \Omega : \boldsymbol{\theta}_0' \in \Theta'\}$ it holds that

$$d_{\mathcal{W}_2}(\nu'_{\boldsymbol{\vartheta}_1'}, \nu'_{\boldsymbol{\vartheta}_2'}) = \|\mathbf{A}'(\boldsymbol{\vartheta}_1' - \boldsymbol{\vartheta}_2')\|_2.$$

Now let $\tilde{\boldsymbol{\theta}}' = \mathbf{A}'\boldsymbol{\theta}'$. Under (A1) and (A1'), we have $d_{\mathcal{W}_2}(\nu_i, \nu_j) = \|\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}_j\|_2$ and $d_{\mathcal{W}_2}(\nu_i', \nu_j') = \|\tilde{\boldsymbol{\theta}}_i' - \tilde{\boldsymbol{\theta}}_j'\|_2$. A natural linear regression model for ν' on ν is to postulate a linear relationship for the regression function $E(\tilde{\boldsymbol{\theta}}'|\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_0)$. For this we adopt a classical multivariate multiple linear regression model (Johnson and Wichern 2007). Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})^T$ be the intercept vector, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \in \mathbb{R}^m$ be slope vectors and define the slope matrix $\mathbf{B}_m = (\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_m) \in \mathbb{R}^{m \times m}$. The linear regression model is given by $\tilde{\boldsymbol{\theta}}' = \boldsymbol{\beta}_0 + \mathbf{B}_m^T\tilde{\boldsymbol{\theta}} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)^T$ is an independent m-dimensional measurement error vector satisfying $E(\boldsymbol{\epsilon}) = \mathbf{0}_m$ and $E(\|\boldsymbol{\epsilon}\|_2^2) = \sigma'^2 < \infty$. Assuming $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T$ are i.i.d. copies of $\boldsymbol{\epsilon}$m at the sample level we have $\tilde{\boldsymbol{\theta}}_i' = \boldsymbol{\beta}_0 + \mathbf{B}_m^T\tilde{\boldsymbol{\theta}}_i + \boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$. The regression function of ν' on ν is defined as the global Fréchet regression function of ν' on the Euclidean predictor parameter $\boldsymbol{\gamma} = \boldsymbol{\beta}_0 + \mathbf{B}_m^T\tilde{\boldsymbol{\theta}}$, given by

$$\nu'_{\oplus}(\boldsymbol{\gamma}_0) = \underset{w \in \Omega}{\arg\min}\ E(\tilde{s}(\boldsymbol{\gamma}, \boldsymbol{\gamma}_0)d^2(\nu'_{\tilde{\boldsymbol{\theta}}'}, w)), \tag{7}$$

where $\boldsymbol{\gamma}_0 \in \mathbb{R}^m$ and the global weights are $\tilde{s}(\boldsymbol{\gamma}, \boldsymbol{\gamma}_0) = 1 + (\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{\gamma}})^T {\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}}^{-1}(\boldsymbol{\gamma}_0 - \boldsymbol{\mu}_{\boldsymbol{\gamma}})$ with $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = E(\boldsymbol{\gamma})$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \mathrm{Var}(\boldsymbol{\gamma})$ is assumed to be positive definite. This allows to incorporate random errors in the parametrization of the response distribution ν' if desired, while capturing a denoised response distribution; see Example 3 below.

An intermediate target assuming knowledge of the $\tilde{\boldsymbol{\theta}}_i$ and both $\boldsymbol{\beta}_0$ and $\mathbf{B}_m$ is then

$$\hat{\nu}'_{\oplus}(\boldsymbol{\gamma}_0) = \underset{w \in \Omega}{\arg\min}\ n^{-1}\sum_{i=1}^n s_{in}(\boldsymbol{\gamma}_0)d^2(\nu_i', w), \tag{8}$$

with empirical weights $s_{in}(\boldsymbol{\gamma}_0) = 1 + (\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}})^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{\gamma}_0 - \bar{\boldsymbol{\gamma}})$, where $\bar{\boldsymbol{\gamma}} = n^{-1}\sum_{i=1}^n \boldsymbol{\gamma}_i$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}} = n^{-1}\sum_{i=1}^n (\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}})(\boldsymbol{\gamma}_i - \bar{\boldsymbol{\gamma}})^T$.

However, the intermediate target in (8) relies on unknown quantities. To resolve this, we utilize as predictors the fitted values of the multivariate multiple linear regression performed on the MDS components between the response and predictor distributions. Let $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_i'$ be the MDS component from the distributions η_i and η_i', respectively, where $i = 1, \ldots, n$.

Denoting the response matrix $\mathbf{Y}_n \in \mathbb{R}^{n\times m}$ whose ith row is given by ${\boldsymbol{\eta}_i'}^T$, $i = 1, \ldots, n$, the fitted values are then given by the rows of

$$\hat{\mathbf{Y}}_n = \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{Y}_n.$$

For the empirical estimate of (8), we utilize as predictors the rows of $\hat{\mathbf{Y}}_n$. Denoting by $\mathbf{Z}_i = \hat{\mathbf{Y}}_n^T\mathbf{e}_i$, $i = 1, \ldots, n$, this leads to

$$\tilde{\nu}'_{\oplus}(\tilde{\boldsymbol{\gamma}}_0) = \underset{w\in\Omega}{\arg\min}\ n^{-1}\sum_{i=1}^{n} \hat{s}_{in}(\tilde{\boldsymbol{\gamma}}_0)d^2(\nu'_i, w), \tag{9}$$

where $\tilde{\boldsymbol{\gamma}}_0 \in \mathbb{R}^m$, $\hat{s}_{in}(\tilde{\boldsymbol{\gamma}}_0) = 1+(\mathbf{Z}_i-\bar{\mathbf{Z}})^T\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}}^{-1}(\tilde{\boldsymbol{\gamma}}_0-\bar{\mathbf{Z}})$, and $\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}} = n^{-1}\sum_{i=1}^n(\mathbf{Z}_i-\bar{\mathbf{Z}})(\mathbf{Z}_i-\bar{\mathbf{Z}})^T$. The next result shows that the global Fréchet regression function in (8) can be consistently recovered.

Theorem 4 *Suppose that (A1), (A1'), (A2), (A3), and the regularity conditions* $(U0)-(U2)$ *in Petersen and Müller (2019) hold. If* $\mathbf{B}_m$ *is invertible, then*

$$\max_{i=1,\ldots,n} d_{\mathcal{W}_2}(\tilde{\nu}'_{\oplus}(\mathbf{Z}_i), \nu'_{\oplus}(\boldsymbol{\gamma}_i)) = O_p\left(n^{-1/(2(\alpha'-1))}\right),$$

for any $\alpha' > \alpha$ *with* α *as in (U2).*

Similarly as in Example 1 in Petersen et al. (2021) and observing the relation $s(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}_0) = s(\boldsymbol{\beta}_0+\mathbf{B}_m^T\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}_0+\mathbf{B}_m^T\tilde{\boldsymbol{\theta}}_0)$ for any point $\tilde{\boldsymbol{\theta}}_0$ in the space $\tilde{\Theta} = \{\mathbf{A}\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta\}$ with $\mathbf{A}$ as in (A1), the following example shows that in this regression framework we can denoise in certain situations the response distribution ν' by employing the predictor distribution ν.

Example 3 *Suppose* $m = 2$, $\Theta \subset \mathbb{R}^2$, *and let* $\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12} \in \mathbb{R}$ *be constants such that* $\beta_{02}+\beta_{12}\vartheta_2 > 0$ *and* $\beta_{02}+\beta_{12}\vartheta_2+\epsilon_2 > 0$ *almost surely for all* $\tilde{\boldsymbol{\theta}} = (\vartheta_1, \vartheta_2)^T$ *and* $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)^T$. *Let* $\tilde{\boldsymbol{\theta}}' = (\beta_{01}+\beta_{11}\vartheta_1+\epsilon_1, \beta_{02}+\beta_{12}\vartheta_2+\epsilon_2)^T$. *Consider the location-scale model with noise at both location and scale levels,*

$$f_{\nu'|\boldsymbol{\theta}'}(\cdot) = \frac{1}{\beta_{02}+\beta_{12}\vartheta_2+\epsilon_2} f_0\left(\frac{\cdot-(\beta_{01}+\beta_{11}\vartheta_1+\epsilon_1)}{\beta_{02}+\beta_{12}\vartheta_2+\epsilon_2}\right),$$

where $f_0 > 0$ *is a density function over* $\mathbb{R}$ *such that* $\int_{\mathbb{R}} sf_0(s)ds = 0$ *and* $\int_{\mathbb{R}} s^2 f_0(s)ds = 1$, *and suppose that* $Cov(\vartheta_1, \vartheta_2) = 0$. *For simplicity consider* $\Omega = \mathcal{W}_2$ *and let* $\boldsymbol{\gamma}_0 = \boldsymbol{\beta}_0+\mathbf{B}_m^T\tilde{\boldsymbol{\theta}}_0$, *where* $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02})^T$, $\tilde{\boldsymbol{\theta}}_0 = (\vartheta_{01}, \vartheta_{02})^T$, *and* $\mathbf{B}_m$ *contains among its columns the vectors* $(\beta_{11}, 0)^T$ *and* $(0, \beta_{12})^T$. *Then the density* $f_{\nu'_{\oplus}(\boldsymbol{\gamma}_0)}$ *of* $\nu'_{\oplus}(\boldsymbol{\gamma}_0)$ *in* (7) *satisfies*

$$f_{\nu'_{\oplus}(\boldsymbol{\gamma}_0)}(\cdot) = \frac{1}{\beta_{02}+\beta_{12}\vartheta_{02}} f_0\left(\frac{\cdot-(\beta_{01}+\beta_{11}\vartheta_{01})}{\beta_{02}+\beta_{12}\vartheta_{02}}\right).$$

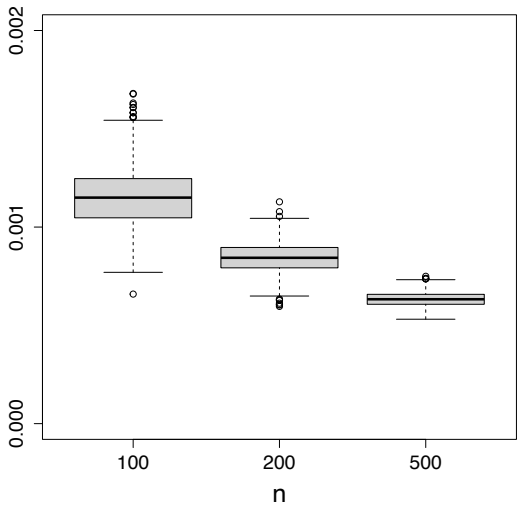

Figure 1: Box plots for the distributions of the average deviations $n^{-1}\sum_{i=1}^{n} d^2_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i)$ for a simulation setting with compactly supported truncated Gaussian distributions on $[0,1]$ obtained from 1000 simulations for increasing sample sizes, where distributions are estimated from $N(n) = 2n$ sample data generated by the respective distribution.

6. Numerical Experiments

We first consider compactly supported densities f_i corresponding to truncated Gaussian distributions on $[0,1]$ with mean parameter $\mu_i = \exp(1+U_i)$ and $\sigma_i = \exp(1+U_i')/2$, where the U_i and U_i' are independent and i.i.d. uniform variates on $[0,1]$. Here we suppose the f_i are unobserved and instead we have available a sample $Y_{i1}, \ldots, Y_{in_i} \overset{iid}{\sim} f_i$ which is generated conditional on (U_i, U_i'), and $n_i \geq N(n) = 2n$. Figure 1 displays boxplots for 1000 simulations and increasing sample sizes n for for the error metric $n^{-1}\sum_{i=1}^{n} d^2_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i)$. It quantifies the recovery of the oracle global Fréchet regression function as if the (μ_i, σ_i) were observed by using the perturbed MDS components as predictors. Clearly, the perturbed global Fréchet regression function is seen to converge to the oracle counterpart that in turns converges to its population level counterpart.

Further simulations show the performance of the methods in recovering the underlying global Fréchet regression function as well as capturing the underlying geometry of the data generation mechanism. For simplicity, we consider the densities f_i to be fully observed in what follows. Since the trailing eigenvalues λ_l, $l \geq m+1$, are numerically not identically zero, instead of employing the estimate $\hat{m}$ as in Section 3, we choose $\hat{m}$ so that the fraction of variance explained defined as $\mathrm{FVE}(\hat{m}) = \sum_{i=1}^{\hat{m}} \lambda_j / \sum_{i=1}^{n} \lambda_j$ first upcrosses 0.999.

First consider a Gaussian location-scale family $\nu|(\mu,\sigma) \sim N(\mu,\sigma)$, where (μ,σ) are random and generated as follows: $\mu = \mu(U) = \exp(1+U)$ with $U \sim \mathrm{Unif}(0,1)$ and $\sigma = \sigma(U) = \mu/2$ depends directly on U, so that this family is parametrized by a single variate U. We generate $n = 500$ random variables $U_i \overset{iid}{\sim} U$, and conditional on U_i we obtain the Gaussian density $f_i \sim N(\mu(U_i), \sigma^2(U_i))$. Therefore, there is only one source of

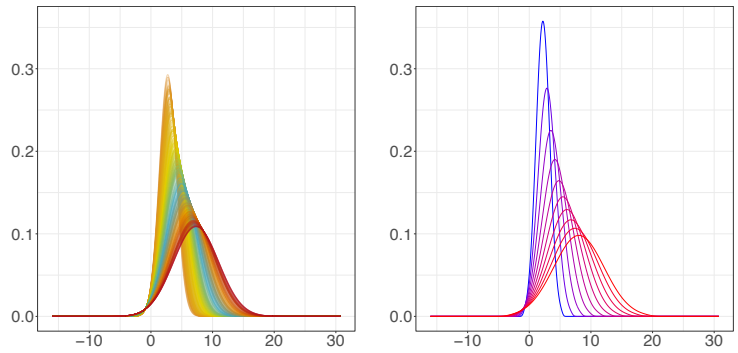

Figure 2: Visualizing Gaussian distributions $N(\mu, \sigma^2)$ in a simulation setting, where $\mu = \exp(1 + U)$ with $U \sim \text{Unif}(0, 1)$ and standard deviation $\sigma = \mu/2$ so that there is a single source of variation. The left panel displays individual densities f_i and the right panel the density functions estimated with global Fréchet regression for a dense grid of MDS components over the observed range.

randomness for the observed distributional data f_i, $i = 1, \ldots, n$. Figure 2 shows the results for this setting, where we employ the frechet R package (Chen et al. 2020) to obtain density estimates $\tilde{f}_{\oplus}(\cdot, \boldsymbol{\eta})$ (4). The left panel displays the observed densities f_i colored according to the values U_i while the right panel contains the estimated global Fréchet regression function (Petersen and Müller 2019) over a dense grid of observed $\boldsymbol{\eta}_i$ values. Clearly, the method is able to keep track of the single intrinsic parameter that originally parametrizes the class of observed densities. It is also seen that the underlying densities f_i can indeed be recovered via global Fréchet regression and specifically that the single underlying parameter is related to a mean-shift with increasing standard deviation for the Gaussian densities. We emphasize that the proposed approach is capable of recovering both shape and parametrization of this distribution family.

We next consider the situation when both the mean μ and standard deviation σ depend on independent uniform random variates U and U' as follows. We generate $\mu_i = \exp(1+U_i)$ and $\sigma_i = \exp(1 + U_i')/2$, $i = 1, \ldots, n$, where U_i and U_i' are i.i.d. copies of U and U', respectively. This data generating mechanism is similar as before but has a secondary source of variation and therefore two underlying parameters. The right panel of Figure 3 shows the individual densities f_i, where it is difficult to disentangle the effect of each underlying parameter. The left panel shows the global Fréchet regression density function when increasing the first MDS component over its observed range while keeping the second one at its mean level, which is identically zero, and vice versa for the right panel. Clearly, the global Fréchet regression density function successfully recovers the mean-shift and vertical variation that is present in the data generating mechanism.

