Submitted to the Annals of Statistics

FRÉCHET REGRESSION FOR RANDOM OBJECTS WITH EUCLIDEAN PREDICTORS

By Alexander Petersen* and Hans-Georg Müller†,‡

*Department of Statistics, University of California, Santa Barbara**
Department of Statistics, University of California, Davis†

Increasingly, statisticians are faced with the task of analyzing complex data that are non-Euclidean and specifically do not lie in a vector space. To address the need for statistical methods for such data, we introduce the concept of Fréchet regression. This is a general approach to regression when responses are complex random objects in a metric space and predictors are in $\mathcal{R}^p$, achieved by extending the classical concept of a Fréchet mean to the notion of a conditional Fréchet mean. We develop generalized versions of both global least squares regression and local weighted least squares smoothing. The target quantities are appropriately defined population versions of global and local regression for response objects in a metric space. We derive asymptotic rates of convergence for the corresponding fitted regressions using observed data to the population targets under suitable regularity conditions by applying empirical process methods. For the special case of random objects that reside in a Hilbert space, such as regression models with vector predictors and functional data as responses, we obtain a limit distribution. The proposed methods have broad applicability. Illustrative examples include responses that consist of probability distributions and correlation matrices, and we demonstrate both global and local Fréchet regression for demographic and brain imaging data. Local Fréchet regression is also illustrated via a simulation with response data which lie on the sphere.

1. Introduction. The regression relationship between a response variable and one or more predictor variables constitutes the target of many statistical methodologies. The most basic form is linear regression, where all variables are real-valued, and the conditional mean of the response variable is linear in the predictors. The linear regression model is quite flexible, includes polynomial fits and categorical predictor variables, among others, and remains one of the most popular tools for data analysis. In addition to the superb interpretability of linear models and simple model fitting via

‡Supported in part by National Science Foundation grants DMS-12-28369 and DMS-14-07852

MSC 2010 subject classifications: Primary 62G05; secondary 62J99, 62G08

Keywords and phrases: Least Squares Regression; Random Objects; Metric Spaces; Local Linear Regression; Functional Connectivity; Densities as Objects

least squares, powerful inferential methods, with well-established theory, are available for estimation and testing. Linear regression ideas also motivate local polynomial smoothing, further adding to their vast applicability.

In recent years, as data types are becoming more complex, attention has turned to regression in more abstract settings. The importance of the analysis of such object data has recently been highlighted (Marron and Alonso 2014; Wang et al. 2007). A setting that is increasingly encountered is that of a response variable taking values in a metric space, which may or may not have algebraic structure. The presence of a metric provides a natural connection to the work of Fréchet (1948), where the Fréchet mean is defined for random elements of a metric space as a direct generalization of the standard mean, which is defined by integration over a probability space. This generalization has been increasingly exploited in statistical analyses due to its inherent flexibility. Specifically, no ambient vector space needs to be assumed and only a distance between data objects is required. As regression can be viewed as the modeling of conditional means, a key feature of our approach is that we introduce the concept of a conditional Fréchet mean, generalizing the classical Fréchet mean.

One important class of random objects, which has been extensively studied, consists of observations on a finite-dimensional differentiable Riemannian manifold. Due to local Euclidean properties of the space, one can mimic both parametric (global) and nonparametric (local) regression techniques for standard Euclidean data quite effectively by local Euclidean approximations. Regression models for this special case have been well studied (Fisher, Lewis and Embleton 1987; Chang 1989; Prentice 1989; Fisher 1995), including intrinsic models for geodesic regression (Fletcher 2013; Niethammer, Huang and Vialard 2011; Cornea et al. 2016), semiparametric regression (Shi et al. 2009) and local kernel regression as a generalization of the classical Nadaraya-Watson smoother (Pelletier 2006; Davis et al. 2007; Hinkle et al. 2012; Yuan et al. 2012). Recently, the extrinsic regression model in Lin et al. (2015) extends the notion of extrinsic means (see, e.g., Ch. 11 and 18 of Patrangenaru and Ellingson 2015), where extrinsic approaches have been reported to have computational advantages (Bhattacharya et al. 2012).

In this paper, however, we go beyond manifolds and our focus is on a more general case of random objects in metric spaces with little structure, where only distances between response objects are computable. To our knowledge, in general metric spaces, the only global or parametric model which has been proposed is that of Faraway (2014), where data are represented as scores in a Euclidean space based on their pairwise distances, followed by the use of classical regression techniques. This method requires a complicated

"backscoring" step, where vectors in Euclidean space are then represented in the original metric space, and its theoretical properties have not been studied. Local regression methods on generic metric spaces are limited to Nadaraya-Watson type estimators (Davis et al. 2007; Hein 2009; Steinke and Hein 2009; Steinke, Hein and Schölkopf 2010) and lack a comprehensive asymptotic analysis. Thus, there is a need for additional statistical models to tackle this type of data that is increasingly common. Accordingly, we present here methodology and theory for both global and local regression analysis of complex random objects.

Specifically, we consider regression relationships between responses which are complex random objects and vectors of real-valued predictors. To this end, we develop a global regression relation as a generalization of multiple linear regression, as well as a class of more flexible local regression methods that generalizes local linear or polynomial regression. As the proposed regression approach for random objects incorporates the geometry implied by the metric and can be viewed as an extension of the Fréchet mean, we refer to our methods as Fréchet regression. *Global Fréchet regression* provides an improvement on the global method of Faraway (2014), as the proposed model defines the regression directly on the object space and does not require backscoring. The global Fréchet regression model constitutes a class of regression functions on arbitrary metric spaces which can be fitted without a tuning parameter or the need for any local smoothing technique. We also propose *local Fréchet regression*, which generalizes local linear estimation to a framework where responses are random objects, extending the available nonparametric regression methodology for object data. A challenge for the development of local Fréchet regression is to define an appropriate population model, which serves as the target to which the fitted local Fréchet regression converges. We establish consistency and rates of convergence for both global and local Fréchet regression.

The proposed global Fréchet regression model is introduced in Section 2, and theory quantifying the convergence rates of these estimators is given in Section 3, along with some concrete examples which are shown to satisfy the necessary regularity conditions. Local Fréchet regression is introduced in Section 4, along with asymptotic convergence theory. All proofs can be found in the online supplement. For the special case where the random objects take values in a Hilbert space, a limiting distribution can be obtained, as demonstrated in Section 5.

Our primary application examples deal with samples of probability distributions and correlation matrices, which are illustrated with data from demography and neuroimaging, with details in Sections 6 and 7, respectively.

Here, we also include a discussion of practical issues, such as a suitable notion of the coefficient of determination R^2 when the responses are random objects. For the space of probability distributions, we utilize the Wasserstein metric to conduct a simulation experiment as well as analyze the evolution of mortality profiles for two countries. For the case where responses are correlation matrices, we examine the relationship between functional connectivity in the brain, as quantified by pairwise correlations of fMRI signals, with age as predictor. Lastly, although the proposed methodology does not require any particular metric structure, it is nevertheless applicable to structured spaces such as manifolds. To demonstrate this, the local Fréchet regression technique is also illustrated with simulated manifold data on the sphere $S^2 \subset \mathcal{R}^3$ in Section 8.

2. Global Fréchet Regression.

2.1. *Preliminaries.* Let (Ω, d) be a metric space. We consider a random process $(X, Y) \sim F$, where X and Y take values in $\mathcal{R}^p$ and Ω, respectively, and F is the joint distribution of (X, Y) on $\mathcal{R}^p \times \Omega$. We denote the marginal distributions of X and Y as F_X and F_Y, respectively, and assume that $\mu = E(X)$ and $\Sigma = \mathrm{Var}(X)$ exist, with Σ positive definite. The conditional distributions $F_{X|Y}$ and $F_{Y|X}$ are also assumed to exist. In this general setting, we refer to Y as a random object. The usual notions of mean and variance were generalized to random objects in metric spaces in Fréchet (1948), where

$$\omega_\oplus = \operatorname*{argmin}_{\omega \in \Omega} E(d^2(Y, \omega)), \quad V_\oplus = E(d^2(Y, \omega_\oplus)) \tag{2.1}$$

were defined, now commonly referred to as Fréchet mean and Fréchet variance, respectively.

Building on these concepts, we introduce the Fréchet regression function of Y given $X = x$,

$$m_\oplus(x) = \operatorname*{argmin}_{\omega \in \Omega} M_\oplus(\omega, x), \quad M_\oplus(\cdot, x) = E(d^2(Y, \cdot)|X = x), \tag{2.2}$$

where we refer to $M_\oplus(\cdot, x)$ as the (conditional) Fréchet function. For the special case $\Omega = \mathcal{R}$, various nonparametric regression methods have been developed which are based on kernel or local linear polynomial fitting (Fan and Gijbels 1996), splines (Craven and Wahba 1979; Marx and Eilers 1996) or other smoothers.

A basic statistical task is to fit a global regression model for response Y and predictor X, in order to provide ease of implementation and interpretation and allow for good options for overall inference and testing. Fitting

of such a global model also does not require the choice of a tuning parameter, as all local fitting methods do, since global models are usually fitted under the assumption that there is no bias. Given that no algebraic structure is assumed, it is not feasible to directly generalize parametric models to a parametric function on Ω, as has been done in the special case when Ω is a Riemannian manifold. However, an alternative solution that we will develop is to recharacterize the standard multiple linear regression model as a function of weighted Fréchet means, where the weights have a known form and vary with x.