Turning to another scenario, we consider a family $\nu|(\alpha, \beta) \sim \Gamma(\alpha, \beta)$ of Gamma distributions with parameters (α, β), where $\alpha \sim \text{Unif}(2.5, 4)$, $\beta \sim \text{Unif}(1.5, 4)$, with α and β independent. Thus the distributional data has two sources of variation. Figure 4 shows the results, where global Fréchet regression is seen to keep track of the two underlying parameters, where the first one is related to a mean-shift to the right with an overall flattening of

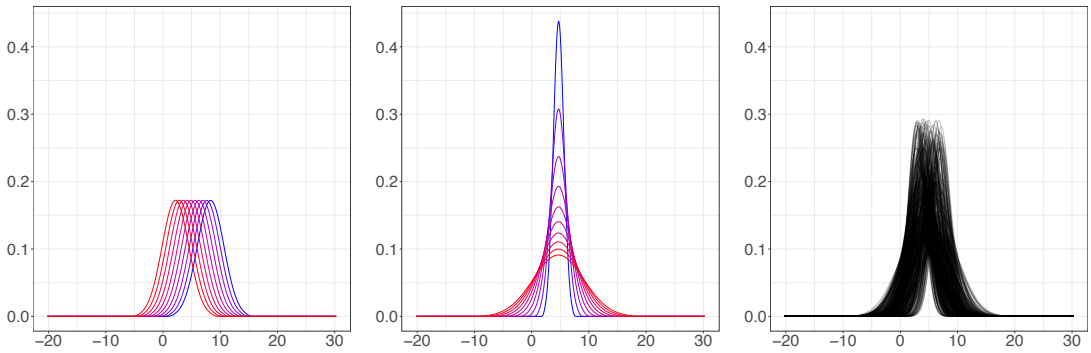

Figure 3: Visualizing Gaussian distributions $N(\mu, \sigma^2)$ in a simulation setting, where $\mu = \exp(1 + U_1)$ and $\sigma = \exp(1 + U_2)/2$ with $U_j \overset{iid}{\sim} \text{Unif}(0, 1)$, $j = 1, 2$, so that in contrast to Figure 2 there are two sources of variation. The left panel shows the density functions obtained from global Fréchet regression when increasing the first MDS component over its observed range while keeping the second MDS component at its mean level at zero. The middle panel shows the density functions obtained analogously when increasing the second MDS component. The generated sample of densities f_i for which the representations in the left and middle panels are obtained is displayed in the right panel.

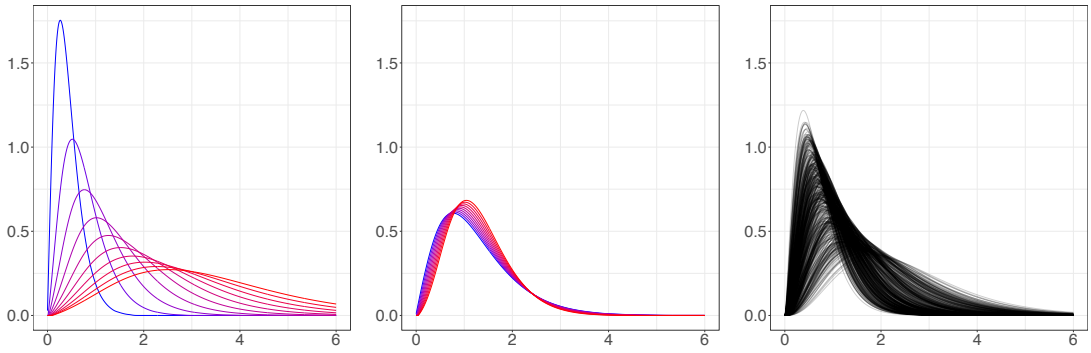

Figure 4: Visualizing Gamma distributions in a simulation setting $\nu|(\alpha, \beta) \sim \Gamma(\alpha, \beta)$, where $\alpha \sim \text{Unif}(2.5, 4)$, $\beta \sim \text{Unif}(1.5, 4)$ and α and β are independent, so that there are two sources of variation. The left panel displays the density functions obtained with global Fréchet regression when increasing the first MDS component over its observed range while keeping the second one at its mean level zero, and vice versa for the middle panel. The right panel displays the individual densities f_i.

the density curve, while the second component is also related to a mean-shift, accompanied by a sharpening of the curve which is related to decreased variance.

We next consider the case when $\nu|(\alpha, \beta) \sim \mathcal{B}(\alpha, \beta)$ has a beta distribution with parameters (α, β), where $\alpha = 2 + \exp(4\sin(U))$, $\beta = 2 + \exp(4\sin(U'))$, and $U, U' \overset{iid}{\sim} \text{Unif}(0, 1)$. Thus, the distributional data has two sources of variation. The results are in Figure 5. The right panel shows the individual densities where it is seen that there is a mean-shift

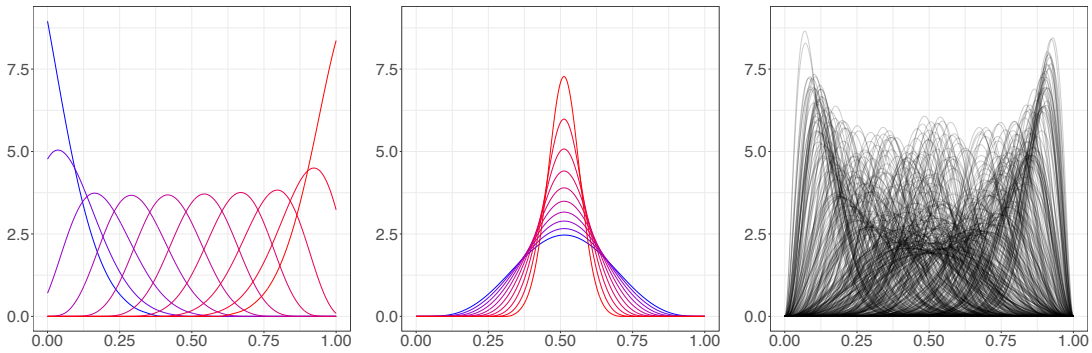

Figure 5: Visualizing Beta distributions in a simulation setting $\nu|(\alpha,\beta) \sim \mathcal{B}(\alpha,\beta)$, where $\alpha = 2+\exp(4\sin(U_1))$ and $\beta = 2+\exp(4\sin(U_2))$ with $U_j \overset{iid}{\sim} \text{Unif}(0,1)$, $j=1,2$, so that there are two underlying parameters. The left panel shows the density functions obtained with global Fréchet regression when increasing the first MDS component over its observed range while keeping the second one at its mean level zero, and vice versa for the middle panel. The right panel shows the individual densities f_i.

as well as vertical variation. Indeed, the density functions recovered from global Fréchet regression keep track of the horizontal variation with the first MDS component (left panel) while the vertical variation is embedded in the second component (middle panel), so that it successfully disentangles the separate sources of variation present in these distributional data.

Finally, we consider the same beta simulation setting as before except that now $\beta = \alpha$ so that there is only a single source of variation in the data. The left panel of Figure 6 shows the individual densities and the right panel the recovered densities from Fréchet regression for a dense grid over the observed MDS components. Clearly, the proposed method recovers the parametrization inherent in this sample of densities.

7. Data Applications

We first illustrate the proposed methodology with the distribution of baby names distribution over the time window spanning from 1980 to 2020 in the United States. These data are available from `https://catalog.data.gov/dataset` and consist of histograms reflecting the popularity of bay names from 1880 to 2020 at a yearly granularity. These histogram data are converted to distributional data in a pre-smoothing step (Petersen and Müller 2016; Cazelles et al. 2018; Han et al. 2020).

We consider the 500 most popular male baby names during 1980 and obtain each corresponding density function from the associated histogram by employing the frechet R package (Chen et al. 2020). The results are shown in Figure 7, where we select two components in the MDS step. The left panel shows that the first component captures baby names that were popular early on and then had a consistent decrease in popularity as well as baby names that were not very popular early on but gained increased popularity consistently

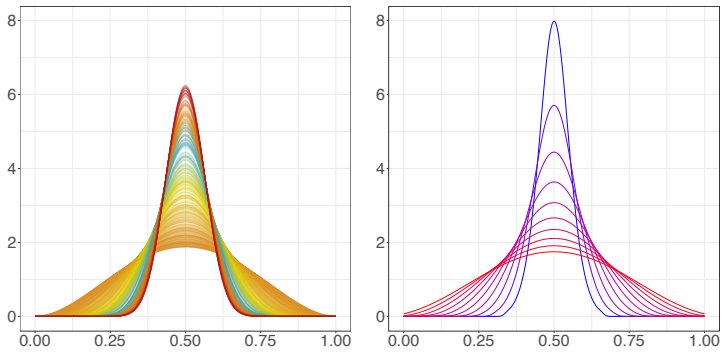

Figure 6: Visualizing Beta distributions in a simulation setting $\nu|(\alpha,\beta) \sim \mathcal{B}(\alpha,\beta)$, where $\alpha = 2 + \exp(4\sin(U))$ and $\beta = \alpha$ with $U \sim \text{Unif}(0,1)$ so that there is only a single source of variation in the data. The left panel displays the individual densities f_i and the right panel the densities recovered by global Fréchet regression when increasing the first MDS component over its observed range.

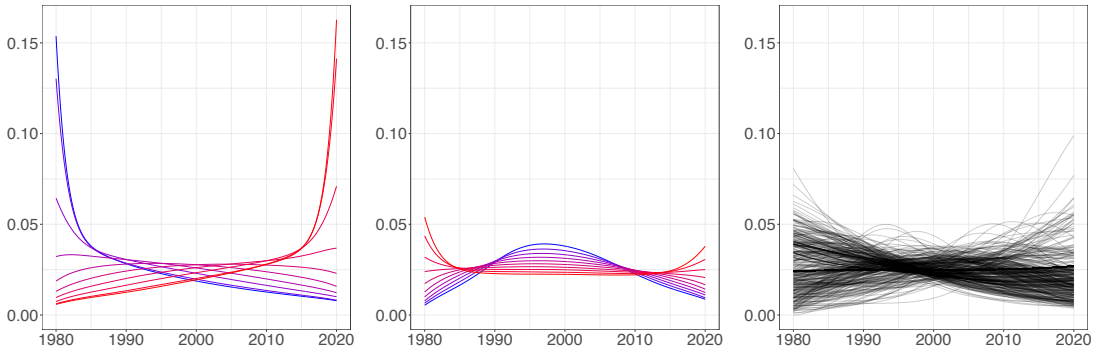

Figure 7: Baby names popularity distribution in the United States from 1980 to 2020 for the 500 most popular names at the beginning of the time window. The left panel displays the distributions recovered with global Fréchet regression function when increasing the first MDS component from its 1% quantile to its 99% quantile (blue to red) while keeping the second MDS component fixed at its mean level 0, and vice versa for the middle panel. The right panel shows the observed smoothed density functions.

across time. The second component reflects baby names that became more popular halfway through the time window followed by a decline in popularity.

As a second illustration, we consider the daily rental bike pickups distribution during weekdays in 2019 at a popular station in the divvy bike system in Chicago. The data is publicly available at `https://www.divvybikes.com/system-data` and contains information such as the bike pickup times for specific bike stations. These data recently gained increased attention and they have been analyzed from a point process perspective (Gervini and Khanal 2019; Gervini 2022; Gajardo and Müller 2022). Each weekday provides a sample of random times coming from an underlying density function which we estimate by first constructing the empirical quantile function and then employing the frechet R package

to map back to density space via kernel smoothing methods, where the smoothing bandwidth is chosen by cross-validation with a minimum bandwidth of 10% of the 24 hours time window to prevent undersmoothing.

Figure 8 shows the results when employing $m = 2$ components. The first MDS component captures the two modes of the demand around 8am and 5pm, clearly due to bike use by commuters, while the second component is seen to capture a boost in the demand around 2pm which may be related to midday commuting or lunch breaks. These findings are similar to those in Gajardo and Müller (2022), where this dataset was analyzed from a temporal Cox point process regression perspective.

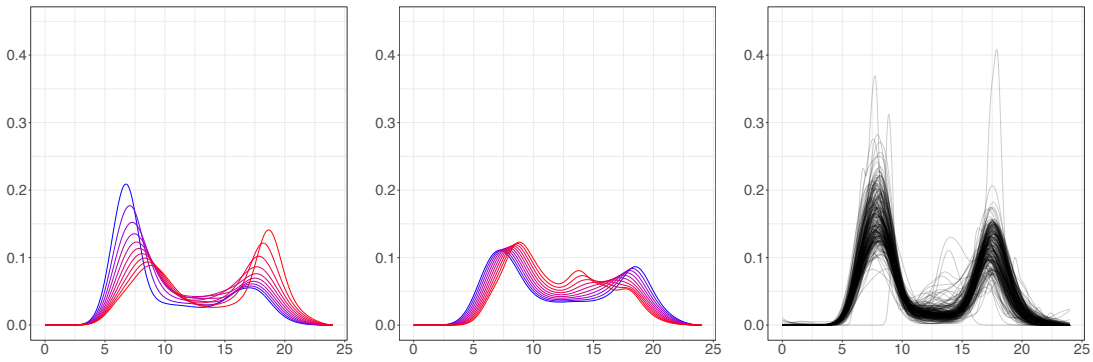

Figure 8: Distribution of daily bike pickups at a station in the Divvy bike system in Chicago during weekdays of 2019, using two MDS components. The left panel displays the recovery obtained from global Fréchet regression function when increasing the first MDS component from its 1% quantile to its 99% quantile (blue to red), while keeping the second MDS component fixed at its mean level 0, and vice versa for the middle panel. The right panel shows the observed smoothed density functions. Here the units on the x-axis are hours of day time 0-24h.

As a third data illustration, we analyze age pyramids across countries in the year 2015. These data consist of histograms corresponding to the population age pyramids for several countries containing the size of the population for age groups defined by age in years and are publicly available at

`https://www.census.gov/programs-surveys/international-programs/data.html`.

They have been previously analyzed from different perspectives such as Functional Data Analysis and Wasserstein geodesic principal component analysis (Delicado 2011; Bigot et al. 2017; Cazelles et al. 2018). The data are converted to smooth age distributions for each country and then the proposed Fréchet-Wasserstein manifold learning approach is applied.

The results are in Figure 9. Here the first MDS component is seen to move distributions with a high fraction of older people towards those with a larger fraction of young people. The second component paints a more nuanced picture, differentiating distributions that have densities with a more pronounced peak at middle age around age 38 from those that don't have such a peak.

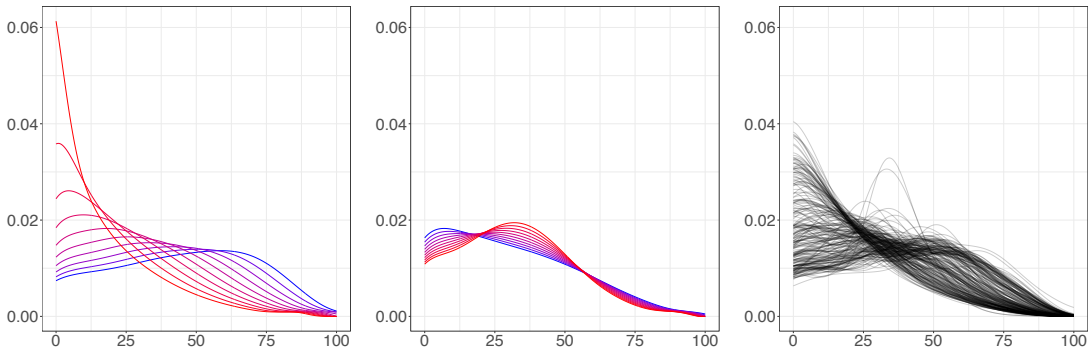

Figure 9: Age distributions corresponding to population age pyramids across a sample of countries in 2015, using two MDS components. The left panel displays the densities obtained from global Fréchet regression when increasing the first MDS component from its 1% quantile to its 99% quantile (blue to red) while keeping the second MDS component fixed at its mean level 0, and vice versa for the middle panel. The right panel shows the smoothed observed density functions of the age distributions. The x-axis represents age in years.

8. Discussion

We develop a Fréchet manifold learning approach for general random objects situated in a metric space by adopting a version of ISOMAP to obtain low-dimensional Euclidean representations, which mimics a parametric representation for a family of random objects. These representations are fully learned from the data without formulating a parametric model or specifying other characteristics such as the shape of a density. Key for the success of this Fréchet manifold learning approach is the utilization of Fréchet regression to obtain a suitable map back from the low-dimensional representation space to the object space.

It is important to note that this map from the low-dimensional representation space to the object space is also available at unobserved Euclidean representation points and allows to study the effect that each component has on the objects in the metric space, in addition to interpolation or extrapolation, and more generally delineating the set of objects of interest. Specifically for the case of families of distributions as random objects, the proposed method will find an empirical parametrization that can be used to determine an underlying parametric family for the observed densities and also as a heuristic diagnostic for a given a priori parametrization.

The proposed method also has a number of limitations. For example, the global Fréchet regression output does not necessarily lie in the manifold $\mathcal{M}$ as the optimization problem that defines the conditional barycenter is over Ω. In such cases the resulting regression is not necessarily an appropriate notion of center conditional on the predictors. The exclusion of such occurrences can be viewed as a modeling assumption regarding the dependency of the random objects lying in $\mathcal{M} = \psi(\Theta) \subset \Omega$ on the parameter space Θ that also depends on the geometry induced by the underlying metric d. Indeed, since global Fréchet regression is a generalization of classical multiple linear regression for responses that are objects lying in a metric space (Petersen and Müller 2019), some generalized version of linearity of the relation

between the random objects and the low-dimensional parameters is implicitly assumed; see, e.g., the location-scale family of probability distributions in Example 1 for an illustration. Additionally, the global isometry condition (in the ambient space metric) (A1) between the manifold and a subset of Euclidean space does not cover general curved manifolds in object space , as the ISOMAP method of constructing geodesics with a neighborhood graph is not part of our approach.

For general metric spaces (Ω, d), the optimization problem involved in global Fréchet regression and the estimation of the regression function may be challenging as it depends on the specifics of the space Ω and the metric d. We remark that for several important statistical object spaces of interest such as probability distributions in 2-Wasserstein space, probability distributions with the Fisher-Rao metric, covariance matrices with the Frobenius and other metrics, Cox point processes and other random objects situated in common metric spaces there are available approaches to address the respective optimization problems for global Fréchet regression (Petersen and Müller 2019; Gajardo and Müller 2022).

Appendix A. Additional Assumptions and Rates of Convergence

For completeness, we list here Assumptions (U0)-(U2) that are part of Petersen and Müller (2019). Let $\|\cdot\|_E$ be the Euclidean norm on $\mathbb{R}^m$ and $B > 0$. With global weights as above

$$s(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = 1 + (\boldsymbol{\theta} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}).$$

and their empirical counterparts

$$s_{in}(\boldsymbol{\theta}_0) = 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}),$$

we define the weighted Fréchet functions

$$M(w, \boldsymbol{\theta}_0) = E(s(\boldsymbol{\theta}, \boldsymbol{\theta}_0) d^2(\nu_{\boldsymbol{\theta}}, w)), \quad M_n(w, \boldsymbol{\theta}_0) = n^{-1} \sum_{i=1}^{n} s_{in}(\boldsymbol{\theta}_0) d^2(\nu_i, w).$$

Then for $x \in \mathbb{R}^m$, $w \in \Omega$, the additional assumptions are as follows.

(U0) Almost surely, for all $\|x\|_E \leq B$, the objects $\nu_\oplus(x)$ and $\hat{\nu}_\oplus(x)$ exist and are unique. Additionally, for any $\varepsilon > 0$,

$$\inf_{\|x\|_E \leq B} \inf_{d(\omega, \nu_\oplus(x)) > \varepsilon} M(\omega, x) - M(\nu_\oplus(x), x) > 0$$

and there exists $\zeta = \zeta(\varepsilon) > 0$ such that

$$P\left(\inf_{\|x\|_E \leq B} \inf_{d(\omega, \hat{\nu}_\oplus(x)) > \varepsilon} M_n(\omega, x) - M_n(\hat{\nu}_\oplus(x), x) \geq \zeta \right) \to 1.$$

(U1) The entropy integral for the space Ω is finite, i.e.

$$\int_0^1 \sqrt{1 + \log N(\epsilon, \Omega, d)}\, d\epsilon < \infty.$$

(U2) There exist $\tau > 0$, $D > 0$, and $\alpha > 1$, possibly depending on B, such that

$$\inf_{\|x\|_E \le B} \inf_{d(\omega, \nu_\oplus(x)) < \tau} \{M(w,x) - M(\nu_\oplus(x), x) - Dd(w, \nu_\oplus(x))^\alpha\} \ge 0.$$

Appendix B. Auxiliary Results and Proofs

Here we provide detailed proofs. For a matrix $\mathbf{R} \in \mathbb{R}^{p\times q}$ with $p, q > 0$ integers, denote by $\|\mathbf{R}\|_{\text{op},2}$ its operator norm. Recall that $\tilde{\boldsymbol{\theta}} = \mathbf{A}\boldsymbol{\theta}$, $\boldsymbol{\Sigma}_0 = \text{Cov}(\tilde{\boldsymbol{\theta}}) \in \mathbb{R}^{m\times m}$, $\lambda_1 \ge \cdots \ge \lambda_n$ are the ordered eigenvalues of $(-1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n$, and $\lambda_{01} \ge \cdots \ge \lambda_{0m}$ are those of $\boldsymbol{\Sigma}_0$.

Lemma 1 *Suppose that (A1) and (A2) hold. Let* $\delta = \min_{j=1,\ldots,m-1}(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$ *whenever* $m > 1$ *and* $\delta = \infty$ *when* $m = 1$*, with* $\lambda_0 = \infty$ *and* δ_0 *as in (A2). Then*

$$|n^{-1}\lambda_l - \lambda_{0l}| = O_p(n^{-1/2}),$$

where $l = 1, \ldots, m$ *and the bound is uniform in* l*. Moreover, for dimensions* $m > 1$ *and for any* $\epsilon \in (0, \delta_0)$*, the event where*

$$n^{-1}\delta \ge \delta_0 - \epsilon,$$

holds with probability tending to 1*.*

Proof [Proof of Lemma 1] Since $\tilde{\boldsymbol{\theta}} = \mathbf{A}\boldsymbol{\theta}$, where $\mathbf{A}$ is defined as in (A1), we have $\tilde{\boldsymbol{\theta}}_i = \mathbf{A}\boldsymbol{\theta}_i$, $i = 1, \ldots, n$, and $\tilde{\boldsymbol{\mu}}_n = n^{-1}\sum_{i=1}^n \tilde{\boldsymbol{\theta}}_i$. Define $\tilde{\mathbf{B}}_n = (\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\mu}}_n \cdots \tilde{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\mu}}_n) \in \mathbb{R}^{m\times n}$, the matrix containing $\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\mu}}_n$ on its ith column. Note the relation

$$-\frac{1}{2}\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n = \tilde{\mathbf{B}}_n^T\tilde{\mathbf{B}}_n,$$

which shows that the rank of $-(1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n$ is at most m, with eigen-equations

$$\tilde{\mathbf{B}}_n^T\tilde{\mathbf{B}}_n\boldsymbol{\nu} = \lambda\boldsymbol{\nu},$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^T$, which imply for all $k = 1, \ldots, n$,

$$(\tilde{\boldsymbol{\theta}}_k - \tilde{\boldsymbol{\mu}}_n)^T \sum_{j=1}^n (\tilde{\boldsymbol{\theta}}_j - \tilde{\boldsymbol{\mu}}_n)\nu_j = \lambda\nu_k.$$

Writing $\tilde{\boldsymbol{\theta}}_k = (\tilde{\theta}_{k1}, \ldots, \tilde{\theta}_{km})^T$, $\tilde{\boldsymbol{\mu}}_n = (\tilde{\mu}_{n1}, \ldots, \tilde{\mu}_{nm})^T$, observe that for any $q = 1, \ldots, m$,

$$\sum_{l=1}^m \sum_{j=1}^n (\tilde{\theta}_{kq} - \tilde{\mu}_{nq})(\tilde{\theta}_{kl} - \tilde{\mu}_{nl})(\tilde{\theta}_{jl} - \tilde{\mu}_{nl})\nu_j = \lambda\nu_k(\tilde{\theta}_{kq} - \tilde{\mu}_{nq}),$$

so that

$$\sum_{l=1}^m \sum_{k=1}^n (\tilde{\theta}_{kq} - \tilde{\mu}_{nq})(\tilde{\theta}_{kl} - \tilde{\mu}_{nl}) \sum_{j=1}^n (\tilde{\theta}_{jl} - \tilde{\mu}_{nl})\nu_j = \lambda \sum_{k=1}^n \nu_k(\tilde{\theta}_{kq} - \tilde{\mu}_{nq}).$$

Setting $y_q = \sum_{k=1}^{n} \nu_k(\tilde{\theta}_{kq} - \tilde{\mu}_{nq})$, $q = 1, \ldots, m$, it follows that

$$\sum_{l=1}^{m} n^{-1} \sum_{k=1}^{n} (\tilde{\theta}_{kq} - \tilde{\mu}_{nq})(\tilde{\theta}_{kl} - \tilde{\mu}_{nl}) y_l = n^{-1}\lambda y_q,$$

which reveals that $n^{-1}\lambda$ is an eigenvalue of the matrix $\hat{\mathbf{\Sigma}}_0 \in \mathbb{R}^{m\times m}$ defined element-wise by

$$[\hat{\mathbf{\Sigma}}_0]_{ql} = n^{-1} \sum_{k=1}^{n} (\tilde{\theta}_{kq} - \tilde{\mu}_{nq})(\tilde{\theta}_{kl} - \tilde{\mu}_{nl}),$$

where $q, l = 1, \ldots, m$. That $n^{-1}\lambda$ is a valid eigenvalue of $\hat{\mathbf{\Sigma}}_0$ requires that $\mathbf{y} = (y_1, \ldots, y_m)^T$ is non-zero or equivalently $\|\mathbf{y}\|_2 \neq 0$. This is easy to verify as otherwise $\|\boldsymbol{\nu}\|_2 = 0$ which contradicts the fact that $\boldsymbol{\nu}$ is a valid eigenvector of $\tilde{\mathbf{B}}_n^T \tilde{\mathbf{B}}_n$.

Observe that under (A1) and since $\mathbf{\Sigma}$ is assumed positive definite, so is $\mathbf{\Sigma}_0$. By the strong law of large numbers, $\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_F = o(1)$ almost surely so that by Weyl's inequality $\hat{\mathbf{\Sigma}}_0$ is positive definite with probability tending to 1. It then suffices to work conditional on this event in what follows. Observing $\hat{\mathbf{\Sigma}}_0 = n^{-1}\tilde{\mathbf{B}}_n \tilde{\mathbf{B}}_n^T$, for any eigenpair $(\lambda, \mathbf{z})$ of $\hat{\mathbf{\Sigma}}_0$ the condition $n^{-1}\tilde{\mathbf{B}}_n \tilde{\mathbf{B}}_n^T \mathbf{z} = \lambda \mathbf{z}$ leads to

$$\tilde{\mathbf{B}}_n^T \tilde{\mathbf{B}}_n \tilde{\mathbf{B}}_n^T \mathbf{z} = n\lambda \tilde{\mathbf{B}}_n^T \mathbf{z}.$$

This implies $n\lambda$ is an eigenvalue of $\tilde{\mathbf{B}}_n^T \tilde{\mathbf{B}}_n$ with corresponding eigenvector $\tilde{\mathbf{B}}_n^T \mathbf{z}$. The latter quantity is not identically zero and therefore is a valid eigenvector since $\|\tilde{\mathbf{B}}_n^T \mathbf{z}\|_2^2 = \mathbf{z}^T \tilde{\mathbf{B}}_n \tilde{\mathbf{B}}_n^T \mathbf{z}$, $\tilde{\mathbf{B}}_n \tilde{\mathbf{B}}_n^T$ is positive definite and $\|\mathbf{z}\|_2 \neq 0$ by construction. Therefore, using that the rank of $\tilde{\mathbf{B}}_n^T \tilde{\mathbf{B}}_n$ is at most m reveals that the positive eigenvalues of $\tilde{\mathbf{B}}_n^T \tilde{\mathbf{B}}_n$ are exactly those of $n\hat{\mathbf{\Sigma}}_0$.

Note that under (A1) and (A2), the ordered eigenvalues of $\tilde{\mathbf{B}}_n^T \tilde{\mathbf{B}}_n$ are $\lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_n = 0$ and the eigenvalues of $\mathbf{\Sigma}_0$ satisfy $\lambda_{01} > \cdots > \lambda_{0m} > 0$. By Weyl's inequality, for any $l = 1, \ldots, m$,

$$|n^{-1}\lambda_l - \lambda_{0l}| \leq \|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\mathrm{op},2} = O_p(n^{-1/2}), \tag{10}$$

which shows the first result. Next, note that

$$\begin{aligned} n^{-1}\delta &\geq \min_{j=1,\ldots,m-1} (\lambda_{0j} - \lambda_{0(j+1)}) - 2\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\mathrm{op},2} \\ &= \delta_0 - 2\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\mathrm{op},2}, \end{aligned} \tag{11}$$

where δ_0 as defined in (A2) is the eigenspacing of $\mathbf{\Sigma}_0$. In view of (10), we have that for any $\epsilon \in (0, \delta_0)$ the event where $\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\mathrm{op},2} \leq \epsilon/2$ holds with probability tending to 1. From (11) we obtain

$$P\left(\|\hat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}_0\|_{\mathrm{op},2} \leq \epsilon/2\right) \leq P\left(n^{-1}\delta \geq \delta_0 - \epsilon\right).$$

The second result follows. ■

Recall the spectral decomposition $(-1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, where $\mathbf{Q} = (\boldsymbol{\nu}_1 \cdots \boldsymbol{\nu}_n) \in \mathbb{R}^{n\times n}$ contains the eigenvectors $\boldsymbol{\nu}_i \in \mathbb{R}^n$ with corresponding ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$, $\mathbf{Q}_m = (\boldsymbol{\nu}_1 \cdots \boldsymbol{\nu}_m) \in \mathbb{R}^{n\times m}$, and $\mathbf{\Lambda}_m = \mathrm{diag}(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^{m\times m}$. Also recall the estimated counterparts $(-1/2)\mathbf{J}_n\hat{\mathbf{D}}_n\mathbf{J}_n = \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^T$, where $\hat{\mathbf{Q}} = (\hat{\boldsymbol{\nu}}_1 \cdots \hat{\boldsymbol{\nu}}_n) \in \mathbb{R}^{n\times n}$ contains the eigenvectors $\hat{\boldsymbol{\nu}}_i \in \mathbb{R}^n$ with corresponding ordered eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n \geq 0$, $\hat{\mathbf{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \in \mathbb{R}^{n\times n}$, $\hat{\mathbf{Q}}_m = (\hat{\boldsymbol{\nu}}_1 \cdots \hat{\boldsymbol{\nu}}_m) \in \mathbb{R}^{n\times m}$, and $\hat{\mathbf{\Lambda}}_m = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_m) \in \mathbb{R}^{m\times m}$. Here $\hat{\boldsymbol{\nu}}_l^T\boldsymbol{\nu}_l \geq 0$ are assumed aligned, $l = 1, \ldots, m$, which does not alter the estimates $\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j)$.