2.2. *Generalizing Linear Regression.* We begin by considering the standard setup for linear regression, for which $\Omega = \mathcal{R}$, and then write $m = m_{\oplus}$ in (2.2). The model for linear regression is

$$m(x) := E(Y|X = x) = \beta_0^* + (\beta_1{}^*)^T(x - \mu), \tag{2.3}$$

where the scalar intercept β_0^* and slope vector $\beta_1{}^*$ are the solutions

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0 \in \mathcal{R}, \beta_1 \in \mathcal{R}^p}{\operatorname{argmin}} \int \left[\int y \mathrm{d}F_{Y|X}(x, y) - (\beta_0 + \beta_1^T(x - \mu)) \right]^2 \mathrm{d}F_X(x). \tag{2.4}$$

Similar to the Fréchet mean, the goal is to characterize the regression values in (2.3) as minimizers of weighted least squares problems, where the weights depend on predictor values and the squared distances depend on response values. Setting $\mu = E(X)$, $\Sigma = \operatorname{Var}(X)$ and $\sigma_{YX} = E[Y(X - \mu)]$, the normal equations for the right-hand side of (2.4) lead to

$$E(Y) - \beta_0 = 0, \quad \sigma_{YX} - \Sigma\beta_1 = 0,$$

with solutions $\beta_1{}^* = \Sigma^{-1}\sigma_{YX}$ and $\beta_0^* = E(Y)$. Plugging these into (2.3),

$$\begin{aligned} m(x) &= E(Y) + \sigma_{YX}^T\Sigma^{-1}(x - \mu) = \int y \left\{1 + (z - \mu)^T\Sigma^{-1}(x - \mu)\right\} \mathrm{d}F(z, y) \\ &= \int y s(z, x) \mathrm{d}F(z, y), \end{aligned} \tag{2.5}$$

where the weight function s is

$$s(z, x) = 1 + (z - \mu)^T\Sigma^{-1}(x - \mu). \tag{2.6}$$

Because $\int s(z, x)\mathrm{d}F(z, y) = 1$, the last line of (2.5) reveals that the standard linear regression function value $m(x)$ is the solution

$$m(x) = \underset{y \in \mathcal{R}}{\operatorname{argmin}}\, E\left[s(X, x)d_E^2(Y, y)\right], \tag{2.7}$$

where d_E is the standard Euclidean metric. This alternative formulation of the linear regression function provides the key to defining the proposed global Fréchet regression function $m_\oplus$ on an arbitrary metric space (Ω, d), by simply replacing the Euclidean metric d_E, which is the default metric for real valued responses, by a more general metric d that is suitable for responses in Ω. The global Fréchet regression model then becomes

$$m_\oplus(x) := \operatorname*{argmin}_{\omega \in \Omega} M(\omega, x), \quad M(\cdot, x) = E\left[s(X, x) d^2(Y, \cdot)\right]. \tag{2.8}$$

Hence, generalizing multiple linear regression to the case of a metric-valued response is achieved by viewing the regression function as a sequence of weighted Fréchet means, with weights that are derived from those of the corresponding standard linear regression. Although Ω is not a linear space, the weight function s is a sensible choice for a number of reasons. First, any coherent generalization of multiple linear regression to a global model for random object regression should result in a regression function passing through the point $(\mu, \omega_\oplus)$, which holds for the proposed model since $s(\cdot, \mu) \equiv 1$ implies that $m_\oplus(\mu) = \omega_\oplus$. Second, in contrast to local regression in metric spaces, where the weights are given by a nonnegative kernel function, the weights given by s can be negative and do not go to zero away from x, both of which are natural properties of a global regression relationship. Lastly, despite being defined as a minimizer of a weighted Fréchet function, the proposed global Fréchet regression function can be computed analytically in some cases, in addition to the obvious case $\Omega = \mathcal{R}$. As an illustrative example, when Ω is the space of probability distributions on the real line equipped with the Wasserstein metric (see Example 1 and Section 6.2 below) and the random objects Y are distributions from a location-scale family with random location ν and scale σ, the global Fréchet regression model is equivalent to modeling the conditional means of ν and σ as linear functions of the predictor x. In fact, when the location-scale family is the Gaussian family, this space has a curved manifold structure, with properties studied extensively in the literature (e.g. Takatsu, 2011). This provides an example of a curved manifold for which the global Fréchet regression relationship is sensible.

2.3. *Estimation.* Assume that $(X_i, Y_i) \sim F$, $i = 1, \ldots, n$, are independent. We take the standard approach to estimate the minimizer in (2.8) by substituting the empirical distribution $\mathrm{d}F_n$ for $\mathrm{d}F$ in the integral in (2.8). Additionally, the unknown parameters μ and Σ in (2.6) are replaced by their empirical estimates $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and $\hat{\Sigma} = n^{-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$, respectively.

The empirical weights

$$s_{in}(x) := 1 + (X_i - \bar{X})^T \hat{\Sigma}^{-1}(x - \bar{X}) \tag{2.9}$$

then lead to the estimator

$$\hat{m}_\oplus(x) = \operatorname*{argmin}_{\omega \in \Omega} M_n(\omega, x) \tag{2.10}$$

of $m_\oplus(x)$ for $x \in \mathcal{R}^p$, where $M_n(\cdot, x) = n^{-1} \sum_{i=1}^n s_{in}(x) d^2(Y_i, \omega)$.

3. Theory. We first consider the estimation of the regression relation in (2.8) by the corresponding estimator in (2.10) in the case of a totally bounded metric space (Ω, d). Recall the functions

$$M(\omega, x) := E\left[s(X, x) d^2(Y, \omega)\right], \quad M_n(\omega, x) := n^{-1} \sum_{i=1}^n s_{in}(x) d^2(Y_i, \omega).$$

With regard to the objects in (2.8) and (2.10), we require the following assumptions for a fixed $x \in \mathcal{R}^p$.

(P0) The objects $m_\oplus(x)$ and $\hat{m}_\oplus(x)$ exist and are unique, the latter almost surely, and, for any $\varepsilon > 0$, $\inf_{d(\omega, m_\oplus(x)) > \varepsilon} M(\omega, x) > M(m_\oplus(x), x)$.

(P1) For $\delta > 0$ small enough,

$$\int_0^1 \sqrt{1 + \log N(\epsilon, B_\delta(m_\oplus(x)), d)}\, d\epsilon < \infty,$$

where $B_\delta(m_\oplus(x))$ is the δ-ball centered at $m_\oplus(x)$ and $N(\epsilon, \Omega, d)$ is the covering number for $B_\delta(m_\oplus(x))$ using open balls of radius ϵ.

(P2) There exist $\eta > 0$, $C > 0$ and $\beta > 1$, possibly depending on x, such that, whenever $d(m_\oplus(x), \omega) < \eta$, we have $M(\omega, x) - M(m_\oplus(x), x) \geq C d(\omega, m_\oplus(x))^\beta$.

Assumption (P0) is common to establish the consistency of an M-estimator such as $\hat{m}_\oplus(x)$; see Chapter 3.2 in Van der Vaart and Wellner (1996). In particular, it ensures that weak convergence of the empirical process M_n to the population process M in turn implies convergence of their minimizers. Furthermore, existence follows immediately if Ω is compact. The conditions on the covering number in (P1) and curvature in (P2) arise from empirical process theory and control the behavior of $M_n - M$ near the minimum in order to obtain rates of convergence.

We also consider uniform convergence results for predictor values x, requiring stronger versions of the above assumptions. Let $\|\cdot\|_E$ be the Euclidean norm on $\mathcal{R}^p$ and $B > 0$.

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

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

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

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

(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

$$\inf_{\|x\|_E \leq B} \inf_{d(\omega, m_\oplus(x)) < \tau} \{M(\omega, x) - M(m_\oplus(x), x) - D d(\omega, m_\oplus(x))^\alpha\} \geq 0.$$

The following examples of classes of random objects correspond to the applications and simulations that will be discussed in Sections 6, 7 and 8.

EXAMPLE 1. *Take Ω to be the set of probability distributions G on $\mathcal{R}$ such that $\int_{\mathcal{R}} x^2 \mathrm{d}G(x) < \infty$, equipped with the Wasserstein metric d_W. For two such distributions G_1 and G_2, the Wasserstein distance is given by*

$$d_W(G_1, G_2)^2 = \int_0^1 (G_1^{-1}(t) - G_2^{-1}(t))^2\, dt,$$

where G_1^{-1} and G_2^{-1} are the quantile functions corresponding to G_1 and G_2, respectively.

EXAMPLE 2. *Take Ω as the set of correlation matrices of a fixed dimension r, i.e. symmetric, positive semidefinite $r \times r$ matrices with unit diagonal, and equip Ω with the Frobenius metric, d_F.*

EXAMPLE 3. *Let Ω be a (bounded) Riemannian manifold of dimension r and let d be the geodesic distance implied by the Riemannian metric.*

Propositions 1 and 2 in the supplement demonstrate that all of the above assumptions are satisfied for the random objects in Examples 1 and 2, with $\beta = \alpha = 2$ in (P2) and (U2). We note that Example 1 refers to objects in

the Wasserstein space, a complex smooth manifold that is characterized by the Wasserstein geodesics (Takatsu 2011) and thus provides an example of random objects on a manifold, for which we obtain consistent estimation of global and local Fréchet regression as demonstrated below. Example 2 refers to a convex space and, at first glance, it seems straightforward to implement local regression using kernel weights on such spaces. This is however not the case; a major difficulty is that global and local regression assign negative weights near the boundaries, where the boundary is a very substantial part of the domain especially in the global regression case.

For Example 3, Proposition 3 shows that (P1) and (U1) hold automatically and, if (P0) (respectively (U0)) holds, then (P2) (respectively (U2)) is equivalent to the Hessian on the tangent space at $m_\oplus(x)$ being positive definite at 0, and in this case we may take $\alpha = \beta = 2$. Thus, for manifolds, local curvatures do not influence the convergence rates below. Uniqueness of Fréchet means for manifolds is challenging in general, but can be guaranteed under certain circumstances, for example restricting the support of the underlying distribution F_Y (Afsari 2011). Alternatively, one can consider Fréchet mean sets (Ziezold 1977); see also the last paragraph in Section 9.

The following two results demonstrate the consistency of our proposed estimators and also provide rates of convergence. All proofs can be found in the Supplement.

THEOREM 1. *Suppose (P0) holds and Ω is bounded. Then, for any fixed $x \in \mathcal{R}$, $d(\hat{m}_\oplus(x), m_\oplus(x)) = o_p(1)$. For $B > 0$, if (U0) holds then $\sup_{\|x\|_E \le B} d(\hat{m}_\oplus(x), m_\oplus(x)) = o_p(1)$.*

THEOREM 2. *Suppose that, for a fixed $x \in \mathcal{R}^p$, (P0)–(P2) hold. Then*

$$d(\hat{m}_\oplus(x), m_\oplus(x)) = O_p\left(n^{-\frac{1}{2(\beta-1)}}\right).$$

Furthermore, for a given $B > 0$, if (U0)–(U2) hold,

$$\sup_{\|x\|_E \le B} d(\hat{m}_\oplus(x), m_\oplus(x)) = O_p\left(n^{-\frac{1}{2(\alpha-1)}}\right).$$

We remark that, if $\beta = 2$ in (P2) and $\alpha = 2$ in (U2), which is the case when Ω is a Euclidean space, this result gives the usual parametric convergence rate of $n^{-1/2}$. In general, the rate of convergence is determined by the local geometry near the minimum as quantified in (P2) and (U2). The proof of the pointwise result follows along the lines of Theorem 3.2.5 in Van der Vaart and Wellner (1996) which deals with M-estimators, where some additional

considerations are needed to deal with the necessary estimation of the mean and covariance of X. The uniform result is more difficult, as an uncountable number of M-estimators are considered simultaneously and no parametric form of the regression function is available. When Ω has a smooth structure, e.g., the Wasserstein space in Example 1 or a smooth Riemannian manifold, one can conceivably also obtain a limiting distribution. We demonstrate this for the case where Ω is a Hilbert space in Section 5.

4. Local Fréchet Regression. As the success of nonparametric regression methods over the last decades has shown, there is often the need for local rather than global fitting of regression functions. Local regression is more flexible but on the other hand requires choosing a tuning parameter that balances bias and variance. As far as we know, to date, local estimation of (2.2) for responses in general metric spaces has been exclusively done with the Nadaraya-Watson estimator (Davis et al. 2007; Hein 2009; Steinke and Hein 2009; Steinke, Hein and Schölkopf 2010)

$$\hat{m}_{\oplus}^{\mathrm{NW}}(x) = \operatorname*{argmin}_{\omega\in\Omega} \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)d^2(Y_i, \omega), \tag{4.1}$$

where K is a smoothing kernel that corresponds to a probability density and h is a bandwidth, with $K_h(\cdot) = h^{-1}K(\cdot/h)$. In this section, the proposed Fréchet regression analysis is extended from the global setting, as described in the previous sections, to a local version. The idea is to adopt the concepts of local linear regression, which has been established for real-valued responses, and then to extend them to the case where responses are random objects, in analogy to the developments in Section 2.2 for global Fréchet regression. Thus, we develop a novel local version of smoothing in general metric spaces which goes beyond the Nadaraya-Watson smoother (4.1). As is the case for Euclidean data, this local Fréchet regression proves to be superior to Nadaraya-Watson smoothing, especially near the boundaries, as demonstrated in the experiments in Sections 6.3 and 8. Moreover, our analysis of these estimators separates bias and stochastic variation of the corresponding estimators.

For ease of representation, we consider here the case of a scalar predictor $X \in \mathcal{R}^p$, where $p = 1$; the local method can also be developed for any p with $p > 1$. The target is again (2.2), where we make no structural assumptions on $m_{\oplus}$. Consider the preliminary case $\Omega = \mathcal{R}$, and again write $m = m_{\oplus}$. In this case, the local linear estimate (Fan and Gijbels 1996) of $m(x)$ is

$\hat{l}(x) = \hat{\beta}_0$, where

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} K_h(X_i - x)(Y_i - \beta_0 - \beta_1(X_i - x))^2.$$

In this sense, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can be viewed as M-estimators of

(4.2)
$$(\beta_0^*, {\beta_1}^*) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \int K_h(z-x) \left[\int y \mathrm{d}F_{Y|X}(z,y) - (\beta_0 + \beta_1(z-x)) \right]^2 \mathrm{d}F_X(z).$$

Defining $\mu_j = E\left[K_h(X-x)(X-x)^j\right]$, $r_j = E\left[K_h(X-x)(X-x)^j Y\right]$ and $\sigma_0^2 = \mu_0\mu_2 - \mu_1^2$, the solutions to (4.2) are

$$\beta_0^* = \sigma_0^{-2}(\mu_2 r_0 - \mu_1 r_1), \quad {\beta_1}^* = \sigma_0^{-2}(\mu_0 r_1 - \mu_1 r_0).$$

This means that $\hat{l}(x) = \hat{\beta}_0$ can be viewed as an estimator of the intermediate target

(4.3)
$$\begin{aligned} \tilde{l}(x) = \beta_0^* &= \frac{\mu_2 r_0 - \mu_1 r_1}{\sigma_0^2} = \frac{1}{\sigma_0^2} \int y K_h(z-x) \left[\mu_2 - \mu_1(z-x)\right] \mathrm{d}F(z,y) \\ &= E[s(X,x,h)Y] \end{aligned}$$

for the weight function

$$s(z,x,h) = \frac{1}{\sigma_0^2} \left\{ K_h(z-x) \left[\mu_2 - \mu_1(z-x)\right] \right\}.$$

Observing that $\int s(z,x,h)\mathrm{d}F(z,y) \equiv 1$, it follows that $\tilde{l}(x)$ in (4.3) corresponds to a localized Fréchet mean,

$$\tilde{l}(x) = \underset{y \in \mathcal{R}}{\operatorname{argmin}} E\left[s(X,x,h)(Y-y)^2\right]. \tag{4.4}$$

The minimizer $\tilde{l}(x)$ in (4.4) can be viewed as a smoothed version of the true regression function, with the bias $m(x) - \tilde{l}(x) = o(1)$ as $h \to 0$. Under mild assumptions on the kernel and distribution F, this bias is $O(h^2)$, which follows from a Taylor expansion argument.

Now we are in a position to define the local regression concept for random objects $Y \in \Omega$, in analogy to the global Fréchet regression. Specifically, (4.4) can be generalized by defining $\tilde{L}_n(\omega) = E\left[s(X,x,h)d^2(Y,\omega)\right]$, where

the dependency on n is through the bandwidth sequence $h = h_n$, and then setting

$$\tilde{l}_{\oplus}(x) = \underset{\omega \in \Omega}{\operatorname{argmin}}\, \tilde{L}_n(\omega).$$

In contrast to Euclidean spaces or Riemannian manifolds (Yuan et al. 2012), no version of a Taylor expansion argument is available on general metric spaces Ω. So one can ask why this weighted Fréchet mean provides a good approximation to the conditional mean in (2.2). It turns out that this is due to the fact (shown in the proof of Theorem 3 below) that

$$\left[\int s(z, x, h) \mathrm{d}F_{X|Y}(z, y)\right] \mathrm{d}F_Y(y) = \mathrm{d}F_{Y|X}(x, y) + O(h^2),$$

so that minimizing $\tilde{L}_n$ is approximately the same as minimizing the conditional Fréchet function $M_{\oplus}(\cdot, x)$.

The target $\tilde{l}_{\oplus}(x)$ can be estimated by using preliminary estimates $\hat{\mu}_j = n^{-1} \sum_{i=1}^n K_h(X_i - x)(X_i - x)^j$, $\hat{\sigma}_0^2 = \hat{\mu}_0 \hat{\mu}_2 - \hat{\mu}_1^2$, and the empirical weights

$$s_{in}(x, h) = \frac{1}{\hat{\sigma}_0^2} K_h(X_i - x) \left[\hat{\mu}_2 - \hat{\mu}_1 (X_i - x)\right].$$

Then, setting $\hat{L}_n(\omega) = n^{-1} \sum_{i=1}^n s_{in}(x, h) d^2(Y_i, \omega)$, the local Fréchet regression estimate is

$$\hat{l}_{\oplus}(x) = \underset{\omega \in \Omega}{\operatorname{argmin}}\, \hat{L}_n(\omega). \tag{4.5}$$

While this local estimation technique is developed here for general metric space data, it is of interest to compare it to other local estimators that have been previously considered for spaces with additional structure, specifically the intrinsic local polynomial (ILPR) estimator for manifold data proposed in Yuan et al. (2012), where covariance matrices as objects are regressed against scalar predictors. Whereas the ILPR estimator requires various technical steps involving exponential, logarithmic and parallel transport maps on the manifold, one advantage of the methodology proposed here is its simplicity, only requiring distances between data objects. In terms of computation on manifolds, the current method also enjoys the distinct advantage of requiring optimization only for a single object, unlike the ILPR for which one has to fit both intercept and derivative terms. It is of course also much more general, providing consistent estimators in unstructured metric spaces. Furthermore, the function to be minimized is merely a weighted least squares problem, potentially with negative weights. Thus, any metric space

for which a Nadaraya-Watson smoother (Hein 2009) is computationally feasible, or any manifold for which the ILPR can be computed, is also feasible for both local and global Fréchet regression. In the manifold case, expressions for the Riemannian gradient and Hessian are available for a variety of complex manifolds (Ferreira et al. 2013), which can be used for Newton-type algorithms, possibly in conjunction with stochastic optimization techniques, such as the annealing algorithm of Yuan et al. (2012).

For a concrete comparison of local Fréchet regression with the ILPR, take Ω to be the space of covariance matrices with d being the Log-Euclidean metric, that is, $d(\omega_1, \omega_2) = d_F(\operatorname{Log}\omega_1, \operatorname{Log}\omega_2)$, where d_F is the Frobenius metric and Log is the inverse of the matrix exponential Exp (Arsigny et al. 2007). In this case, both the ILPR and local Fréchet regression estimates can be computed analytically. For a sample (X_i, Y_i), with Y_i a positive definite covariance matrix, both methods yield the estimate

$$\hat{m}_\oplus(x) = \operatorname{Exp}\left(\frac{\hat{\mu}_2\hat{\mathrm{r}}_0 - \hat{\mu}_2\hat{\mathrm{r}}_1}{\hat{\sigma}_0^2}\right),$$

where $\hat{r}_j = n^{-1}\sum_{i=1}^n K_h(X_i - x)(X_i - x)^j \operatorname{Log}(Y_i)$. That these two methods coincide is not altogether surprising due to the metric being the Euclidean metric on transformed matrices. However, it shows that in this situation local Fréchet regression gives a sensible and intuitive estimate which coincides with the previously established manifold-based estimator.

Returning to theory, in order to obtain the rate of convergence for the quantity $d(m_\oplus(x), \hat{l}_\oplus(x))$, we need to quantify the convergence of the bias term $d(m_\oplus(x), \tilde{l}_\oplus(x))$ and the stochastic term $d(\tilde{l}_\oplus(x), \hat{l}_\oplus(x))$. This requires the assumptions below. Recall that $M_\oplus(\omega, x) = E(d^2(Y, \omega)|X = x)$. For simplicity, we assume that the marginal density f of X, within the joint distribution F, has unbounded support, and consider points $x \in \mathcal{R}$ for which $f(x) > 0$. We need the following assumptions.

(K0) The kernel K is a probability density function, symmetric around zero. Furthermore, defining $K_{kj} = \int_{\mathcal{R}} K^k(u)u^j\, du$, $|K_{14}|$ and $|K_{26}|$ are both finite.

(L0) The object $m_\oplus(x)$ exists and is unique. For all n, $\tilde{l}_\oplus(x)$ and $\hat{l}_\oplus(x)$ exist and are unique, the latter almost surely. Additionally, for any $\varepsilon > 0$,

$$\inf_{d(\omega, m_\oplus(x)) > \varepsilon} \{M_\oplus(\omega, x) - M_\oplus(m_\oplus(x), x)\} > 0,$$

$$\liminf_n \inf_{d(\omega, \tilde{l}_\oplus(x)) > \varepsilon} \left\{\tilde{L}_n(\omega) - \tilde{L}_n(\tilde{l}_\oplus(x))\right\} > 0.$$

(L1) The marginal density f of X, as well as the conditional densities g_y of $X|Y = y$, exist and are twice continuously differentiable, the latter for

all $y \in \Omega$, and $\sup_{x,y} |g_y''(x)| < \infty$. Additionally, for any open $U \subset \Omega$, $\int_U \mathrm{d}F_{Y|X}(x,y)$ is continuous as a function of x.