Recall the estimate $\tilde{m}$ of the dimension m defined by $\tilde{m} = \sup\{l = 1, \ldots, n\colon \lambda_l > 0\}$.

Lemma 2 *Suppose that (A1) and (A2) hold. Then, for any sequence $\alpha_n \to \infty$ as $n \to \infty$,*

$$\tilde{m} = m + O_p(\alpha_n^{-1}).$$

Proof [Proof of Lemma 2] Define $\mathbf{\Phi} = (\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_n) \in \mathbb{R}^{m\times n}$. The relation $(-1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n = \mathbf{J}_n\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{J}_n$ along with $\mathrm{rank}(\mathbf{\Phi}) \leq m$ implies $\mathrm{rank}((-1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n) \leq m$ and therefore $\lambda_j = 0$ for all $j = m+1, \ldots, n$. This implies $\tilde{m} \leq m$ and $\lambda_{\tilde{m}+1} = 0$. Let $M > 0$ and α_n a sequence such that $\alpha_n \to \infty$. Then

$$\begin{aligned} P\left(\alpha_n|\tilde{m} - m| > M\right) &\leq P\left(\tilde{m} > m + M/\alpha_n\right) + P\left(\tilde{m} < m - M/\alpha_n\right) \\ &\leq P\left(\tilde{m} \geq m + 1\right) + P\left(\tilde{m} < m\right) \\ &\leq P\left(n^{-1}\lambda_m = 0\right) \\ &\leq P\left(\lambda_{0m} \leq |n^{-1}\lambda_m - \lambda_{0m}|\right) \\ &= o(1), \end{aligned}$$

as $n \to \infty$, where the second inequality holds for large enough n, the third is due to $\tilde{m} \leq m$, and the last follows from Lemma 1 and $\lambda_{0m} > 0$. The result follows. ■

Proof [Proof of Theorem 1] From Lemma 2, the event where $\hat{m} = m$ holds with probability tending to 1 and therefore it suffices to work conditionally on this event. From Theorem 3.2 along with arguments in the proof of Corollary 3.3 in Hamm et al. (2023), and in view of (A1), there exists an invertible matrix $\mathbf{R}_n \in \mathbb{R}^{m\times m}$ and a vector $\mathbf{b}_n \in \mathbb{R}^m$ such that the MDS components satisfy $\boldsymbol{\eta}_i = \mathbf{R}_n\mathbf{A}\boldsymbol{\theta}_i + \mathbf{b}_n$. Let $\mathbf{A}_n = \mathbf{R}_n\mathbf{A} \in \mathbb{R}^{m\times m}$. Since $\mathbf{A}_n$ is invertible, we obtain

$$\tilde{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^n(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^T = n^{-1}\sum_{i=1}^n \mathbf{A}_n(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T\mathbf{A}_n^T = \mathbf{A}_n\hat{\mathbf{\Sigma}}\mathbf{A}_n^T,$$

and

$$\bar{\boldsymbol{\eta}} = n^{-1}\sum_{i=1}^n \boldsymbol{\eta}_i = \mathbf{A}_n\bar{\boldsymbol{\theta}} + b_n.$$

For any $j = 1, \dots, n$, it follows that

$$\begin{aligned}
\tilde{s}_{in}(\boldsymbol{\eta}_j) &= 1 + (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^T \tilde{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\eta}_j - \bar{\boldsymbol{\eta}}) \\
&= 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \mathbf{A}_n^T (\mathbf{A}_n^T)^{-1} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{A}_n^{-1} \mathbf{A}_n (\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}}) \\
&= 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}}) \\
&= s_{in}(\boldsymbol{\theta}_j),
\end{aligned}$$

which implies $\tilde{\nu}_{\oplus}(\boldsymbol{\eta}_j) = \hat{\nu}_{\oplus}(\boldsymbol{\theta}_j)$. In particular, this shows that (4) has a unique solution at points $\boldsymbol{\eta} = \boldsymbol{\eta}_j$, $j = 1, \dots, n$. Let $R > 0$ be such that $\mathrm{Diam}(\Theta) < R$. Then

$$\max_{i=1,\dots,n} d(\tilde{\nu}_{\oplus}(\boldsymbol{\eta}_i), \nu_{\oplus}(\boldsymbol{\theta}_i)) = \max_{i=1,\dots,n} d(\hat{\nu}_{\oplus}(\boldsymbol{\theta}_i), \nu_{\oplus}(\boldsymbol{\theta}_i)) \leq \sup_{\|\boldsymbol{\theta}_0\|_2 \leq R} d(\hat{\nu}_{\oplus}(\boldsymbol{\theta}_0), \nu_{\oplus}(\boldsymbol{\theta}_0)),$$

and the first result follows from Theorem 2 in Petersen and Müller (2019). Next,

$$\begin{aligned}
\tilde{s}_{in}(\boldsymbol{\eta}_{\mathbf{p}}) &= 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{A}_n^{-1} \left(\sum_{j=1}^n p_j \boldsymbol{\eta}_j - \bar{\boldsymbol{\eta}} \right) \\
&= 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{A}_n^{-1} \left(\sum_{j=1}^n (p_j - 1/n)(\mathbf{A}_n \boldsymbol{\theta}_j + \mathbf{b}_n) \right) \\
&= 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}^{-1} \left(\sum_{j=1}^n (p_j - 1/n) \boldsymbol{\theta}_j \right) \\
&= 1 + (\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\theta}_{\mathbf{p}} - \bar{\boldsymbol{\theta}}) \\
&= s_{in}(\boldsymbol{\theta}_{\mathbf{p}}).
\end{aligned}$$

Therefore $\tilde{\nu}_{\oplus}(\boldsymbol{\eta}_{\mathbf{p}}) = \hat{\nu}_{\oplus}(\boldsymbol{\theta}_{\mathbf{p}})$, which implies existence and uniqueness of (4) at any point belonging to the convex hull of $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n$. Since $\|\boldsymbol{\theta}_{\mathbf{p}}\|_2 \leq \sum_{j=1}^n p_j \|\boldsymbol{\theta}_j\|_2 \leq R$, we obtain

$$\begin{aligned}
\sup_{\mathbf{p}=(p_1,\dots,p_n)\colon p_j \geq 0, \|\mathbf{p}\|_1 = 1} d(\tilde{\nu}_{\oplus}(\boldsymbol{\eta}_{\mathbf{p}}), \nu_{\oplus}(\boldsymbol{\theta}_{\mathbf{p}})) &\leq \sup_{\mathbf{p}=(p_1,\dots,p_n)\colon p_j \geq 0, \|\mathbf{p}\|_1 = 1} d(\hat{\nu}_{\oplus}(\boldsymbol{\theta}_{\mathbf{p}}), \nu_{\oplus}(\boldsymbol{\theta}_{\mathbf{p}})) \\
&\leq \sup_{\|\boldsymbol{\theta}_0\|_2 \leq R} d(\hat{\nu}_{\oplus}(\boldsymbol{\theta}_0), \nu_{\oplus}(\boldsymbol{\theta}_0)).
\end{aligned}$$

The second result follows from Theorem 2 in Petersen and Müller (2019). ■

Proof [Proof of Corollary 2] The first result follows directly from Theorem 1. Recall that $\vartheta_{(1)} = \min_{i=1,\dots,n} \vartheta_i$, $\vartheta_{(n)} = \max_{i=1,\dots,n} \vartheta_i$, and F_ϑ is the CDF of ϑ. For any $\epsilon > 0$, observe

$$\Big\{ \vartheta_{(1)} \leq \epsilon \wedge \vartheta_{(n)} \geq 1 - \epsilon \Big\} = \Big\{ [\epsilon, 1 - \epsilon] \subseteq [\vartheta_{(1)}, \vartheta_{(n)}] \Big\}, \tag{12}$$

and

$$\begin{aligned}
&P\left(\vartheta_{(1)} \leq \epsilon \wedge \vartheta_{(n)} \geq 1-\epsilon\right) \\
&= 1 - P\left(\vartheta_{(1)} > \epsilon \vee \vartheta_{(n)} < 1-\epsilon\right) \\
&\geq 1 - P(\vartheta_{(1)} > \epsilon) - P(\vartheta_{(n)} < 1-\epsilon) + P(\vartheta_{(1)} > \epsilon \wedge \vartheta_{(n)} < 1-\epsilon) \\
&= 1 - (1-F_\vartheta(\epsilon))^n - [F_\vartheta(1-\epsilon)]^n + (F_\vartheta(1-\epsilon) - F_\vartheta(\epsilon))^n \\
&\geq 1 - (1-F_\vartheta(\epsilon))^n - [F_\vartheta(1-\epsilon)]^n, \qquad (13)
\end{aligned}$$

Since the density f_ϑ of ϑ is continuous and strictly positive over the compact set $\Theta = [0,1]$, there exist constants $0 < L < M < \infty$ such that $L \leq \inf_{s\in\Theta} f_\vartheta(s) \leq \sup_{s\in\Theta} f_\vartheta(s) \leq M$. Recall that $\epsilon = \epsilon_n = 1 - F_\vartheta^{-1}(1-n^{-\rho})$ with $\rho \in (0,1)$. By a Taylor expansion,

$$0 < 1 - F_\vartheta(\epsilon) \leq 1 - F_\vartheta(1 - F_\vartheta^{-1}(1-n^{-\rho})) \leq 1 - \frac{L}{M} n^{-\rho},$$

and $0 < F_\vartheta(1-\epsilon) \leq (1-n^{-\rho})$. Similarly, $0 < F_\vartheta(1-\epsilon) - F_\vartheta(\epsilon) \leq 1 - n^{-\rho}(1+L/M)$. From (13) we obtain

$$P\left(\vartheta_{(1)} \leq \epsilon \wedge \vartheta_{(n)} \geq 1-\epsilon\right) \geq 1 - (1 - \frac{L}{M} n^{-\rho})^n - (1-n^{-\rho})^n.$$

Elementary calculations show that for any two constants $\rho_1, c_0 > 0$ and since $\rho \in (0,1)$ we have $\lim_{n\to\infty}(1-c_0 n^{-\rho})^n n^{\rho_1} = o(1)$ as $n \to \infty$. Therefore, for large enough n,

$$P\left(\vartheta_{(1)} \leq \epsilon \wedge \vartheta_{(n)} \geq 1-\epsilon\right) \geq 1 - n^{-\rho_1}.$$

This along with (12) leads to the second result. ■

Lemma 3 *Under the conditions of Theorem 2, suppose that there exists $\rho \in (0, 1/3)$ such that $nN^{-\rho} = o(1)$ as $n \to \infty$. Let $\epsilon > 0$ and consider $\hat{m} = \sup\{l = 1, \ldots, n\colon n^{-1}\hat{\lambda}_l \geq nN^{-\rho}\epsilon\}$. Let α_n be a positive scalar sequence such that $\alpha_n \to \infty$ as $n \to \infty$. Then*

$$\begin{aligned}
\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F &= O_p(n^2 N^{-1/2}), \\
n^{-1}\|\hat{\mathbf{\Lambda}}_{\hat{m}} - \mathbf{\Lambda}_{\hat{m}}\|_{op,2} &= O_p(nN^{-1/2}), \\
\hat{m} &= m + O_p(\alpha_n^{-1}), \\
\|\hat{\tilde{\mathbf{\Sigma}}}^{-1} - \tilde{\mathbf{\Sigma}}^{-1}\|_{op,2} &= O_p(nN^{-1/2}).
\end{aligned}$$

Proof [Proof of Lemma 3] From Theorem 5.1 in Bobkov and Ledoux (2019), one has

$$E(\hat{D}_{ii}^2 | \boldsymbol{\theta}_i) = E(d_{\mathcal{W}_2}^2(\hat{\nu}_i, \nu_i) | \boldsymbol{\theta}_i) \leq 2(N+1)^{-1} J_2(\nu_i),$$

where $J_2(\nu_i) = \int_{\mathbb{R}} F_{\nu_i}(x)(1-F_{\nu_i}(x))/f_{\nu_i}(x)dx$ is the J_2 functional. Since by assumption $J_2(\nu_i) \leq \sup_{\nu_0\in\Omega} J_2(\nu_0) < \infty$, by a conditioning argument it follows that for any $i = 1, \ldots, n$,

$$E(d_{\mathcal{W}_2}^2(\hat{\nu}_i, \nu_i)) = O(N^{-1}), \qquad (14)$$

where the bound is uniform in i. By the triangle inequality,

$$d_{\mathcal{W}_2}(\hat{\nu}_i, \hat{\nu}_j) - d_{\mathcal{W}_2}(\nu_i, \nu_j) \leq d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j),$$

where $i, j = 1, \ldots, n$. Since $(\Omega, d_{\mathcal{W}_2})$ is totally bounded, it is bounded and thus there exists $T > 0$ such that $d_{\mathcal{W}_2}(\nu_i, \nu_j) \leq T$ for all $i, j = 1, \ldots, n$. Also, $d_{\mathcal{W}_2}(\hat{\nu}_i, \hat{\nu}_j) \leq d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j) + T$. Therefore

$$d^2_{\mathcal{W}_2}(\hat{\nu}_i, \hat{\nu}_j) - d^2_{\mathcal{W}_2}(\nu_i, \nu_j) \leq (d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j))^2 + 2T(d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j)).$$

Similarly,

$$d^2_{\mathcal{W}_2}(\nu_i, \nu_j) - d^2_{\mathcal{W}_2}(\hat{\nu}_i, \hat{\nu}_j) \leq (d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j))^2 + 2T(d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j)).$$

Hence

$$|[\hat{\mathbf{D}}_n]_{ij} - [\mathbf{D}_n]_{ij}| \leq (d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j))^2 + 2T(d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j)).$$

This combined with (14), Jensen's inequality, and the fact that $(x+y)^2 \leq 2(x^2+y^2)$ holds for all $x, y \in \mathbb{R}$ leads to

$$\begin{aligned}
&E|[\hat{\mathbf{D}}_n]_{ij} - [\mathbf{D}_n]_{ij}| \\
&\leq 2E(d^2_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d^2_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j)) + 2T(\{E(d^2_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i))\}^{1/2} + \{E(d^2_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j))\}^{1/2}) \\
&= O(N^{-1/2}),
\end{aligned}$$

where the bound is uniform in i, j and n. Therefore, since $\|\mathbf{z}\|_2 \leq \|\mathbf{z}\|_1$ holds for all $\mathbf{z} \in \mathbb{R}^{n^2}$, we obtain

$$\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F = O_p(n^2 N^{-1/2}), \tag{15}$$

and the first result follows. By Weyl's inequality and using that $\|\mathbf{J}_n\|_{\text{op},2} = 1$, we have

$$n^{-1}|\hat{\lambda}_l - \lambda_l| \leq n^{-1}\|(-1/2)\mathbf{J}_n\hat{\mathbf{D}}_n\mathbf{J}_n + (1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n\|_{\text{op},2} \leq (2n)^{-1}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F, \tag{16}$$

where $l = 1, \ldots, n$. This along with (15) shows the second result. Next, for large enough n and setting $\epsilon_n = nN^{-\rho}\epsilon$, observe

$$\begin{aligned}
&P(\hat{m} < m) \\
&\leq P\left(n^{-1}\hat{\lambda}_m < \epsilon_n\right) \\
&= P\left(n^{-1}(\hat{\lambda}_m - \lambda_m) + n^{-1}\lambda_m < \epsilon_n\right) \\
&\leq P\left(n^{-1}\lambda_m - \epsilon_n < n^{-1}|\hat{\lambda}_m - \lambda_m|\right) \\
&\leq P\left((n^{-1}\lambda_m - \lambda_{0m}) + \lambda_{0m} - \epsilon_n < n^{-1}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F\right) \\
&\leq P\left(-|n^{-1}\lambda_m - \lambda_{0m}| + \lambda_{0m} - \epsilon_n < n^{-1}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F\right) \\
&\leq P\left((\lambda_{0m} - \epsilon_n)/2 < n^{-1}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F\right) + P\left((\lambda_{0m} - \epsilon_n)/2 < |n^{-1}\lambda_m - \lambda_{0m}|\right) \\
&\leq P\left(\lambda_{0m}/4 < n^{-1}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F\right) + P\left(\lambda_{0m}/4 < |n^{-1}\lambda_m - \lambda_{0m}|\right),
\end{aligned}$$

where the third inequality follows from (16) and the last inequality is due to $\epsilon_n < \lambda_{0m}/2$ for large enough n since $\epsilon_n = o(1)$ as $n \to \infty$. This along with Lemma 1, (15), and observing that $\rho \in (0, 1/3)$ implies $nN^{-1/2} \leq nN^{-\rho} = o(1)$ so that $n^{-1}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F = o_p(1)$, whence

$$P\left(\hat{m} < m\right) = o(1), \tag{17}$$

as $n \to \infty$. From (15) and since $\lambda_{m+1} = 0$, which was shown in the proof of Lemma 1,

$$\begin{aligned} P\left(\hat{m} \geq m+1\right) \leq P\left(n^{-1}\hat{\lambda}_{m+1} \geq \epsilon_n\right) &= P\left(n^{-1}(\hat{\lambda}_{m+1} - \lambda_{m+1}) \geq \epsilon_n\right) \\ &\leq P\left(N^{1/2}n^{-2}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F \geq 2\epsilon N^{1/2-\rho}\right) \\ &= o(1), \end{aligned} \tag{18}$$

as $n \to \infty$, where the last equality is due to $N^{1/2-\rho} \to \infty$ as $n \to \infty$. Let $M > 0$. Similarly as in the proof of Lemma 2, we have

$$P\left(\alpha_n|\hat{m} - m| > M\right) \leq P\left(\hat{m} \geq m+1\right) + P\left(\hat{m} < m\right),$$

which along with (17) and (18) leads to the third result.