(L2) There exists $\eta_1 > 0$, $C_1 > 0$ and $\beta_1 > 1$ such that

$$M_\oplus(\omega, x) - M_\oplus(m_\oplus(x), x) \geq C_1 d(\omega, m_\oplus(x))^{\beta_1},$$

provided $d(\omega, m_\oplus(x)) < \eta_1$.

(L3) There exists $\eta_2 > 0$, $C_2 > 0$ and $\beta_2 > 1$ such that

$$\liminf_n \left[\tilde{L}_n(\omega) - \tilde{L}_n(\tilde{l}_\oplus(x))\right] \geq C_2 d(\omega, \tilde{l}_\oplus(x))^{\beta_2},$$

provided $d(\omega, \tilde{l}_\oplus(x)) < \eta_2$.

Assumptions (K0) and (L1) are common in local regression estimation and imply that the smoothed marginal distribution

$$\left(\int s(z, x, h) \mathrm{d}F_{X|Y}(z|y)\right) \mathrm{d}F_Y(y)$$

converges to $\mathrm{d}F_{Y|X}(x,y)$ as $h \to 0$, while (L2) and (L3) provide the rate for the bias and stochastic terms, respectively. While (L1) is a distributional assumption, (L2) and (L3) can be shown to hold for Examples 1–3 in Section 3, using arguments similar to those in Propositions 1–3 in the supplement. In these cases, it is easy to verify that $C_j = 1$, $\beta_j = 2$ and η_j arbitrary, $j = 1, 2$, are admissible in (L2) and (L3). We now state our main results for local Fréchet regression, where the first result is for the bias, the second for the stochastic deviation and the corollary combines these results to obtain an overall rate of convergence.

THEOREM 3. *If (K0), (L0), (L1), (L2) and (P1) hold, then*

$$d(m_\oplus(x), \tilde{l}_\oplus(x)) = O(h^{2/(\beta_1 - 1)})$$

as $h = h_n \to 0$.

THEOREM 4. *If (K0), (L0), (L3) and (P1) hold, and if* $h \to 0$ *and* $nh \to \infty$, *then*

$$d(\tilde{l}_\oplus(x), \hat{l}_\oplus(x)) = O_p\left[(nh)^{-\frac{1}{2(\beta_2 - 1)}}\right].$$

COROLLARY 1. *Under the assumptions of Theorem 3 and Theorem 4, among bandwidth sequences* $h = n^{-\gamma}$, *the optimal sequence is obtained for* $\gamma^* = (\beta_1 - 1)/(4\beta_2 + \beta_1 - 5)$ *and yields the rate*

$$d(m_\oplus(x), \hat{l}_\oplus(x)) = O_p\left(n^{-2/(\beta_1 + 4\beta_2 - 5)}\right).$$

We note that for $\beta_1 = \beta_2 = 2$, one obtains the result

$$d(m_\oplus(x), \hat{l}_\oplus(x)) = O_p\left[(nh)^{-\frac{1}{2}} + h^2\right]$$

that is familiar for local regression with real valued responses, and with $\gamma^* = 1/5$ leads to the rate $d(m_\oplus(x), \hat{l}_\oplus(x)) = O_p(n^{-2/5})$. While the above results are pointwise, we remark that the same rate will hold uniformly in x over compact intervals by suitably strengthening assumptions (L0), (L2) and (L3).

5. Limiting Distributions when Ω is a separable Hilbert space. A case of particular interest arises when the random objects are functions that are assumed to be (almost surely) square-integrable, e.g., $\Omega = L^2[0,1]$) (Faraway 1997). Going beyond functional data as responses, we more generally assume that Ω is a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\|\cdot\|_\Omega$. As before, let F be a distribution on $\mathbb{R} \times \Omega$ with $(X, Y) \sim F$. As this setting enables linear operations, the minimizing objects in (2.8) and (2.10) can be given explicitly under mild assumptions on the moments of F. Unsurprisingly, for the case of functional data, the minimizer of (2.10) corresponds to the estimator given in Faraway (1997). Our developments in the following are for global Fréchet regression, but using essentially the same arguments can be extended to local Fréchet regression, by utilizing the tools developed in Section 4.

We will use the following notation. For $q > 1$, let Ω^q be the q-fold Cartesian product of Ω, with inner product $\langle \alpha, \alpha' \rangle_q = \sum_{l=1}^q \langle \alpha_l, \alpha'_l \rangle$ for $\alpha, \alpha' \in \Omega^q$, so that Ω^q is also a Hilbert space. For a $p \times p$ matrix A, $x \in \mathcal{R}^p$, $\omega \in \Omega$ and $\alpha \in \Omega^p$, we define $A\alpha \in \Omega^p$ with elements $(A\alpha)_l = \sum_{m=1}^p A_{lm}\alpha_m$, $\alpha^T x = \sum_{l=1}^p x_l \alpha_l \in \Omega$ and $x\omega \in \Omega^p$ with elements $(x\omega)_l = x_l \omega$.

THEOREM 5. *A. Let $(X, Y) \sim F$ and suppose that $E\|Y\|_\Omega^2 < \infty$. Then there exist unique elements $\gamma_0 \in \Omega$ and $\gamma_1 \in \Omega^p$ which satisfy, for all $\omega \in \Omega$ and $\alpha \in \Omega^p$,*

$$E\langle Y, \omega \rangle = \langle \gamma_0, \omega \rangle \quad \text{and} \quad E\langle (X - \mu)Y, \alpha \rangle_p = \langle \gamma_1, \alpha \rangle_p.$$

With $\Sigma = \mathrm{Var}(X)$ and defining $\beta_1 := \Sigma^{-1}\gamma_1$ and $\beta_0 = \gamma_0$, the solution to (2.8) is

$$m_\oplus(x) = \beta_0 + \beta_1^T(x - \mu). \tag{5.1}$$

B. Define estimators $\hat{\gamma}_0 = \bar{Y} = n^{-1}\sum_{i=1}^n Y_i$, $\hat{\gamma}_1 = n^{-1}\sum_{i=1}^n (X_i - \bar{X})Y_i$, $\hat{\beta}_1 = \hat{\Sigma}^{-1}\hat{\gamma}_1$ and $\hat{\beta}_0 = \hat{\gamma}_0$. The solution of (2.10) is given by

$$\hat{m}_\oplus(x) = \hat{\beta}_0 + \hat{\beta}_1^T(x - \bar{X}). \tag{5.2}$$

Results (5.1) and (5.2) demonstrate that explicit solutions of the minimization problems that define the global Fréchet regression are available for the case of responses that are random objects in Hilbert space. Moreover, in this situation one can also obtain limiting distributions, as follows.

THEOREM 6. *Set* $\beta = (\beta_0, \beta_1^T)^T$ *and* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1^T)^T$. *Under the assumptions of Theorem 5,*

$$\sqrt{n}(\hat{\beta} - \beta) \rightsquigarrow \mathcal{G},$$

where '$\rightsquigarrow$*' denotes weak convergence and* $\mathcal{G}$ *is a zero mean Gaussian process on* Ω^{p+1}. *The covariance structure of* $\mathcal{G}$ *is defined by projection covariances* $\mathrm{Cov}(\langle \mathcal{G}, \alpha \rangle_{p+1}) = l_\alpha^T C_\alpha l_\alpha$, *where* $\alpha \in \Omega^{p+1}$, C_α *is the covariance matrix of the vector defined in (S.14) in the supplement and* l_α *can be constructed using the expressions in (S.15) in the supplement.*

We next consider weak convergence of the process

$$\mathcal{M}_n(x) = \sqrt{n}(\hat{m}_\oplus(x) - m_\oplus(x))$$

as x varies in $\mathcal{R}^p$. For any $U \subset \mathcal{R}^p$, define the function space

$$l_\Omega^\infty(U) = \{g : U \to \Omega : \sup_{x \in U} \|g(x)\|_\Omega < \infty\}$$

with norm $\|g\|_U = \sup_{x \in U} \|g\|_\Omega$.

COROLLARY 2. *Let* $B > 0$ *be arbitrary, and define* $V_B = \{x \in \mathcal{R}^p : \|x\|_E \leq B\}$. *Under the assumptions of Theorem 5,*

$$\sup_{x \in V_B} \|\hat{m}_\oplus(x) - m_\oplus(x)\|_\Omega = O_p(n^{-1/2}).$$

Additionally, there is a zero-mean Gaussian process $\mathcal{M}$ *on* V_B *such that*

$$\mathcal{M}_n \rightsquigarrow \mathcal{M} \quad \text{in} \quad l_\Omega^\infty(V_B),$$

where $\mathcal{M}_n$ *is restricted to* V_B.

These results show that one can take advantage of the additional structure that is available in the case of Hilbertian objects to obtain limit distributions of the estimates. Limit distributions are not available for general object spaces due to the lack of a linear structure. Generally, even for the simpler case of Fréchet means, limit results cannot be directly obtained, except in special cases. For example, for random objects that fall on manifolds satisfying certain regularity conditions, local linear approximations sometimes make it possible to derive limit theorems (Barden, Le and Owen 2013).

6. Fréchet Regression for Probability Distributions with the Wasserstein metric.

6.1. *Computational Details.* Here, the space Ω is the set of distribution functions equipped with the Wasserstein metric, as outlined in Example 1 in Section 3. To implement the minimization required by (2.10) using a sample (X_i, Y_i), $i = 1, \ldots, n$, of covariates and distributions, first define $Q(\omega)$ to be the quantile function corresponding to ω, for any $\omega \in \Omega$, and let Q^{-1} be the inverse map, mapping quantile functions to their associated distribution function. Set $\hat{g}_x = n^{-1} \sum_{i=1}^n s_{in}(x) Q(Y_i)$, where the weights $s_{in}(x)$ are given in (2.9). Note that $\hat{g}_x \in L^2[0,1]$, and let d_{L^2} be the standard L^2 metric on this space. The global Fréchet regression estimator is

$$\hat{m}_\oplus(x) = \operatorname*{argmin}_{\omega \in \Omega} d^2_{L^2}(\hat{g}_x, Q(\omega)) = Q^{-1}\left(\operatorname*{argmin}_{q \in Q(\Omega)} d^2_{L^2}(\hat{g}_x, q) \right),$$

where we refer to the proof of Proposition 1 in the supplement for details.

Now, let u_j, $j = 1, \ldots, M$, be an equispaced grid on $[0,1]$ and let $g_j = \hat{g}_x(u_j)$. Then compute

$$q^* = \operatorname*{argmin}_{q \in \mathcal{R}^M} \|g - q\|^2_E,$$

subject to the constraint $q_1 \leq \cdots \leq q_M$. This optimization problem is a quadratic program and can be solved using a variety of techniques. The solution q^* represents a discretized version of the approximation of the quantile function $Q(\hat{m}_\oplus(x))$. Similar arguments hold for the computation of the local Fréchet estimator.

6.2. *Simulation.* To assess the performance of the global Fréchet regression estimator in (2.10), it is first necessary to determine a generative model that produces suitably simulated data. The space of distributions with the Wasserstein metric provides an ideal setting for this. The responses Y are distributions with quantile functions $Q(Y)$ and the predictors are random variables $X \in \mathcal{R}$. For notational simplicity, the quantile function corresponding to Y will also be denoted as Y. The regression function is

$$m_\oplus(x)(\cdot) = E(Y(\cdot)|X = x) = \mu_0 + \beta x + (\sigma_0 + \gamma x)\Phi^{-1}(\cdot),$$

where Φ is the standard normal distribution function, μ_0, $\beta \in \mathcal{R}$ and σ_0 and γ satisfy $\sigma_0 + \gamma x > 0$ for all x in the support of F_X. This corresponds to the response distributions being, on average, a normal distribution with parameters that depend linearly on x.

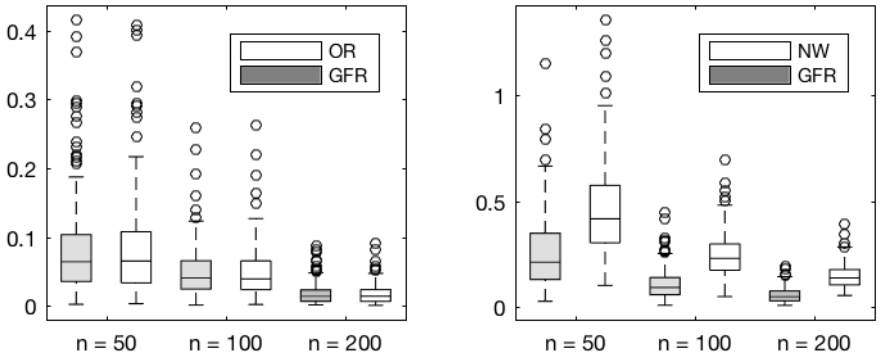
The random response Y is generated conditional on X by adding noise to the quantile functions, which we will demonstrate in two settings. In the first, the distribution parameters $\mu|X \sim \mathcal{N}(\mu_0 + \beta X, v_1)$ and $\sigma|X \sim \mathrm{Gam}((\sigma_0 + \gamma X)^2/v_2, v_2/(\sigma_0 + \gamma X))$ are independently sampled, and the corresponding distribution is $Y = \mu + \sigma\Phi^{-1}$. In the second setting, after sampling the distribution parameters as in the previous setting, the resulting distribution is "transported" in Wasserstein space following a simplified version of the algorithm outlined in Section 8.1 of Panaretos and Zemel (2016). Specifically, random transport maps T (increasing diffeomorphisms of the real line) are generated by sampling uniformly from the collection of transport maps $T_k(x) = x - \sin(kx)/|k|$, for $k \in \{-l, \ldots, l\} \setminus \{0\}$, with $Y = T \circ (\mu + \sigma\Phi^{-1})$. This second setting is significantly more complex, as the observed distributions are no longer Gaussian.

Random samples of pairs (X_i, Y_i), $i = 1, \ldots, n$ were generated by sampling $X_i \sim \mathcal{U}(-1, 1)$, setting $\mu_0 = 0$, $\sigma_0 = 3$, $\beta = 3$ and $\gamma = 0.5$, and following the above procedure for the two simulation settings. In the first setting, the parameter variances were set at $v_1 = 0.25$ and $v_2 = 1$. In the second, the values were $v_1 = 1$ and $v_2 = 2$, with $l = 2$ used for generating the transport maps. In each setting, 200 runs were executed for three sample sizes $n = 50, 100, 200$. For the r-th simulation of a particular sample size, with $\hat{m}^r_\oplus(x)$ denoting the fitted distribution function, the quality of the estimation was measured quantitatively by the integrated squared errors

$$\mathrm{ISE}_r = \int_{-1}^{1} d_W^2(\hat{m}^r_\oplus(x), m_\oplus(x))\, \mathrm{d}x.$$

In the first simulation setting, we verify that global Fréchet regression is performing as expected by comparing to the best-case scenario where one knows the finite-dimensional generating model. That is, we compute the mean μ_i and standard deviation of σ_i of the distribution Y_i and regress them linearly against X_i, while restricting the estimates of σ_0 and γ such that the regression line is positive on $[-1, 1]$. Thus, we can compare this "oracle" linear regression with global Fréchet regression by computing its integrated squared error for each simulation run. These errors are shown for both methods in boxplots in Figure 1a. It is clear that global Fréchet regression performs just as well as the oracle procedure. Sign-rank tests were performed to test the hypothesis of no difference between the methods for each sample size, with the smallest of the three p-values being 0.51.

In the second simulation setting, the random transportation renders the oracle linear regression technique above inadmissable, since the standard deviation of the transported distribution no longer has a linear relationship



(a) (b)

Fig 1: Boxplots of integrated squared errors for 200 simulation runs and three sample sizes n. The left panel compares global Fréchet regression (GFR) with the oracle linear regression (OR), while the right shows results for global Fréchet regression and the Nadaraya-Watson smoother (NW).

with X. However, the global Fréchet regression model still holds true. Figure 1b shows the decreasing integrated squared errors for increasing sample sizes, demonstrating the validity and utility of global Fréchet regression for this complex regression setting. Furthermore, at the suggestion of a referee, we compared our results with the Nadaraya-Watson estimator in (4.1), where the bandwidth was chosen in the interval $[0.2, 0.7]$ so as to minimize the average ISE over all simulations. This resulted in bandwidth choices 0.5, 0.45, and 0.35 for $n = 50, 100$, and 200, respectively. The corresponding ISE values in Figure 1b demonstrate that this approach is inferior to the global Fréchet fits, which is expected if the global model holds, analogous to the situation in the Euclidean case.

6.3. *Application to Mortality Profiles.* Many studies and analyses have been motivated by a desire to understand human longevity. Of particular interest is the evolution of the distributions of age-at-death over calendar time. The Human Mortality Database provides such data in the form of yearly lifetables, differentiated by country. Currently, this database includes yearly mortality and population data for 37 countries that are available at <www.mortality.org>. For a given country and calendar year, the probability distribution for mortality can be represented by its density. A first step is to estimate this density from the data in the lifetables for a specified country. Consider a country for which lifetables are available for

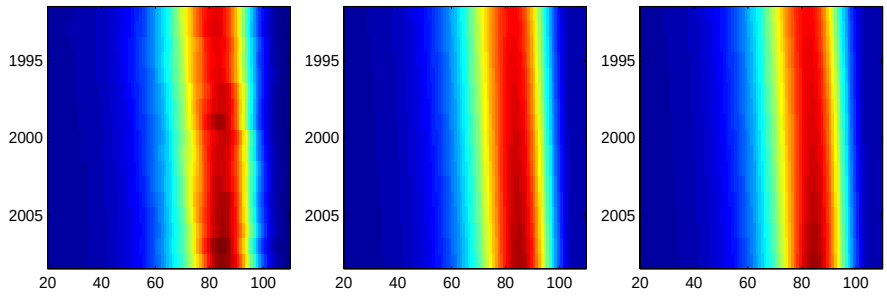


Fig 2: (a) Yearly mortality densities for Chile for the years 1992–2008; (b) Global Fréchet regression fits of yearly mortality densities using $X_i = t_i$; (c) Global Fréchet regression fits using $\alpha_i = (t_i, t_i^2)^T$.

the years t_i, $i = 1, \ldots, n$. For integer-valued ages j, $j = 0, \ldots, 110$, the lifetable provides the size of the population m_j which is at least j years old, normalized so that $m_0 = 100000$. These values can be used to construct a histogram for age-at-death, which in turn can be smoothed using a local linear smoother to obtain an estimate of the density. This smoothing step was performed in `Matlab` using the `hades` package, available at <http://www.stat.ucdavis.edu/hades/>. Each density was estimated for ages in the interval $[20, 110]$, with the value 2 as a common smoothing bandwidth.

As an initial example, we consider the data for Chile, which has mortality data available for the years 1992–2008. Using the procedure outlined above, mortality density estimates Y_i were obtained for the years $t_i = 1991 + i$, $i = 1, \ldots, 17$. These estimates are shown as a heat map in Figure 2, linearly interpolating between years for continuity. The variation from year to year is marked by a steady increase in both the location and height of the peak in mortality. The global Fréchet regression fits using calendar year as predictor for linear ($X_i = t_i$) and quadratic ($X_i = (t_i, t_i^2)^T$) models are shown in Figures 2b and 2c, respectively. Similar to the least squares regression plane, these fits provide a smooth visualization of the evolution of mortality and remove the noise that is visible in the raw density data. There seems to be little gain in fitting a quadratic model, as the global Fréchet regression fits with linear and quadratic predictors are very similar. Leave-one-out prediction errors were 0.088 for the linear fit and 0.0972 for the quadratic fit, indicating that the simpler linear model is indeed preferable.