Next, since $\hat{m} = m + o_p(1)$ as $n \to \infty$ and both $\hat{m}$ and m are integer-valued, the event where $\hat{m} = m$ holds with probability tends to 1 and it suffices to work conditional on this event. Since $\boldsymbol{\eta}_i = \boldsymbol{\Lambda}_m^{1/2}\mathbf{Q}_m^T\mathbf{e}_i$, $i = 1, \ldots, n$, where the $\mathbf{e}_i$ are the canonical basis of $\mathbb{R}^m$, and $\sum_{i=1}^n \mathbf{e}_i\mathbf{e}_i^T = \mathbf{I}_m$, we have

$$n^{-1}\sum_{i=1}^n \boldsymbol{\eta}_i\boldsymbol{\eta}_i^T = n^{-1}\boldsymbol{\Lambda}_m^{1/2}\mathbf{Q}_m^T\mathbf{Q}_m\boldsymbol{\Lambda}_m^{1/2} = n^{-1}\boldsymbol{\Lambda}_m.$$

Similarly as in Oh et al. (2010), from $\mathbf{J}_n\mathbf{1}_n = \mathbf{0}_n$ and $(-1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$ one has $\boldsymbol{\Lambda}_m^{1/2}\mathbf{Q}_m^T\mathbf{1}_n = \mathbf{0}_m$, whence

$$\bar{\boldsymbol{\eta}} = n^{-1}\sum_{i=1}^n \boldsymbol{\eta}_i = n^{-1}\boldsymbol{\Lambda}_m^{1/2}\mathbf{Q}_m^T\mathbf{1}_n = \mathbf{0}_m.$$

Thus

$$\tilde{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n (\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})(\boldsymbol{\eta}_i - \bar{\boldsymbol{\eta}})^T = n^{-1}\sum_{i=1}^n \boldsymbol{\eta}_i\boldsymbol{\eta}_i^T = n^{-1}\boldsymbol{\Lambda}_m,$$

and combining with Lemma 1,

$$\|\tilde{\boldsymbol{\Sigma}}\|_{\mathrm{op},2} = \|\boldsymbol{\Sigma}_0\|_{\mathrm{op},2} + O_p(n^{-1/2}). \tag{19}$$

Similarly, since $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{\Lambda}}_m^{1/2}\hat{\mathbf{Q}}_m^T\mathbf{e}_i$, $i = 1, \ldots, n$, we also have $n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\eta}}_i = \mathbf{0}_m$ and

$$\tilde{\hat{\boldsymbol{\Sigma}}} = n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\eta}}_i\hat{\boldsymbol{\eta}}_i^T = n^{-1}\hat{\boldsymbol{\Lambda}}_m.$$

Therefore

$$\|\tilde{\hat{\boldsymbol{\Sigma}}} - \tilde{\boldsymbol{\Sigma}}\|_{\mathrm{op},2} = n^{-1}\|\hat{\boldsymbol{\Lambda}}_m - \boldsymbol{\Lambda}_m\|_{\mathrm{op},2} \leq \frac{1}{2n}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F = O_p(nN^{-1/2}), \tag{20}$$

where the last inequality follows from (15) and (16).

Since $\boldsymbol{\Sigma}_0$ is positive definite, which is due to (A2), there exists $0 < L_0 < M_0 < \infty$ such that $\|\boldsymbol{\Sigma}_0\|_{\mathrm{op},2} \in (L_0, M_0)$. Now, for any $\varepsilon \in (0, L_0)$ and in view of (19), we have that both events $\|\tilde{\boldsymbol{\Sigma}}\|_{\mathrm{op},2} \geq L_0 - \varepsilon > 0$ and $\|\tilde{\boldsymbol{\Sigma}}\|_{\mathrm{op},2} \leq M_0 + \varepsilon < \infty$ hold with probability tending to 1. It then suffices to work conditional on both events. From (20), the fact that $\|\tilde{\boldsymbol{\Sigma}}^{-1}\|_{\mathrm{op},2} = \|\tilde{\boldsymbol{\Sigma}}\|_{\mathrm{op},2}^{-1} \geq (M_0+\epsilon)^{-1}$, and using Lemma A.3 in Facer and Müller (2003) along with the relation $nN^{-1/2} \leq nN^{-\rho} = o(1)$ as $n \to \infty$, we obtain

$$\|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\|_{\mathrm{op},2} \leq c\|\tilde{\boldsymbol{\Sigma}}^{-1}\|_{\mathrm{op},2}^2\|\tilde{\hat{\boldsymbol{\Sigma}}} - \tilde{\boldsymbol{\Sigma}}\|_{\mathrm{op},2} \leq \frac{c}{L_0 - \varepsilon}\|\tilde{\hat{\boldsymbol{\Sigma}}} - \tilde{\boldsymbol{\Sigma}}\|_{\mathrm{op},2} = O_p(nN^{-1/2}),$$

where $c > 0$ is a constant. This shows the fourth result. ■

Proof [Proof of Theorem 2] Similarly as in the proof of Lemma 3 we work conditional on the event where $\hat{m} = m$, which holds with probability tending to 1 as $n \to \infty$. Also, from the proof of Lemma 3 we have $n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\eta}}_i = \mathbf{0}_m$ and $n^{-1}\sum_{i=1}^n \boldsymbol{\eta}_i = \mathbf{0}_m$, so that $\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) = 1 + \hat{\boldsymbol{\eta}}_i^T\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\hat{\boldsymbol{\eta}}_j$ and $\tilde{s}_{in}(\boldsymbol{\eta}_j) = 1 + \boldsymbol{\eta}_i^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}_j$. Observe

$$\begin{aligned}
&\|n^{-1}\sum_{i=1}^n[\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j)\hat{Q}_i - \tilde{s}_{in}(\boldsymbol{\eta}_j)Q_i]\Big\|_{L^2(0,1)} \\
&\leq \|n^{-1}\sum_{i=1}^n[(\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j))\hat{Q}_i]\|_{L^2(0,1)} + \|n^{-1}\sum_{i=1}^n \tilde{s}_{in}(\boldsymbol{\eta}_j)(\hat{Q}_i - Q_i)]\|_{L^2(0,1)} \\
&\leq n^{-1}\sum_{i=1}^n|\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j)|\ \|\hat{Q}_i\|_{L^2(0,1)} + \|n^{-1}\sum_{i=1}^n[s_{in}(\boldsymbol{\theta}_j)(\hat{Q}_i - Q_i)]\|_{L^2(0,1)} \\
&\leq n^{-1}\sum_{i=1}^n|\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j)|\ \|\hat{Q}_i\|_{L^2(0,1)} + \max_{i,j=1,\ldots,n}|s_{in}(\boldsymbol{\theta}_j)|\ n^{-1}\sum_{i=1}^n\|\hat{Q}_i - Q_i\|_{L^2(0,1)},
\end{aligned} \tag{21}$$

where the second inequality is due to $\tilde{s}_{in}(\boldsymbol{\eta}_j) = s_{in}(\boldsymbol{\theta}_j)$, which was shown in the proof of Theorem 1. From (14) we have

$$n^{-1}\sum_{i=1}^n\|\hat{Q}_i - Q_i\|_{L^2(0,1)} = O_p\left(N^{-1/2}\right), \tag{22}$$

so that it suffices to control the first term in the upper bound of (21), which we pursue next.

Observe

$$\begin{aligned}
|\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j)| &= |\hat{\boldsymbol{\eta}}_i^T\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_i^T\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}_j| \\
&\leq |(\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i)^T\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)| + |(\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i)^T\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\boldsymbol{\eta}_j| \\
&\quad + |\boldsymbol{\eta}_i^T\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}(\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j)| + |\boldsymbol{\eta}_i^T(\tilde{\hat{\boldsymbol{\Sigma}}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1})\boldsymbol{\eta}_j|.
\end{aligned} \tag{23}$$

Define auxiliary matrices $\mathbf{\Psi}_1 = (\hat{\boldsymbol{\eta}}_1 \cdots \hat{\boldsymbol{\eta}}_n) \in \mathbb{R}^{m\times n}$, $\mathbf{\Psi}_2 = (\boldsymbol{\eta}_1 \cdots \boldsymbol{\eta}_n) \in \mathbb{R}^{m\times n}$, and recall that the estimated MDS components $\mathbf{\Psi}_1$ satisfy $\mathbf{\Psi}_1 = \hat{\mathbf{\Lambda}}_m^{1/2}\hat{\mathbf{Q}}_m^T$. This implies $\hat{\boldsymbol{\eta}}_i = (\hat{\lambda}_1^{1/2}\hat{\nu}_{1i}, \ldots, \hat{\lambda}_m^{1/2}\hat{\nu}_{mi})^T$, where $\hat{\nu}_{li}$ is the ith component of $\hat{\boldsymbol{\nu}}_l$, $l = 1, \ldots, m$, $i = 1, \ldots, n$. Similarly we have $\boldsymbol{\eta}_i = (\lambda_1^{1/2}\nu_{1i}, \ldots, \lambda_m^{1/2}\nu_{mi})^T$ with ν_{li} the ith component of $\boldsymbol{\nu}_l$. Note that

$$
\begin{aligned}
&\|\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i\|_2^2 \\
&\le \sum_{l=1}^{m}(\hat{\lambda}_l^{1/2}\hat{\nu}_{li} - \lambda_l^{1/2}\nu_{li})^2 \\
&\le 2\sum_{l=1}^{m}\hat{\lambda}_l(\hat{\nu}_{li} - \nu_{li})^2 + 2\sum_{l=1}^{m}(\hat{\lambda}_l^{1/2} - \lambda_l^{1/2})^2\nu_{li}^2 \\
&\le 2\sum_{l=1}^{m}\lambda_l(\hat{\nu}_{li} - \nu_{li})^2 + 2\|\hat{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_m\|_{\mathrm{op},2}\sum_{l=1}^{m}(\hat{\nu}_{li} - \nu_{li})^2 + 2\|\hat{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_m\|_{\mathrm{op},2}^2\sum_{l=1}^{m}\lambda_l^{-1}\nu_{li}^2 \\
&\le 2\lambda_1 n^{-1}\sum_{l=1}^{m}\|\hat{\boldsymbol{\nu}}_l - \boldsymbol{\nu}_l\|_2^2 + 2n^{-1}\|\hat{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_m\|_{\mathrm{op},2}\sum_{l=1}^{m}\|\hat{\boldsymbol{\nu}}_l - \boldsymbol{\nu}_l\|_2^2 + 2mn^{-1}\lambda_m^{-1}\|\hat{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_m\|_{\mathrm{op},2}^2,
\end{aligned}
\tag{24}
$$

where the third inequality is due to the fact that $|\hat{\lambda}_l^{1/2} - \lambda_l^{1/2}| \le \lambda_l^{-1/2}|\hat{\lambda}_l - \lambda_l|$ and $|\hat{\lambda}_l - \lambda_l| \le \|\hat{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_m\|_{\mathrm{op},2}$, and the fourth inequality follows from the relation $a_j^2 \le n^{-1}\sum_{i=1}^{n}a_i^2$, which is valid for any scalars $a_j \in \mathbb{R}$, $j = 1, \ldots, n$, along with the fact that $\|\boldsymbol{\nu}_l\|_2 = 1$.

By Corollary 1 in Yu et al. (2015), for any $l = 1, \ldots, m$ we have

$$
\|\hat{\boldsymbol{\nu}}_l - \boldsymbol{\nu}_l\|_2 \le 2^{3/2}\delta^{-1}\|(-1/2)\mathbf{J}_n\hat{\mathbf{D}}_n\mathbf{J}_n + (1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n\|_{\mathrm{op},2} \le 2^{1/2}\delta^{-1}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_{\mathrm{op},2},
$$

where δ is defined as in Lemma 1 and we use that $\|\mathbf{J}_n\|_{\mathrm{op},2} = 1$. Using (24),

$$
\begin{aligned}
\|\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i\|_2^2 &\le 4m\lambda_1 n^{-1}\delta^{-2}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_{\mathrm{op},2}^2 + 4mn^{-1}\delta^{-2}\|\hat{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_m\|_{\mathrm{op},2}\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_{\mathrm{op},2}^2 \\
&\quad + 2mn^{-1}\lambda_m^{-1}\|\hat{\mathbf{\Lambda}}_m - \mathbf{\Lambda}_m\|_{\mathrm{op},2}^2 \\
&= U_n,
\end{aligned}
$$

where $i = 1, \ldots, n$, and U_n is defined through the last equality and does not depend on i. Using Lemma 1 and Lemma 3,

$$
\|\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i\|_2^2 \le U_n = O_p\left(n^3N^{-1}\right). \tag{25}
$$

Also note that for any $i = 1, \ldots, n$,

$$
\|\boldsymbol{\eta}_i\|_2^2 = \|\mathbf{\Lambda}_m^{1/2}\mathbf{Q}_m^T\mathbf{e}_i\|_2^2 = \sum_{l=1}^{m}\lambda_l\nu_{li}^2 \le n^{-1}\sum_{l=1}^{m}\lambda_l\sum_{i=1}^{n}\nu_{li}^2 \le n^{-1}\lambda_1,
$$

which combined with Lemma 1 leads to

$$
\|\boldsymbol{\eta}_i\|_2 \le n^{-1}\lambda_1 = O_p(1). \tag{26}
$$

From the proof of Lemma 3, we have $\|\tilde{\boldsymbol{\Sigma}}^{-1}\|_{\text{op},2} = O_p(1)$, whence with Lemma 3,

$$\|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\|_{\text{op},2} \leq \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\|_{\text{op},2} + \|\tilde{\boldsymbol{\Sigma}}^{-1}\|_{\text{op},2} = O_p(1). \tag{27}$$