Next, we consider the data for Luxembourg, with mortality lifetable data

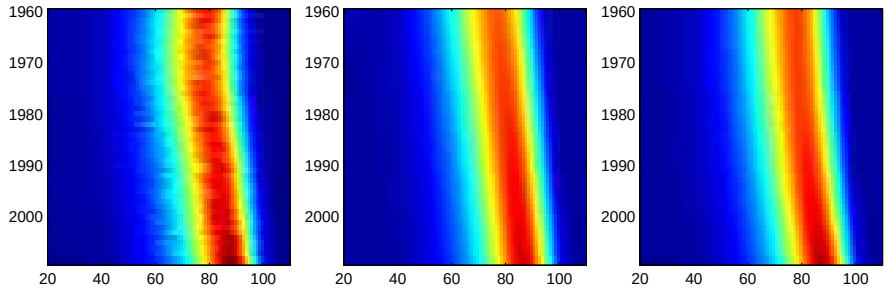
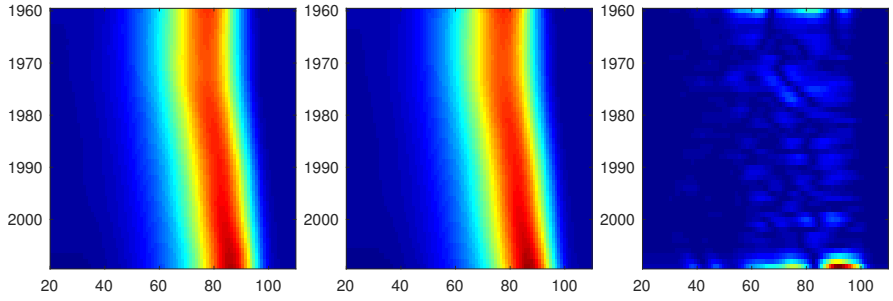


Fig 3: (a) Yearly mortality densities for Luxembourg for the years 1960–2009 (b),(c) Global Fréchet regression fits of yearly mortality densities using $X_i = t_i$ and $X_i = (t_i, t_i^2)^T$, respectively.

ranging from 1960–2009. The density estimates for these years are shown in Figure 3a. We find a slightly more complicated evolution of mortality for Luxembourg compared to Chile. For example, the mode of the density does not steadily increase over the years; rather, the mode seems to carve out a curved path. Figure 3b and 3c show the global Fréchet regression fits for the linear and quadratic global Fréchet model, with $X_i = t_i$ for the linear and $X_i = (t_i, t_i^2)^T$ for the quadratic model. The quadratic fit is better at capturing the shape of the peak dynamics observed in the raw sample of densities. The adjusted Fréchet R^2 values (for details on these extensions of the coefficient of determination see Section 6.4) are 0.971 and 0.975 for the linear and quadratic models, respectively. Average leave-one-out prediction errors were 0.56 for the linear and 0.27 for the quadratic model.

While the quadratic model seems to be indeed better for both fitting and prediction than the linear model, it still does not capture some aspects of the mortality distributions for Luxembourg, particularly between 1970 and 1980. Therefore, local fitting methods will likely prove superior. Figure 4 shows the Nadaraya-Watson kernel regression (4.1) and the local Fréchet (4.5) fits, using bandwidths $h = 5$ and $h = 7$, respectively. These bandwidths were chosen by minimizing the average leave-one-out prediction error over a grid, with minimum values of 0.196 and 0.168, respectively, for Nadaraya-Watson and local Fréchet fits. This represents a 14% improvement in prediction using the local Fréchet fit as compared to the Nadaraya-Watson at the best tuning parameter choices. From the plot of the absolute differences between these estimates in Figure 4c, the superiority of local Fréchet regression for the most part can be attributed to its improved performance near the boundaries.



(a) (b) (c)

Fig 4: (a), (b) Nadaraya-Watson and local Fréchet estimates for Luxembourg (c) Absolute difference between local fits.

Specifically, the Nadaraya-Watson estimator appears to underestimate the mode of the mortality distribution in the years preceding 2009.

6.4. *Inference and Model Selection.* Many of the standard inferential tools that are available for ordinary linear regression depend on the algebraic structure of $\mathcal{R}$, and thus are not directly extendable to Fréchet regression for metric-valued data. However, one tool which does generalize is the coefficient of determination, R^2. Recall that in multiple linear regression modeling with real valued responses, R^2 is usually interpreted as the fraction of variance of the response which is explained by a linear relationship with the predictor variables, i.e.

$$R^2 = 1 - \frac{\mathrm{Var}(Y - \beta_0^* - (\beta_1{}^*)^T(X - \mu))}{\mathrm{Var}(Y)}.$$

Using the generalized notions of mean and variance in (2.1), we define a corresponding Fréchet R^2 coefficient of determination as

$$R^2_\oplus := 1 - \frac{E\left[d^2(Y, m_\oplus(X))\right]}{V_\oplus}.$$

Given a random sample (X_i, Y_i), $i = 1, \ldots, n$, $R^2_\oplus$ can be estimated by

$$\hat{R}^2_\oplus = 1 - \frac{\sum_{i=1}^n d^2(Y_i, \hat{m}_\oplus(X_i))}{\sum_{i=1}^n d^2(Y_i, \hat{\omega}_\oplus)},$$

where

$$\hat{\omega}_\oplus = \underset{\omega \in \Omega}{\mathrm{argmin}}\, n^{-1} \sum_{i=1}^n d^2(Y_i, \omega)$$

is the sample Fréchet mean. The values $R^2_\oplus$ has similar interpretations as the ordinary coefficient of determination R^2 and likely is also useful for inference and model selection.

In the setting of global Fréchet regression, the null hypothesis of no effect is equivalent to testing $\mathcal{H}_0 : R^2_\oplus = 0$, for which the estimate $\hat{R}^2_\oplus$ can be used as a test statistic. In order to obtain a p-value, a permutation test can be performed (Lehmann and D'Abrera 2006; Higgins 2004; Bradley 1968). First, the values X_i, $i = 1, \ldots, n$, are permuted to form a new sample $\tilde{X}_i$, $i = 1, \ldots, n$. For each new predictor sample, a global Fréchet regression is fitted, using the pairs $(\tilde{X}_i, Y_i)$, and the value $\hat{R}^2_\oplus$ is computed for each of these regression fits. By performing a large number of such permutations and fits, one then obtains an empirical approximation of the null distribution of the test statistic and a p-value by calculating the quantile of the actually observed $\hat{R}^2_\oplus$ within this null distribution.

Another potential application of the coefficient $\hat{R}^2_\oplus$ is model selection, where one can mimic the use of the adjusted R^2 in linear regression by fitting Fréchet regression models that use various subsets of the predictor variables. For a fitted submodel $\mathcal{M}$ using $q \leq p$ predictor values, the adjusted Fréchet R^2 is then

$$\hat{R}^2_{\oplus,\mathrm{adj}}(\mathcal{M}) = \hat{R}^2_\oplus - (1 - \hat{R}^2_\oplus)\frac{q}{n-q-1}.$$

Let $\mathcal{C}_q$ be the class of submodels using q predictors, $1 \leq q \leq p$. Computing

$$q^* = \operatorname*{argmax}_{1 \leq q \leq p} \max_{\mathcal{M} \in \mathcal{C}_q} \hat{R}^2_{\oplus,\mathrm{adj}}(\mathcal{M})$$

the final model can then be taken as $\mathcal{M}^* = \operatorname{argmax}_{\mathcal{M} \in \mathcal{C}_{q^*}} \hat{R}^2_{\oplus,\mathrm{adj}}(\mathcal{M})$. Another alternative for model selection is to minimize prediction error, which can be estimated by k-fold cross validation.

7. Fréchet Regression for Correlation Matrices as Random Objects.

7.1. *Computational Details.* Here we consider a space of random objects Ω which consists of correlation matrices, i.e., the space of square $r \times r$ symmetric positive semidefinite matrices with unit diagonal, for some positive integer r, and equip Ω with the Frobenius metric d_F. Positive definite matrices have been studied previously from the random object perspective under different metrics (Arsigny et al. 2007; Pigoli et al. 2014). From a sample (X_i, Y_i), $i = 1, \ldots, n$, the minimization in (2.10) can be reformulated by

setting $\hat{B}(x) = n^{-1} \sum_{i=1}^{n} s_{in}(x) Y_i$ and computing (see proof of Proposition 2 in the supplement for details)

$$\hat{m}_{\oplus}(x) = \operatorname*{argmin}_{\omega \in \Omega} d_F(\hat{B}(x), \omega)^2.$$

Thus, the problem is reduced to finding the correlation matrix which is nearest to the matrix $\hat{B}(x)$. This problem has been well studied (Higham 2002; Qi and Sun 2006; Borsdorf and Higham 2010), and in our implementations we used the alternating projections algorithm, written by Nicholas Higham and available at <https://nickhigham.wordpress.com/2013/02/13/the-nearest-correlation-matrix/>, to carry out this optimization.

7.2. *Functional Connectivity in the Brain.* In recent years, the problem of identifying functional connectivity between brain voxels or regions has received a great deal of attention, especially for resting state fMRI (Allen et al. 2014; Ferreira and Busatto 2013; Lee, Smyser and Shimony 2013; Sheline and Raichle 2013). Subjects are asked to relax while undergoing a fMRI brain scan, where blood-oxygen-level dependent signals are recorded and then processed to yield voxel-specific time courses of signal strength. The connectivity between brain regions is usually quantified by the temporal correlation between representative time signals of the two regions. Higher levels of correlation are reflective of higher connectivity, giving rise to the question of which subject-specific factors might explain observed variations in connectivity. When considering $r > 2$ brain regions, the resulting number of pairwise correlations is $r(r-1)/2$, so that standard statistical models are inadequate for investigating the relationship between several predictors and the connectivity response. Fréchet regression can be employed to directly address this issue by viewing the functional connectivity measurements in a natural way as random elements of the space of correlation matrices.

The data for our analysis come from a study of 174 cognitively normal elderly patients, each of whom underwent an fMRI scan at the UC Davis Imaging Research Center. Preprocessing of the recorded BOLD (blood oxygenation-level-dependent) signals was implemented by adopting the standard procedures of slice-timing correction, head motion correction and normalization, in addition to linear detrending to account for signal drift and band-pass filtering to include only frequencies between 0.01 and 0.08 Hz.

Of particular interest regarding functional connectivity in the resting state is the so-called default-mode network (DMN), including the study of age-related effects (Ferreira and Busatto 2013). In one such study, Mevel et al. (2013) investigated disruptions between anterior-posterior components in the DMN as subjects age and found a decrease in connectivity between

a seed region in the left ventral medial prefrontal cortex (lvmPFC) and three other voxels located within the right vmPFC/orbitofrontal (rvmPFC), left ventral posterior cingulate cortex (lvPCC) and right precuneus/PCC (rpPCC) regions.