In view of (23), (25), and (26), it follows that

$$\begin{aligned}
|\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j)| &\leq \|\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i\|_2 \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\|_{\text{op},2} \|\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_2 + \|\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i\|_2 \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\|_{\text{op},2} \|\boldsymbol{\eta}_j\|_2 \\
&\quad + \|\boldsymbol{\eta}_i\|_2 \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\|_{\text{op},2} \|\hat{\boldsymbol{\eta}}_j - \boldsymbol{\eta}_j\|_2 + \|\boldsymbol{\eta}_i\|_2 \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\|_{\text{op},2} \|\boldsymbol{\eta}_j\|_2 \\
&\leq U_n \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\|_{\text{op},2} + n^{-1} U_n^{1/2} \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1}\|_{\text{op},2} \lambda_1 + (n^{-1}\lambda_1)^2 \|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\|_{\text{op},2} \\
&= R_n,
\end{aligned}$$

where $j = 1, \ldots, n$, and R_n as defined through the last equality is uniform in j. Thus

$$n^{-1} \sum_{i=1}^{n} |\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j)| \; \|\hat{Q}_i\|_{L^2(0,1)} \leq R_n \; n^{-1} \sum_{i=1}^{n} \|\hat{Q}_i - Q_i\|_{L^2(0,1)} + R_n \; n^{-1} \sum_{i=1}^{n} \|Q_i\|_{L^2(0,1)},$$

where $R_n = O_p(n^{3/2} N^{-1/2})$, which follows from (25), (27), Lemma 1 and Lemma 3. This combined with (22), the fact that $E(\|Q_1\|_{L^2(0,1)}) \leq [E(\|Q_1\|^2_{L^2(0,1)})]^{1/2} < \infty$ and the weak law of large numbers leads to

$$n^{-1} \sum_{j=1}^{n} \left[n^{-1} \sum_{i=1}^{n} |\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j)| \; \|\hat{Q}_i\|_{L^2(0,1)} \right] = O_p(n^{3/2} N^{-1/2}).$$

Combining this with (21), (22), $\max_{i,j=1,\ldots,n} |s_{in}(\boldsymbol{\theta}_j)| = O_p(1)$, which follows by employing the developments in the proof of Theorem 1 in Petersen and Müller (2019) along with compactness of Θ, and the fact that

$$d_{\mathcal{W}_2}(\tilde{\hat{\nu}}_{\oplus}(\hat{\boldsymbol{\eta}}_j), \tilde{\nu}_{\oplus}(\boldsymbol{\eta}_j)) \leq \|n^{-1} \sum_{i=1}^{n} [\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) \hat{Q}_i - \tilde{s}_{in}(\boldsymbol{\eta}_j) Q_i]\|_{L^2(0,1)},$$

which is due to properties of the orthogonal projection from the Hilbert space $L^2(0,1)$ onto the closed and convex subset $\mathcal{Q}_\Omega$, leads to the first result.

Next, analogous arguments as in the proof of Theorem 1 show that $d_{\mathcal{W}_2}(\tilde{\nu}_{\oplus}(\boldsymbol{\eta}_j), \nu_{\oplus}(\boldsymbol{\theta}_j)) \leq \sup_{\|\boldsymbol{\theta}_0\|_2 \leq R} d_{\mathcal{W}_2}(\hat{\nu}_{\oplus}(\boldsymbol{\theta}_0), \nu_{\oplus}(\boldsymbol{\theta}_0))$ with $R = \text{diam}(\Theta) < \infty$. The second result follows from the triangle inequality, using that Ω is totally bounded, and analogous arguments as the ones in Proposition 1 in Petersen and Müller (2019) which show that the conditions of Theorem 2 are satisfied, whence $\sup_{\|\boldsymbol{\theta}_0\|_2 \leq R} d_{\mathcal{W}_2}(\hat{\nu}_{\oplus}(\boldsymbol{\theta}_0), \nu_{\oplus}(\boldsymbol{\theta}_0)) = O_p\left(n^{-1/(2(\alpha'-1))}\right)$ for any $\alpha' > 2$.

Finally, the last result follows by analogous arguments as before by observing the relation

$$|\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_{\mathbf{p}}) - \tilde{s}_{in}(\boldsymbol{\eta}_{\mathbf{p}})| \leq \sum_{j=1}^{n} p_j |\tilde{\hat{s}}_{in}(\hat{\boldsymbol{\eta}}_j) - \tilde{s}_{in}(\boldsymbol{\eta}_j)|,$$

where $\sum_{j=1}^n p_j = 1$ with $p_j > 0$ for all $j = 1, \ldots, n$, and $\boldsymbol{\eta}_{\mathbf{p}} = \sum_{j=1}^n p_j \boldsymbol{\eta}_j$. ■

Proof [Proof of Corollary 3]

Note that $d_{\mathcal{W}_2}(\hat{\nu}_i, \hat{\nu}_j) \leq \mathcal{T}$ and $d_{\mathcal{W}_2}(\nu_i, \nu_j) \leq \mathcal{T}$. Similarly as in the proof of Lemma 3, we have

$$|[\hat{\mathbf{D}}_n]_{ij} - [\mathbf{D}_n]_{ij}| \leq 2\mathcal{T}(d_{\mathcal{W}_2}(\hat{\nu}_i, \nu_i) + d_{\mathcal{W}_2}(\hat{\nu}_j, \nu_j)),$$

and

$$\begin{aligned} E(\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F^2) &= \sum_{i=1}^n \sum_{j=1}^n E\{([\hat{\mathbf{D}}_n]_{ij} - [\mathbf{D}_n]_{ij})^2\} \\ &\leq 8\mathcal{T}^2 \sum_{i=1}^n \sum_{j=1}^n E(d_{\mathcal{W}_2}^2(\hat{\nu}_i, \nu_i)) + E(d_{\mathcal{W}_2}^2(\hat{\nu}_j, \nu_j)) \\ &= O(n^2 N^{-1}), \end{aligned}$$

where the last equality follows from (14). This shows that $\|\hat{\mathbf{D}}_n - \mathbf{D}_n\|_F = O_p(nN^{-1/2})$ and analogous arguments as in the proof of Lemma 3 imply $n^{-1}\|\hat{\boldsymbol{\Lambda}}_{\hat{m}} - \boldsymbol{\Lambda}_{\hat{m}}\|_{\text{op},2} = O_p(N^{-1/2})$, $\hat{m} = m + O_p(\alpha_n^{-1})$, and $\|\tilde{\hat{\boldsymbol{\Sigma}}}^{-1} - \tilde{\boldsymbol{\Sigma}}^{-1}\|_{\text{op},2} = O_p(N^{-1/2})$. The result follows from $E(\|Q_1\|_{L^2(0,1)}^2) \leq \mathcal{T}^2 < \infty$ and analogous arguments as in the proof of Theorem 2. ■

Proof [Proof of Theorem 3] Similar to the proof of Lemma 2, the event where $\tilde{m} = m$ holds with probability tending to 1 and it thus suffices to work conditional on this event. Similar to the proof of Theorem 1, there exists an invertible matrix $\mathbf{R}_n \in \mathbb{R}^{m \times m}$ and a vector $\mathbf{b}_n \in \mathbb{R}^m$ such that $\boldsymbol{\eta}_i = \mathbf{R}_n \tilde{\boldsymbol{\theta}}_i + \mathbf{b}_n$ for all $i = 1, \ldots, n$. Define

$$\tilde{\boldsymbol{\Theta}}_n = \begin{pmatrix} \tilde{\boldsymbol{\theta}}_1^T \\ \vdots \\ \tilde{\boldsymbol{\theta}}_n^T \end{pmatrix} \in \mathbb{R}^{n \times m},$$

and $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^T$, and $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^T$. Recall that the MDS components are centered across the sample, i.e. $\sum_{i=1}^n \boldsymbol{\eta}_i = \mathbf{0}_{m \times 1}$ so that $\mathbf{1}_n^T \mathbf{M}_n = \mathbf{0}_{1 \times m}$. Since

$$\mathbf{Y}_n = (\mathbf{1}_n \ \tilde{\boldsymbol{\Theta}}_n) \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\epsilon}_n,$$

it follows that the least squares predictions are given by

$$\begin{aligned} \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{Y}_n &= (\mathbf{1}_n \ \mathbf{M}_n) \begin{pmatrix} n^{-1} & \mathbf{0}_{1\times m} \\ \mathbf{0}_{m\times 1} & (\mathbf{M}_n^T\mathbf{M}_n)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1}_n^T \\ \mathbf{M}_n^T \end{pmatrix} \mathbf{Y}_n \\ &= \begin{pmatrix} n^{-1}\mathbf{1}_n & \mathbf{M}_n(\mathbf{M}_n^T\mathbf{M}_n)^{-1} \end{pmatrix} \begin{pmatrix} n & \mathbf{1}_n^T\tilde{\boldsymbol{\Theta}}_n \\ \mathbf{0}_{m\times 1} & \mathbf{M}_n^T\tilde{\boldsymbol{\Theta}}_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} \\ &+ \begin{pmatrix} n^{-1}\mathbf{1}_n & \mathbf{M}_n(\mathbf{M}_n^T\mathbf{M}_n)^{-1} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n \epsilon_i \\ \mathbf{M}_n^T\boldsymbol{\epsilon}_n \end{pmatrix} \\ &= \tilde{\mathbf{R}}_1 + \tilde{\mathbf{R}}_2, \end{aligned}$$

where $\mathbf{R}_1$ and $\mathbf{R}_2$ are defined by the last equality. Let $\tilde{\mathbf{b}}_n = \mathbf{R}_n^T \mathbf{b}_n$ and observe $\mathbf{M}_n = \tilde{\mathbf{\Theta}}_n \mathbf{R}_n^T + \mathbf{1}_n \mathbf{b}_n^T = \tilde{\mathbf{\Theta}}_n \mathbf{R}_n^T + \mathbf{1}_n \tilde{\mathbf{b}}_n^T \mathbf{R}_n^T$. Since $\mathbf{1}_n^T \mathbf{M}_n = \mathbf{0}_{1\times m}$, it follows that $\mathbf{1}_n \tilde{\mathbf{b}}_n^T \mathbf{R}_n^T = -n^{-1} \mathbf{1}_n \mathbf{1}_n^T \tilde{\mathbf{\Theta}}_n \mathbf{R}_n^T$ which shows that $\mathbf{M}_n = \mathbf{J}_n \tilde{\mathbf{\Theta}}_n \mathbf{R}_n^T$. Thus $\mathbf{M}_n^T \mathbf{M}_n = \mathbf{R}_n (\tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n) \mathbf{R}_n^T$ and $\mathbf{M}_n (\mathbf{M}_n^T \mathbf{M}_n)^{-1} \mathbf{M}_n^T = \mathbf{J}_n \tilde{\mathbf{\Theta}}_n (\tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n)^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n$ and therefore

$$\tilde{\mathbf{R}}_1 = \begin{pmatrix} \mathbf{1}_n & \tilde{\mathbf{\Theta}}_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \beta_0 + \tilde{\boldsymbol{\theta}}_1^T \boldsymbol{\beta} \\ \vdots \\ \beta_0 + \tilde{\boldsymbol{\theta}}_n^T \boldsymbol{\beta} \end{pmatrix},$$

which corresponds to the regression function $E(Y|\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_0)$ at the sample points $\tilde{\boldsymbol{\theta}}_0 = \tilde{\boldsymbol{\theta}}_i$. Next, note that for any $j = 1, \ldots, n$,

$$\mathbf{e}_j^T \tilde{\mathbf{R}}_2 = n^{-1} \sum_{i=1}^n \epsilon_i + \mathbf{e}_j^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n (\tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n)^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \boldsymbol{\epsilon}_n = n^{-1} \sum_{i=1}^n \epsilon_i + U_{jn},$$

where U_{jn} is defined through the last equality. By conditioning and using that $E(U_{jn}|\tilde{\mathbf{\Theta}}_n) = 0$ and Weyl's inequality, we obtain

$$\begin{aligned}
\mathrm{Var}(U_{jn}) = E(\mathrm{Var}(U_{jn}|\tilde{\mathbf{\Theta}}_n)) &= \sigma^2 n^{-1} E(\mathbf{e}_j^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n (n^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n)^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \mathbf{e}_j) \\
&\leq \sigma^2 n^{-1} E\left(\lambda_{\min}\left(n^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n \right) \| \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \mathbf{e}_j \|_2^2 \right) \\
&\leq \sigma^2 n^{-1} E\left([\lambda_{\min}(\mathbf{\Sigma}_0) + \| n^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n - \mathbf{\Sigma}_0 \|_{\mathrm{op},2}] \| \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \mathbf{e}_j \|_2^2 \right) \\
&= \sigma^2 n^{-1} E\left([\lambda_{\min}(\mathbf{\Sigma}_0) + \| n^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n - \mathbf{\Sigma}_0 \|_{\mathrm{op},2}] \| \tilde{\boldsymbol{\theta}}_j - \bar{\tilde{\boldsymbol{\theta}}}_n \|_2^2 \right) \\
&= O(n^{-1}),
\end{aligned}$$

where $\bar{\tilde{\boldsymbol{\theta}}}_n = n^{-1} \sum_{i=1}^n \tilde{\boldsymbol{\theta}}_i$ and the bound is uniform in j which follows from the compactness of Θ along with standard arguments and noting that $n^{-1} \tilde{\mathbf{\Theta}}_n^T \mathbf{J}_n \tilde{\mathbf{\Theta}}_n$ is the sample covariance matrix of the $\tilde{\boldsymbol{\theta}}_i$. Therefore

$$\begin{aligned}
\mathbf{e}_j^T \tilde{\mathbf{R}}_2 &= O_p(n^{-1/2}), \\
\mathbf{e}_j^T \mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n &= E(Y|\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_j) + O_p(n^{-1/2}).
\end{aligned}$$

Recall that $P_i = \mathbf{e}_i^T \mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n$ is the ith prediction, $i = 1, \ldots, n$, and the average squared prediction $S_n = n^{-1} \sum_{i=1}^n (P_i - E(Y|\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}_i))^2$. Observe

$$\begin{aligned}
ES_n &= n^{-1} \sum_{i=1}^n E[(\mathbf{e}_i^T \tilde{\mathbf{R}}_2)^2] \\
&= n^{-1} \sum_{i=1}^n E\Big[\Big(n^{-1} \sum_{i=1}^n \epsilon_i + \mathbf{e}_i^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n (\tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n)^{-1} \tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \boldsymbol{\epsilon}_n\Big)^2\Big] \\
&= n^{-1}\sigma^2 + n^{-1} \sum_{i=1}^n E\Big[\Big(\mathbf{e}_i^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n (\tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n)^{-1} \tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \boldsymbol{\epsilon}_n\Big)^2\Big] \\
&= n^{-1}\sigma^2 + \sigma^2 n^{-1} \sum_{i=1}^n \mathrm{Var}(U_{in}) \\
&= O(n^{-1}),
\end{aligned}$$

where the third equality follows from $\mathbf{J}_n \mathbf{1}_n = \mathbf{0}_n$ and the bound follows from the fact that $\mathrm{Var}(U_{in}) = O(n^{-1})$ uniformly in i, which was shown before. Thus $S_n = O_p(n^{-1})$. ■