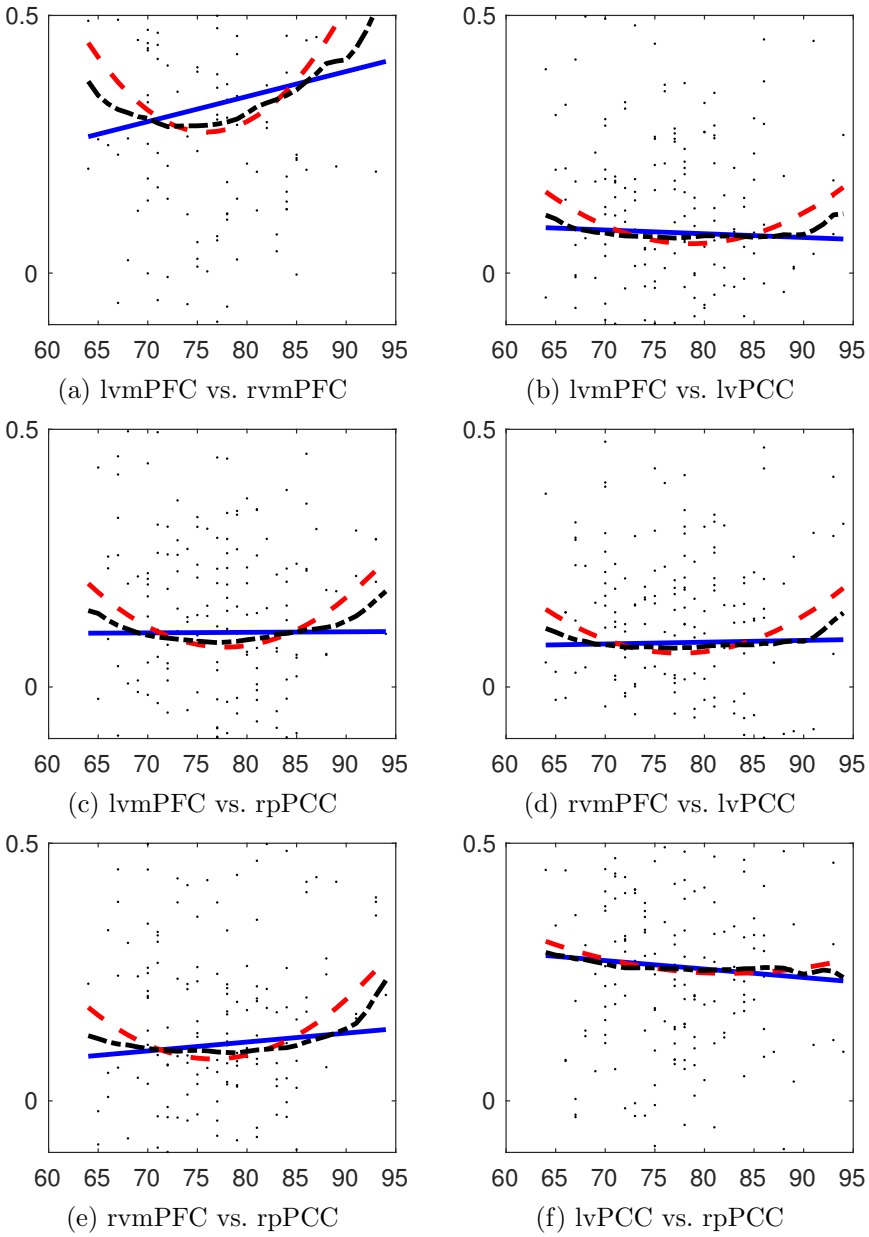
To construct a connectivity correlation matrix for each subject, signals at these $r = 4$ locations were extracted and their temporal correlations computed. These signals are taken over the interval [0, 470] (in seconds), with $T = 236$ measurements available at 2 second intervals. Hence, for the ith subject, the data are in the form of an $T \times r$ signal matrix S_i where the rows correspond to consecutive time points and the columns to distinct voxels. Define $s_{ijk} = (S_i)_{jk}$ and $\bar{s}_{ik} = T^{-1} \sum_{j=1}^{T} s_{ijk}$. The connectivity correlation matrix Y_i for the ith subject as it is routinely calculated for analyzing connectivity in fMRI has the elements

$$(Y_i)_{kl} = \frac{\sum_{j=1}^{T}(s_{ijk} - \bar{s}_{ik})(s_{ijl} - \bar{s}_{il})}{\left[\left(\sum_{j=1}^{T}(s_{ijk} - \bar{s}_{ik})^2\right)\left(\sum_{j=1}^{T}(s_{ijl} - \bar{s}_{il})^2\right)\right]^{1/2}}.$$

In our regression model, we use age as a predictor of connectivity and fit both linear and quadratic models, i.e. $X_i = Z_i$ and $X_i = (Z_i, Z_i^2)^T$, where Z_i is the age of subject i, $i = 1, \ldots, 174$.

Since it is unclear whether the global Fréchet regression model (2.8) holds, we also fit the regression nonparametrically using the Nadaraya-Watson smoother (4.1) over a range of bandwidths. One notable difference between the current data and those used in Mevel et al. (2013) is the age range. The current analysis includes only elderly subjects, aged 64 to 94 years, while Mevel et al. (2013) included subjects between 19 and 80 years of age. It has been observed previously (Onoda, Ishihara and Yamaguchi 2012; Ferreira and Busatto 2013) that age-related effects are more difficult to detect in later years. Thus, the goal for our analysis is to investigate if the decreases in connectivity observed in Mevel et al. (2013) are also found among a group of strictly elderly subjects, or whether the pattern is different.

For each regression fit, the estimated mean-square prediction error (MSPE) was calculated using five-fold cross validation, averaged over 50 runs. The MSPE values for the linear and quadratic models were 0.6489 and 0.6386, respectively. For the Nadaraya-Watson fit, the minimum MSPE over a grid of bandwidths was 0.6393, for bandwidth $h = 7$. The linear model had a p-value of 0.58 and $\hat{R}^2_{\oplus} = 0.0041$, while the quadratic model was a much better fit, with a p-value of 0.014 and $\hat{R}^2_{\oplus} = 0.0288$. Figure 5 illustrates the regression fits for each component of the lower subdiagonal of the correlation matrix. The visual and numerical results suggest that the quadratic



(a) lvmPFC vs. rvmPFC (b) lvmPFC vs. lvPCC

(c) lvmPFC vs. rpPCC (d) rvmPFC vs. lvPCC

(e) rvmPFC vs. rpPCC (f) lvPCC vs. rpPCC

Fig 5: Component-wise scatterplots and Fréchet regression fits for voxel-to-voxel correlation as a function of age. The linear, quadratic, and Nadaraya-Watson ($h = 7$) fits are represented by solid, dashed, and dot-dashed lines, respectively. Note that these fits are derived from Fréchet regression analysis where entire correlation matrices serve as random object responses.

global Fréchet regression model is adequate for these data, as the fit is quite similar to the Nadaraya-Watson estimator without requiring any bandwidth selection. Thus, age-related changes in connectivity seem to be more subtle in later years, with subjects over 85 demonstrating greater connectivity between some regions than younger subjects between the ages of 75 and 85. While some studies have found increased connectivity with age (Ferreira and Busatto 2013), the quadratic model reveals that simple linear associations between age and connectivity may be inadequate.

8. Local Fréchet Regression for Spherical Data. As a final illustration, we implement local Fréchet regression for a situation where the random object responses lie in a Riemannian manifold object space. Specifically, choose $\Omega = S^2$ as the unit sphere in $\mathcal{R}^3$, with geodesic distance $d(y, z) = \arccos(z^T y)$ and consider the regression function

$$m_\oplus(x) = ((1-x^2)^{1/2}\cos(\pi x), (1-x^2)^{1/2}\sin(\pi x), x), \quad x \in (0,1),$$

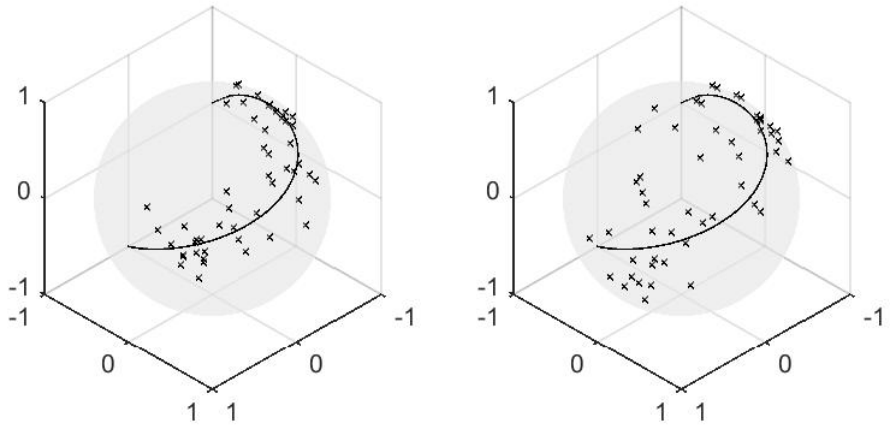
which maps a spiral on the sphere. To generate a random sample (X_i, Y_i), $i = 1, \ldots, n$, $X_i \sim \mathcal{U}(0,1)$ was first sampled, followed by a bivariate normal random vector U_i on the tangent space $T_{m_\oplus(X_i)}\Omega$. Finally, with $\|\cdot\|_E$ being the Euclidean norm,

$$Y_i = \operatorname{Exp}_{m_\oplus(X_i)}(U_i) = \cos(\|U_i\|_E) m_\oplus(X_i) + \sin(\|U_i\|_E)\frac{U_i}{\|U_i\|_E}.$$

Random samples of size $n = 50, 100, 200$ were generated under two noise scenarios, with 200 runs for each simulation. In both noise scenarios, the components of U_i were independent, with each having a variance of 0.2 and 0.35 in the low and high noise scenarios, respectively. Figure 6 shows two sample data sets of size 50 for the two noise scenarios.

For estimation, a grid of bandwidths $h \in (0.05, 0.3)$ was used for the smoothing, with K being the Epanechnikov kernel; this estimation was performed for both local Fréchet regression and the Nadaraya-Watson smoother. The necessary optimization was performed using the trust regions algorithm as implemented in the ManOpt toolbox for Matlab (Boumal et al. 2014). While we found this to be an adequate computational tool for our simulations, it may be necessary in some scenarios to implement a stochastic optimization scheme, such as the annealing algorithm (Yuan et al. 2012). We also implemented an alternative regression method for response data on a nonlinear manifold via smoothing splines (Su et al. 2012), with code provided by one of the authors.

To compare local Fréchet regression with Nadaraya-Watson and spline smoothing, for each combination of noise setting and sample size, the mean



(a) Low Noise, $n = 50$ (b) High Noise, $n = 50$

Fig 6: Sample simulation data sets of size $n = 50$ under low (left) and high (right) noise settings. The true regression curve is shown by the solid line.