Proof [Proof of Theorem 4] Similar to the proof of Theorem 1, there exists an invertible matrix $\mathbf{R}'_n \in \mathbb{R}^{m \times m}$ and a vector $\mathbf{b}'_n \in \mathbb{R}^m$ such that $\boldsymbol{\eta}'_i = \mathbf{R}'_n \tilde{\boldsymbol{\theta}}'_i + \mathbf{b}'_n$ for all $i = 1, \ldots, n$. Since the MDS components $\boldsymbol{\eta}'_i$ are centered, we have $\mathbf{b}'_n = -n^{-1} \sum_{i=1}^n \mathbf{R}'_n \tilde{\boldsymbol{\theta}}'_i$ and thus $\boldsymbol{\eta}'_i = \mathbf{R}'_n(\tilde{\boldsymbol{\theta}}'_i - \bar{\tilde{\boldsymbol{\theta}}}')$ holds for all $i = 1, \ldots, n$, where $\bar{\tilde{\boldsymbol{\theta}}}' = n^{-1} \sum_{i=1}^n \tilde{\boldsymbol{\theta}}'_i$. Recall that $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})^T$ is the intercept vector, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m \in \mathbb{R}^m$ are slope vectors and $\mathbf{B}_m = (\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_m) \in \mathbb{R}^{m \times m}$. By the multivariate multiple linear regression model,

$$\tilde{\boldsymbol{\theta}}'_i = \begin{pmatrix} & \boldsymbol{\beta}_1^T \\ \boldsymbol{\beta}_0 & \vdots \\ & \boldsymbol{\beta}_m^T \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\boldsymbol{\theta}}_i \end{pmatrix} + \boldsymbol{\epsilon}_i,$$

where $i = 1, \ldots, n$. Thus, using that $\boldsymbol{\eta}'_i = \mathbf{R}'_n(\tilde{\boldsymbol{\theta}}'_i - \bar{\tilde{\boldsymbol{\theta}}}')$, it follows that

$$\boldsymbol{\eta}'_i = \mathbf{R}'_n \begin{pmatrix} \boldsymbol{\beta}_1^T(\tilde{\boldsymbol{\theta}}_i - \bar{\tilde{\boldsymbol{\theta}}}) \\ \vdots \\ \boldsymbol{\beta}_m^T(\tilde{\boldsymbol{\theta}}_i - \bar{\tilde{\boldsymbol{\theta}}}) \end{pmatrix} + \mathbf{R}'_n(\boldsymbol{\epsilon}_i - \bar{\boldsymbol{\epsilon}}), \tag{28}$$

where $\bar{\boldsymbol{\epsilon}} = n^{-1} \sum_{i=1}^n \boldsymbol{\epsilon}_i$ and $i = 1, \ldots, n$. Recall that the response matrix $\mathbf{Y}_n \in \mathbb{R}^{n \times m}$ is given by

$$\mathbf{Y}_n = \begin{pmatrix} \boldsymbol{\eta}'^T_1 \\ \vdots \\ \boldsymbol{\eta}'^T_n \end{pmatrix},$$

which contains on each row the MDS components from the sample of distributions $\nu_1', \ldots, \nu_n'$. With $\tilde{\boldsymbol{\Theta}}_n$ as in the proof of Theorem 3, in view of (28), it follows that

$$\mathbf{Y}_n = \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n \mathbf{B}_{m'} {\mathbf{R}_n'}^T + \mathbf{J}_n \begin{pmatrix} {\boldsymbol{\epsilon}_1}^T \\ \vdots \\ {\boldsymbol{\epsilon}_n}^T \end{pmatrix} {\mathbf{R}_n'}^T,$$

Define the auxiliary matrix $\mathbf{E}_n \in \mathbb{R}^{n \times m}$ whose ith row is given by $\boldsymbol{\epsilon}_i^T$, $i = 1, \ldots, n$. The (multivariate) fitted values are given by

$$\mathbf{X}_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{Y}_n = \begin{pmatrix} n^{-1}\mathbf{1}_n & \mathbf{M}_n (\mathbf{M}_n^T \mathbf{M}_n)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0}_n^T \\ \mathbf{M}_n^T \end{pmatrix} (\tilde{\boldsymbol{\Theta}}_n \mathbf{B}_m {\mathbf{R}_n'}^T + \mathbf{E}_n {\mathbf{R}_n'}^T) = \mathbf{H}_n \mathbf{U}_n,$$

where $\mathbf{H}_n = \mathbf{M}_n (\mathbf{M}_n^T \mathbf{M}_n)^{-1} \mathbf{M}_n^T$, $\mathbf{U}_n = \tilde{\boldsymbol{\Theta}}_n \mathbf{B}_m {\mathbf{R}_n'}^T + \mathbf{E}_n {\mathbf{R}_n'}^T$, and the first equality follows by using that $\mathbf{J}_n \mathbf{1}_n = \mathbf{0}_n$ and $\mathbf{M}_n^T \mathbf{1}_n = \mathbf{0}_m$ which implies $\mathbf{M}_n^T \mathbf{J}_n = \mathbf{M}_n^T$. Recall the m-dimensional predictors $\mathbf{Z}_i = (\mathbf{H}_n \mathbf{U}_n)^T \mathbf{e}_i$, $i = 1, \ldots, n$, with estimated covariance $\tilde{\boldsymbol{\Sigma}}_{\mathbf{Z}} = n^{-1} \mathbf{U}_n^T \mathbf{H}_n \mathbf{U}_n \in \mathbb{R}^{m \times m}$, which follows by noting that $\mathbf{H}_n \mathbf{1}_n = \mathbf{0}_n$ so that $\bar{\mathbf{Z}} = \mathbf{0}_m$. The global weights evaluated at points $x = \mathbf{Z}_j$, with $j \in \{1, \ldots, n\}$, are given by

$$\hat{s}_{in}(x) = 1 + (\mathbf{Z}_i - \bar{\mathbf{Z}})^T \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}^{-1} (\tilde{\boldsymbol{\gamma}}_0 - \bar{\mathbf{Z}}) = 1 + n \mathbf{e}_i^T \mathbf{H}_n \mathbf{U}_n (\mathbf{U}_n^T \mathbf{H}_n \mathbf{U}_n)^{-1} \mathbf{U}_n^T \mathbf{H}_n \mathbf{e}_j.$$

Observe

$$(\mathbf{U}_n^T \mathbf{H}_n \mathbf{U}_n)^{-1} = (\mathbf{U}_n^T \mathbf{M}_n (\mathbf{M}_n^T \mathbf{M}_n)^{-1} \mathbf{M}_n^T \mathbf{U}_n)^{-1} = (\mathbf{M}_n^T \mathbf{U}_n)^{-1} (\mathbf{M}_n^T \mathbf{M}_n) (\mathbf{U}_n^T \mathbf{M}_n)^{-1},$$

so that

$$\mathbf{H}_n \mathbf{U}_n (\mathbf{U}_n^T \mathbf{H}_n \mathbf{U}_n)^{-1} \mathbf{U}_n^T \mathbf{H}_n = \mathbf{H}_n.$$

From the proof of Theorem 3, we have $\mathbf{H}_n = \mathbf{M}_n (\mathbf{M}_n^T \mathbf{M}_n)^{-1} \mathbf{M}_n^T = \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n (\tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n)^{-1} \tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n$, whence

$$\hat{s}_{in}(x) = 1 + n \mathbf{e}_i^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n (\tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n)^{-1} \tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \mathbf{e}_j.$$

This shows that the global weights $\hat{s}_{in}(x)$, $i = 1, \ldots, n$, at a point $x = \mathbf{Z}_j$ are exactly the ones obtained by employing the predictor distribution parameters $\tilde{\boldsymbol{\Theta}}_n$ at a point $x = \tilde{\boldsymbol{\Theta}}_n^T \mathbf{e}_j$. Denote by $\mathbf{U}_i = \boldsymbol{\beta}_0 + \mathbf{B}_m^T \tilde{\boldsymbol{\Theta}}_n^T \mathbf{e}_i \in \mathbb{R}^m$, $\bar{\mathbf{U}} = \boldsymbol{\beta}_0 + n^{-1} \sum_{i=1}^n \mathbf{U}_i = n^{-1} \mathbf{B}_m^T \tilde{\boldsymbol{\Theta}}_n^T \mathbf{1}_n$ and $\boldsymbol{\Sigma}_{\mathbf{U}} = n^{-1} \sum_{i=1}^n (\mathbf{U}_i - \bar{\mathbf{U}})(\mathbf{U}_i - \bar{\mathbf{U}})^T = n^{-1} \mathbf{B}_m^T \tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n \mathbf{B}_m$. Observe

$$1 + (\mathbf{U}_i - \bar{\mathbf{U}})^T \boldsymbol{\Sigma}_{\mathbf{U}}^{-1} (\mathbf{U}_j - \bar{\mathbf{U}}) = 1 + n (\mathbf{e}_i^T - n^{-1} \mathbf{1}_n^T) \tilde{\boldsymbol{\Theta}}_n (\tilde{\boldsymbol{\Theta}}_n^T \mathbf{J}_n \tilde{\boldsymbol{\Theta}}_n)^{-1} \tilde{\boldsymbol{\Theta}}_n^T (\mathbf{e}_j - n^{-1} \mathbf{1}_n) = \hat{s}_{in}(x),$$

where the first equality uses that the slope matrix $\mathbf{B}_m$ is invertible. Thus the global weights $\hat{s}_{in}$ at points $x = \mathbf{Z}_j$ are exactly the same as the ones obtained by employing as predictors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$ at a point $x = \boldsymbol{\gamma}_j$, $j = 1, \ldots, n$. This implies $\tilde{\nu}_{\oplus}'(\mathbf{Z}_j) = \hat{\nu}_{\oplus}'(\boldsymbol{\gamma}_j)$. The result follows from Theorem 2 in Petersen and Müller (2019) and using that $\|\boldsymbol{\gamma}\|_2 = \|\boldsymbol{\beta}_0 + \mathbf{B}_m^T \tilde{\boldsymbol{\theta}}\|_2$ is uniformly bounded due to the compactness of Θ. ■

Appendix C. Computational Details

This section provides a general algorithm (see Algorithm 1) to obtain the estimated regression map $\tilde{\nu}_{\oplus}(\boldsymbol{\eta})$ in (4), employing global Fréchet regression (2) for the case of fully observed random objects situated in a general metric space (Ω, d). We note that computing the MDS components requires obtaining the eigendecomposition of a dense $n \times n$ matrix, which may be challenging for very large sample sizes n. Furthermore, as pointed out in Section 8, the optimization problem defining the global Fréchet regression depends on both the space Ω and the metric d. For the most important special cases there are algorithms that can solve this optimization problem, see e.g. Petersen and Müller (2019) for more details. Such spaces include covariance matrices with the Frobenius and other metrics, Hilbert spaces with the canonical metric, where the solution can in principle be found explicitly, and probability measures in 2-Wasserstein space, where the solution is equivalent to solving a quadratic optimization program.

For the important case of distributional data under uncertainty as discussed in Section 4, where only samples of observations coming from each underlying probability distribution are available, we refer to Algorithm 2. We remark that the optimization problem (6) can be solved along the lines of the quadratic program described in Petersen and Müller (2019). Indeed, denoting by $Q(\hat{\tilde{\nu}}_{\oplus}(\boldsymbol{\eta}))$ the quantile function corresponding to the probability measure $\hat{\tilde{\nu}}_{\oplus}(\boldsymbol{\eta})$, similar to the proof of Proposition 1 in Petersen and Müller (2019),

$$Q(\hat{\tilde{\nu}}_{\oplus}(\boldsymbol{\eta})) = \underset{q \in \{Q(w_0): w_0 \in \Omega\}}{\arg\min} ||n^{-1} \sum_{i=1}^{n} \tilde{\hat{s}}_{in}(\boldsymbol{\eta}) \hat{Q}_i - q||_{L^2(0,1)},$$

where $||\cdot||_{L^2}$ is the L^2 norm. Thus solving this problem for q over a dense grid of points in $(0,1)$ leads to the quadratic constrained optimization problem considered in Petersen and Müller (2019).

For the case where Ω consists of probability distributions with bi-Lipschitz quantile functions over $(0,1)$, a similar quadratic program has been considered in Gajardo and Müller (2022). In the applications and numerical examples presented here, we solve the optimization problem (6) by utilizing the frechet R package (Chen et al. 2020). While the theoretical asymptotic results for the estimation of the dimension m are tied to the growth rate of the lower bound $N(n)$ with sample size n, in practice one can select $\hat{m}$ based on a fraction-of-variance explained approach where m is selected such that $\sum_{j=1}^{\hat{m}} \hat{\lambda}_j / \sum_{j=1}^{n} \hat{\lambda}_j$ first upcrosses α, for large values of α such as 95%.

Input : Random objects $\nu_1, \dots, \nu_n$ and the metric d.
Output: The estimated map $\tilde{\nu}_\oplus$.

Obtain $\mathbf{\Lambda}, \mathbf{Q}$ from the spectral decomposition $(-1/2)\mathbf{J}_n\mathbf{D}_n\mathbf{J}_n = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ with corresponding ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$.
$\tilde{m} \leftarrow \sup\{l = 1, \dots, n \colon \lambda_l > 0\}$
for $j \leftarrow 1$ **to** $\tilde{m}$ **do**
  $\boldsymbol{\nu}_j \leftarrow$ j-th column of $\mathbf{Q}$
end
$\mathbf{Q}_{\tilde{m}} \leftarrow (\boldsymbol{\nu}_1 \cdots \boldsymbol{\nu}_{\tilde{m}}) \in \mathbb{R}^{n \times \tilde{m}}$
$\mathbf{\Lambda}_{\tilde{m}} \leftarrow \operatorname{diag}(\lambda_1, \dots, \lambda_{\tilde{m}}) \in \mathbb{R}^{\tilde{m} \times \tilde{m}}$
$\mathbf{\Psi}_{\tilde{m}} \leftarrow \mathbf{\Lambda}_{\tilde{m}}^{1/2}\mathbf{Q}_{\tilde{m}}^T$
for $i \leftarrow 1$ **to** n **do**
  $\boldsymbol{\eta}_i \leftarrow$ i-th column of $\mathbf{\Psi}_{\tilde{m}}$
end
$\tilde{\mathbf{\Sigma}} \leftarrow n^{-1}\sum_{i=1}^n \boldsymbol{\eta}_i\boldsymbol{\eta}_i^T$
For $\boldsymbol{\eta} \in \mathbb{R}^{\tilde{m}}$, construct the map $\tilde{\nu}_\oplus(\boldsymbol{\eta})$ as follows:
for $i \leftarrow 1$ **to** n **do**
  $\tilde{s}_{in}(\boldsymbol{\eta}) \leftarrow 1 + \boldsymbol{\eta}_i^T\tilde{\mathbf{\Sigma}}^{-1}\boldsymbol{\eta}$
end
$\tilde{\nu}_\oplus(\boldsymbol{\eta}) \leftarrow \underset{w \in \Omega}{\arg\min}\ n^{-1}\sum_{i=1}^n \tilde{s}_{in}(\boldsymbol{\eta}) d^2(\nu_i, w)$

Algorithm 1: Case of fully observed random objects.

Appendix D. Additional Simulations

We investigate here the effect of fitting the global Fréchet regression function with fewer MDS components than the true underlying low dimension m. The simulation example from Figure 3 in Section 6 illustrates the situation of Gaussian densities with two sources of variation, namely a mean shift and a variance shift, which leads to vertical variation. Since the random data generating mechanism produces each variation component independently (i.e., the mean and variance parameters are generated independently), the underlying correct dimension is $m = 2$ and the manifold satisfies the isometry condition (A1).

Investigating the case when one erroneously assumes that there is only one component, Figure 10 shows the densities resulting from global Fréchet regression fitted with only a single MDS component. This demonstrates that in this underspecified case the inverse map is not able to track the underlying geometry of the manifold, which is generated from two independent sources of variation. Instead, it produces a mixture of both mean and variance shift as if both variations were dependent on each other, while they were independently generated. Unsurprisingly, the image space of this inverse map does not coincide with the correct manifold. Specifically, it is unable to produce Gaussian densities with large mean and small variance (i.e., a large vertical variation pushed to the right), see e.g. the right panel in Figure 3, and neither does the isometry condition hold when setting $m = 1$.

Input : Sample of scalars $Y_{i1}, \ldots, Y_{in_i}$, $i = 1, \ldots, n$, and estimated dimension $\hat{m}$.
Output: The estimated map $\hat{\tilde{\nu}}_{\oplus}$.

for $i \leftarrow 1$ **to** n **do**
  $\hat{F}_i \leftarrow$ empirical CDF of the sample $Y_{i1}, \ldots, Y_{in_i}$
  $\hat{Q}_i \leftarrow$ left-continuous generalized inverse of $\hat{F}_i$
end
for $i \leftarrow 1$ **to** n **do**
  for $j \leftarrow 1$ **to** n **do**
    $[\hat{\mathbf{D}}_n]_{ij} \leftarrow \int_0^1 (\hat{Q}_i(t) - \hat{Q}_j(t))^2 dt$
  end
end
Obtain $\hat{\mathbf{\Lambda}}, \hat{\mathbf{Q}}$ from the spectral decomposition $(-1/2)\mathbf{J}_n\hat{\mathbf{D}}_n\mathbf{J}_n = \hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^T$ with corresponding ordered eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$.
for $j \leftarrow 1$ **to** $\hat{m}$ **do**
  $\hat{\boldsymbol{\nu}}_j \leftarrow$ j-th column of $\hat{\mathbf{Q}}$
end
$\hat{\mathbf{Q}}_{\hat{m}} \leftarrow (\hat{\boldsymbol{\nu}}_1 \cdots \hat{\boldsymbol{\nu}}_{\hat{m}}) \in \mathbb{R}^{n \times \hat{m}}$
$\hat{\mathbf{\Lambda}}_{\hat{m}} \leftarrow \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{m}}) \in \mathbb{R}^{\hat{m} \times \hat{m}}$
for $i \leftarrow 1$ **to** n **do**
  $\hat{\boldsymbol{\eta}}_i \leftarrow$ i-th column of $\hat{\mathbf{\Lambda}}_{\hat{m}}^{1/2}\hat{\mathbf{Q}}_{\hat{m}}^T$
end
$\hat{\tilde{\mathbf{\Sigma}}} \leftarrow n^{-1}\sum_{i=1}^n (\hat{\boldsymbol{\eta}}_i - n^{-1}\sum_{k=1}^n \hat{\boldsymbol{\eta}}_k)(\hat{\boldsymbol{\eta}}_i - n^{-1}\sum_{k=1}^n \hat{\boldsymbol{\eta}}_k)^T$
Let $\boldsymbol{\eta} \in \mathbb{R}^{\hat{m}}$.
for $i \leftarrow 1$ **to** n **do**
  $\hat{\tilde{s}}_{in}(\boldsymbol{\eta}) \leftarrow 1 + (\hat{\boldsymbol{\eta}}_i - n^{-1}\sum_{k=1}^n \hat{\boldsymbol{\eta}}_k)^T \hat{\tilde{\mathbf{\Sigma}}}^{-1}(\boldsymbol{\eta} - n^{-1}\sum_{k=1}^n \hat{\boldsymbol{\eta}}_k)$
end
$\hat{\tilde{\nu}}_{\oplus}(\boldsymbol{\eta}) \leftarrow \underset{w \in \Omega}{\arg\min}\ n^{-1}\sum_{i=1}^n \tilde{\hat{s}}_{in}(\boldsymbol{\eta}) d^2_{\mathcal{W}_2}(\hat{\nu}_i, w)$ where $\hat{\nu}_i$ has quantile function $\hat{Q}_i$.

Algorithm 2: Fréchet-Wasserstein Manifold Learning Under Uncertainty.

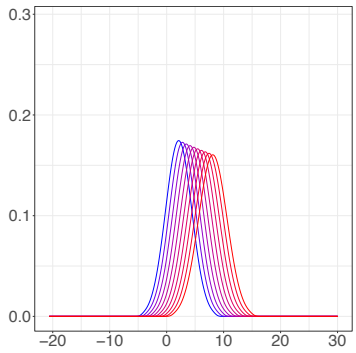

Figure 10: Density functions obtained with global Fréchet regression when using only a single MDS component (blue to red from lower to higher MDS values) for the Gaussian simulation setting in Figure 3, which has two independent sources of variation, corresponding to mean and vertical shifts.

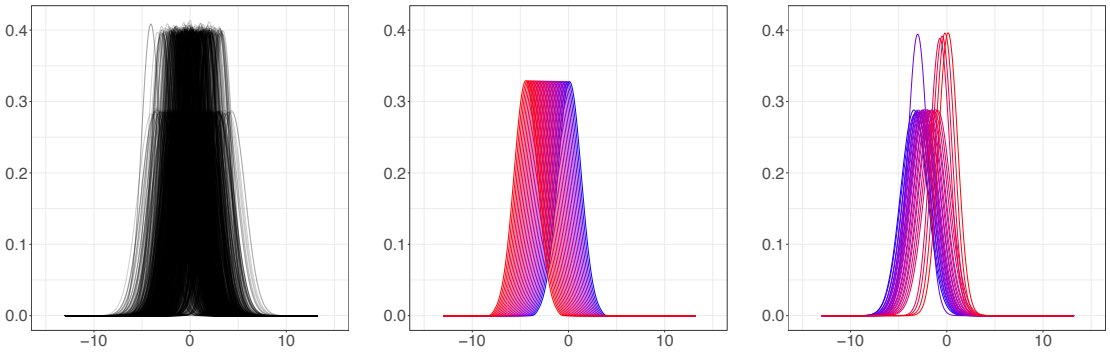

Figure 11: Comparison between the proposed inverse map back to distribution space with a naive nearest neighbor approach that predicts the inverse map as the distribution whose MDS component is closest to the unobserved MDS point. The simulation setting consists of Gaussian variates generated with two independent sources of variation giving rise to distinct mean and vertical shifts. The left panel shows the observed distributions f_i, the middle panel the proposed inverse map at unobserved MDS points along the horizontal line $(\psi_1, 0)^T \in \mathbb{R}^2$ with ψ_1 ranging from 0 to the 99% quantile of the observed first MDS components and the right panel the naive nearest neighbor approach to construct the inverse map.

Appendix E. Comparison with a naive nearest-neighbor approach

Upon the suggestion of a referee, we compared the proposed inverse map in the distributional case with a naive nearest neighbor approach where the predicted inverse map at an unobserved MDS point $\boldsymbol{\eta}$ is given by the distribution f_i such that its MDS component $\boldsymbol{\eta}_i$ is closest to $\boldsymbol{\eta}$ in the Euclidean norm. For this comparison, we considered a Gaussian location-scale family $\nu|(\mu, \sigma) \sim N(\mu, \sigma)$, where the parameters (μ, σ) were independently generated according to the following scheme.

First, random variables $\mu_i \overset{iid}{\sim} \mu$, $i = 1, \ldots, n$, are generated, where $\mu \sim \text{truncN}([a_1, b_1], \mu_1, \sigma_1)$ has a truncated Gaussian distribution over the compact interval $[a_1, b_1]$, $a_1 < b_1$, $\mu_1 \in \mathbb{R}$ and $\sigma_1 > 0$. Secondly, random variables $\sigma_i \overset{iid}{\sim} \sigma$ are generated from a mixture of truncated Gaussian variates over the compact interval $[a_2, b_2]$, $0 < a_2 < b_2$, as follows. For each $i = 1, \ldots, n$, with probability $p \in (0, 1)$ generate $\sigma_i \sim \text{truncN}([a_2, b_2], \mu_2, \sigma_2)$ and otherwise, with probability $1 - p$, generate $\sigma_i \sim \text{truncN}([a_2, b_2], \mu_3, \sigma_3)$, where $\mu_j \in \mathbb{R}$, $\sigma_j > 0$, $j = 1, 2$. The sample of Gaussian distributional objects is then given by $f_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, where we consider $n = 900$, $\mu_1 = 0$, $\sigma_1 = 1.5$, $a_1 = -10$, $b_1 = 10$, $p = 0.5$, $\mu_2 = 1.4$, $\sigma_2 = 0.01$, $\mu_3 = 1$, $\sigma_3 = 0.01$, $a_2 = 0.6$, $b_2 = 6$. Thus, there are two sources of variation in the distributional data consisting of a mean shift and a vertical variation as shown in the left panel of Figure 11.

We compared both inverse map methods across the unobserved MDS region consisting of the horizontal line $(\eta_1, 0)^T \in \mathbb{R}^2$ with η_1 ranging from 0 to the 99% quantile of the observed first MDS components, where $\boldsymbol{\eta} = (\eta_1, \eta_2)^T$. Figure 11 displays the results for the comparison between the proposed inverse method (middle panel) and the nearest neighbor approach (right panel). Clearly, the nearest neighbor inverse map approach is not able to disentangle both forms of variation and instead contains a mixture of both, especially towards the boundaries of the observed MDS components where a vertical variation in addition to mean shift is being predicted. However, the global Fréchet regression function is able to fully capture the mean shift single variation that is present along the horizontal line where the second MDS component is held constant at its mean value 0.

Another issue that affects the nearest neighbor approach is that it leads to an inverse map that is piecewise constant within the distribution space, since the nearest neighbors remain the same over smaller regions along the MDS line. This is problematic when there are sparsely populated or empty regions in the MDS space, as then the prediction would remain the same (at a constant level in the distribution space). The proposed inverse map does not inherit these problems and correctly captures the horizontal mean-shift variation represented by the first MDS component in a continuous fashion.

Acknowledgments

This research was conducted while Álvaro Gajardo was a PhD student at UC Davis and was supported in part by NSF grants DMS-2014626 and DMS-2310450. We wish to thank the reviewers for their insightful comments that led to substantial improvements in the paper.

References

Pedro C. Álvarez Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

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

P. Bickel and B. Li. Local polynomial regression on unknown manifolds. *Complex Datasets And Inverse Problems: Tomography, Networks And Beyond, IMS Lecture Notes-*

Monograph Series, 54:177–186, 2007.

Jérémie Bigot and Benjamin Charlier. On the consistency of Fréchet means in deformable models for curve and image analysis. *Electronic Journal of Statistics*, 5:1054 – 1089, 2011.

Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electronic Journal of Statistics*, 12(2):2253 – 2289, 2018.

Jrmie Bigot, Ral Gouet, Thierry Klein, and Alfredo Lpez. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1 – 26, 2017.

Sergey Bobkov and Michel Ledoux. *One-dimensional Empirical Measures, Order Statistics and Kantorovich Transport Distances*. American Mathematical Society, New York, 2019.

B M. Bolstad, R.A. Irizarry, M. Åstrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

Ingwer Borg and Patrick JF Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media, 2005.

Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, and Nicolas Papadakis. Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456, 2018.

D. Chen and Hans-Georg Müller. Nonlinear manifold representations for functional data. *The Annals of Statistics*, 40(1):1–29, 2012.

Yaqing Chen, Álvaro Gajardo, Jianing Fan, Qixian Zhong, Paromita Dubey, Kyunghee Han, Satarupa Bhattacharjee, and Hans-Georg Müller. *frechet: Statistical Analysis for Random Objects and Non-Euclidean Data*, 2020. URL `https://CRAN.R-project.org/package=frechet`. R package version 0.2.0.

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.

Alexander Cloninger, Keaton Hamm, Varun Khurana, and Caroline Moosmüller. Linearized wasserstein dimensionality reduction with approximation guarantees. *Applied and Computational Harmonic Analysis*, 74:101718, 2025.

Xiongtao Dai and Hans-Georg Müller. Principal component analysis for functional data on Riemannian manifolds and spheres. *The Annals of Statistics*, 46(6B):3334–3361, 2018.

P. Delicado. Dimensionality reduction when data are density functions. *Computational Statistics and Data Analysis*, 55(1):401–420, 2011.

Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

P. Drineas, A. Javed, M. Magdon-Ismail, G. Pandurangan, R. Virrankoski, and A. Savvides. Distance matrix reconstruction from incomplete distance information for sensor network localization. In *2006 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks*, volume 2, pages 536–544, 2006.

Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102 – 1123, 2009.

Paromita Dubey and Hans-Georg Müller. Modeling time-varying random objects and dynamic networks. *Journal of the American Statistical Association*, 540(117):2252–2267, 2022.

Bradley Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, 1978.

Bradley Efron. Curvature and inference for maximum likelihood estimates. *The Annals of Statistics*, 46(4):1664–1692, 2018.

Matthew R. Facer and Hans-Georg Müller. Nonparametric estimation of the location of a maximum in a response surface. *Journal of Multivariate Analysis*, 87(1):191–217, 2003.

Jianing Fan and Hans-Georg Müller. Conditional Wasserstein barycenters and interpolation/extrapolation of distributions. *IEEE Transactions on Information Theory*, 71(5): 363–410, 2025.

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

Álvaro Gajardo and Hans-Georg Müller. Cox point process regression. *IEEE Transactions on Information Theory*, 68(2):1133–1156, 2022.

Daniel Gervini. Spatial kriging for replicated temporal point processes. *Spatial Statistics*, 51:100681, 2022.

Daniel Gervini and Manoj Khanal. Exploring patterns of demand in bike sharing systems via replicated point process models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):585–602, 2019.

Aritra Ghosal, Wendy Meiring, and Alexander Petersen. Fréchet single index models for object response regression. *Electronic Journal of Statistics*, 17:1074–1112, 2023a.

Rahul Ghosal, Vijay R Varma, Dmitri Volfson, Inbar Hillel, Jacek Urbanek, Jeffrey M Hausdorff, Amber Watts, and Vadim Zipunnikov. Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's Disease. *Biostatistics*, 24(3):539–561, 2023b.

Cedric E Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, 11(2):725–750, 2017.

J.C. Gower. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985.

Florian F Gunsilius. Distributional synthetic controls. *Econometrica*, 91(3):1105–1117, 2023.

Keaton Hamm, Nick Henscheid, and Shujie Kang. Wassmap: Wasserstein isometric mapping for image manifold learning. *SIAM Journal on Mathematics of Data Science*, 5(2): 475–501, 2023.

Kyunghee Han, Hans-Georg Müller, and Byeong U. Park. Additive functional regression for densities as responses. *Journal of the American Statistical Association*, 115(530): 997–1010, 2020.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015.

Adel Javanmard and Andrea Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, 2013.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 2007.

Jurgen Kleffe. Principal components of random variables with values in a separable Hilbert space. *Mathematische Operationsforschung und Statistik*, 4:391–406, 1973.

Kuang-Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41 (1):221–249, 2013.

Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45(3):1059–1095, 2017.

Zhenhua Lin and Fang Yao. Intrinsic Riemannian functional data analysis. *The Annals of Statistics*, 47(6):3533–3577, 2019.

Marcos Matabuena, Alexander Petersen, Juan C Vidal, and Francisco Gude. Glucodensities: a new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464, 2021.

Hans-Georg Müller. Peter Hall, functional data analysis and random objects. *The Annals of Statistics*, 44(5):1867 – 1887, 2016.

Sewoong Oh, Andrea Montanari, and Amin Karbasi. Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pages 1–5. IEEE, 2010.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019.

Matteo Pegoraro and Mario Beraha. Projected statistical methods for distributional data on the real line with the Wasserstein metric. *Journal of Machine Learning Research*, 23: 1–59, 2022.

Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.

Alexander Petersen, Xi Liu, and Afshin A. Divani. Wasserstein F-tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics*, 49 (1):590 – 611, 2021.

Alexander Petersen, Chao Zhang, and Piotr Kokoszka. Modeling probability density functions as data objects. *Econometrics and Statistics*, 21:159–178, 2022.

James O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, 2016.

Samuel Stanley Wilks. Order statistics. *Bulletin of the American Mathematical Society*, 54 (1):6–50, 1948.

Yingcun Xia, Howell Tong, WK Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

Gale Young and Aiston S Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Qi Zhang, Lingzhou Xue, and Bing Li. Dimension reduction and data visualization for Fréchet regression. *arXiv preprint arXiv:2110.00467*, 2021.

Changbo Zhu and Hans-Georg Müller. Spherical autoregressive models, with application to distributional and compositional time series. *Journal of Econometrics*, 239(2):105389, 2024.