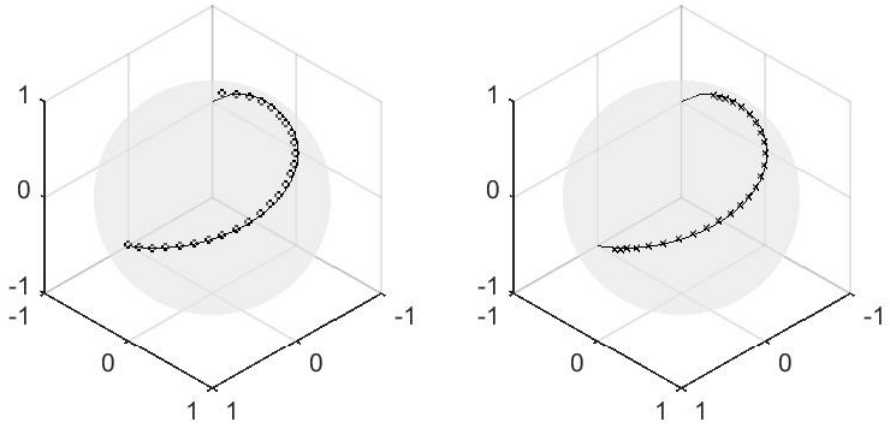
integrated squared error (MISE) of each method was computed across a range of tuning parameters. For our method and Nadaraya-Watson smoothing, this was done over the grid of bandwidths. For the spline method, the three parameters and their values used for each simulation were $T = 50t+1$, $t = 1, 5, 10, 20$, $\epsilon = 10^{-l}$, $l = 2, \ldots, 6$ and $\lambda = 10^k$, $k = -5, \ldots, 5$. The minimum MISE values are shown in Table 1. We see that local Fréchet regression outperforms the other methods in every setting, while the spherical spline method is not a close competitor. Additionally, one can get a sense of the bias of the two Fréchet estimation techniques by taking Fréchet averages of the fits $\hat{m}_{\oplus}(x)$ across simulations, for a grid of levels $x \in [0, 1]$. For example, these averaged local Fréchet and Nadaraya-Watson regression fits, using the bandwidths which minimize MISE, are shown in Figure 7 for the low noise setting with $n = 100$. Again, the local Fréchet method is found to be superior, especially in terms of performance near the boundaries.

9. Discussion. The proposed global and local Fréchet regression models are new tools for the analysis of random objects that are increasingly encountered in modern data analysis. They extend the fundamental notion of a Fréchet mean to that of conditional Fréchet means. We provide theoretical justifications including rates of convergence for both global and local versions. The rates of convergence are optimal for both global and local

TABLE 1

Best MISE values (multiplied by 100 for clarity) for local Fréchet regression (LF), Nadaraya-Watson (NW) and spherical spline (SS) fits. In parentheses, the minimizing bandwidths h are given for the first two methods, while the minimizing triples (T, ϵ, λ) are given for the spline method.

Noise	n	NW	LF	SS
Low	50	1.34(0.13)	0.97(0.22)	5.47 (51, 0.01, 0.01)
	100	0.74(0.13)	0.51(0.19)	5.46 (51, 0.01, 1000)
	200	0.45(0.09)	0.31(0.15)	5.42 (51, 0.01, 100)
High	50	3.00(0.19)	2.61(0.34)	16.99 (251, 0.01, 1)
	100	1.73(0.16)	1.41(0.26)	16.26 (51, 0.001, 0.00001)
	200	0.99(0.13)	0.76(0.21)	13.32 (51, 0.001, 0.00001)



(a) Local Fréchet regression (b) Nadaraya-Watson smoothing

Fig 7: Fréchet-averaged regression curve fits for local Fréchet regression (left, circle markers) and Nadaraya-Watson smoothing (right, 'x' markers), with true regression curve given for reference (solid). These are from the low noise simulation with $n = 100$.

versions in the sense that in the special case of Euclidean objects they correspond to the known optimal rates, and under the same regularity conditions as satisfied for Euclidean objects, the rates remain the same for objects in general metric spaces; we demonstrate this to be the case for the Wasserstein space of distributions as one of many example spaces.

For practical applications of the global Fréchet regression model, we introduce the concept of the Fréchet coefficient of determination, $R^2_\oplus$, and

explore its potential use for testing. We focus in this paper on estimation, and future work will be needed to develop formal tests, confidence sets and predictor selection. For the development of the local version of Fréchet regression it proved necessary to revisit what is meant by the concept of a local regression and to clarify the nature of the target. In data examples, local Fréchet regression proved competitive with previously discussed local smoothing methods for special object spaces. An interesting special case for which we obtain limit distributions is the case of responses that live in a Hilbert space, such as functional data. Indeed, as pointed out by a referee, this model may prove useful in the case of responses which lie on a Hilbert manifold as an extrinsic regression technique in infinite dimensions (see Chapters 11 and 18 of Patrangenaru and Ellingson (2015)).

Conditional Fréchet means and the associated regression approaches have a wide range of applications that include responses that lie in a Riemannian manifold as a special case. For this case we show that our general and straightforward approach is not only theoretically competitive but also works well in simulations. In this and other situations, uniqueness of the Fréchet mean is sometimes not guaranteed, e.g., in the case of a uniform distribution on the sphere, in contrast to other cases that we explored, where it is unique. For manifolds, it is often assumed that Ω is complete in order to prove existence of a Fréchet mean (Bhattacharya and Patrangenaru 2003). Recently, Le and Barden (2014) showed that the cut locus of a minimizer of the Fréchet function necessarily has probability zero, lending further insight into the distributional limitations which allow for existence and uniqueness of Fréchet means. When conditional Fréchet means are not unique, one may need to deal with sets of Fréchet means that consist of many elements (Patrangenaru and Ellingson 2015; Ziezold 1977). Extensions that fall within the framework that we outline also include special types of linear models such as analysis of variance and, more generally, regression models that include indicators among the predictors, as well as polynomial regression models or models with interactions.

SUPPLEMENTARY MATERIAL

Supplement: Proofs of Theoretical Results
(). The supplement includes four sections of proofs. The first section contains proofs of propositions verifying that our theoretical assumptions hold for the examples included in Section 3. The other three contain proofs for each of Sections 3–5.

References.

Afsari, B. (2011). Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society* **139** 655–673.

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T. and Calhoun, V. D. (2014). Tracking Whole-Brain Connectivity Dynamics in the Resting State. *Cerebral Cortex* **24** 663–676.

Arsigny, V., Fillard, P., Pennec, X. and Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* **29** 328–347.

Barden, D., Le, H. and Owen, M. (2013). Central limit theorems for Fréchet means in the space of phylogenetic trees. *Electronic Journal of Probability* **18** 1–25.

Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds - I. *Annals of Statistics* **31** 1–29.

Bhattacharya, R. N., Ellingson, L., Liu, X., Patrangenaru, V. and Crane, M. (2012). Extrinsic analysis on manifolds is computationally faster than intrinsic analysis with applications to quality control by machine vision. *Applied Stochastic Models in Business and Industry* **28** 222–235.

Borsdorf, R. and Higham, N. J. (2010). A preconditioned Newton algorithm for the nearest correlation matrix. *IMA Journal of Numerical Analysis* **30** 94–107.

Boumal, N., Mishra, B., Absil, P.-A., Sepulchre, R. et al. (2014). Manopt, a matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research* **15** 1455–1459.

Bradley, J. V. (1968). *Distribution-free Statistical Tests*. NJ, Prentice-Hall.

Chang, T. (1989). Spherical regression with errors in variables. *Annals of Statistics* **17** 293–306.

Cornea, E., Zhu, H., Kim, P. and Ibrahim, J. G. (2016). Regression models on Riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B*.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematics* **31** 377–403. MR516581 (81g:65018)

Davis, B. C., Fletcher, P. T., Bullitt, E. and Joshi, S. (2007). Population shape regression from random design data. In *ICCV 2007. IEEE 11th International Conference on Computer Vision* 1–7.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall, London. MR1383587 (97f:62063)

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics* **39** 254–261. MR1462586

Faraway, J. J. (2014). Regression for non-Euclidean data using distance matrices. *Journal of Applied Statistics* **41** 2342–2357.

Ferreira, L. K. and Busatto, G. F. (2013). Resting-state functional connectivity in normal brain aging. *Neuroscience & Biobehavioral Reviews* **37** 384–400.

Ferreira, R., Xavier, J., Costeira, J. P. and Barroso, V. (2013). Newton algorithms for Riemannian distance related problems on connected locally symmetric manifolds. *IEEE Journal of Selected Topics in Signal Processing* **7** 634–645.

Fisher, N. I. (1995). *Statistical analysis of circular data*. Cambridge University Press.

Fisher, N. I., Lewis, T. and Embleton, B. J. (1987). *Statistical analysis of spherical data*. Cambridge University Press.

Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision* **105** 171–185.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace dis-

tancié. In *Annales de l'Institut Henri Poincaré* **10** 215–310.

Hein, M. (2009). Robust Nonparametric Regression with Metric-Space valued Output. In *Advances in Neural Information Processing Systems* 718–726.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Brooks/Cole Pacific Grove, CA.

Higham, N. J. (2002). Computing the nearest correlation matrix – a problem from finance. *IMA Journal of Numerical Analysis* **22** 329–343.

Hinkle, J., Muralidharan, P., Fletcher, P. T. and Joshi, S. (2012). Polynomial regression on Riemannian manifolds. In *Computer Vision–ECCV 2012* 1–14. Springer.

Le, H. and Barden, D. (2014). On the measure of the cut locus of a Fréchet mean. *Bulletin of the London Mathematical Society* bdu025.

Lee, M., Smyser, C. and Shimony, J. (2013). Resting-state fMRI: a review of methods and clinical applications. *American Journal of Neuroradiology* **34** 1866–1872.

Lehmann, E. L. and D'Abrera, H. J. (2006). *Nonparametrics: statistical methods based on ranks*. Springer New York.

Lin, L., Thomas, B. S., Zhu, H. and Dunson, D. B. (2015). Extrinsic local regression on manifold-valued data. *arXiv preprint arXiv:1508.02201*.

Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal* **56** 732–753.

Marx, B. and Eilers, B. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science* **11** 89–121.

Mevel, K., Landeau, B., Fouquet, M., La Joie, R., Villain, N., Mézenge, F., Perrotin, A., Eustache, F., Desgranges, B. and Chételat, G. (2013). Age effect on the default mode network, inner thoughts, and cognitive abilities. *Neurobiology of Aging* **34** 1292–1301.

Niethammer, M., Huang, Y. and Vialard, F.-X. (2011). Geodesic regression for image time-series. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011* 655–662. Springer.

Onoda, K., Ishihara, M. and Yamaguchi, S. (2012). Decreased functional connectivity by aging is associated with cognitive decline. *Journal of Cognitive Neuroscience* **24** 2186–2198.

Panaretos, V. M. and Zemel, Y. (2016). Amplitude and phase variation of point processes. *The Annals of Statistics* **44** 771–812.

Patrangenaru, V. and Ellingson, L. (2015). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press.

Pelletier, B. (2006). Non-parametric regression estimation on closed Riemannian manifolds. *Journal of Nonparametric Statistics* **18** 57–67.

Pigoli, D., Aston, J. A., Dryden, I. L. and Secchi, P. (2014). Distances and inference for covariance operators. *Biometrika* **101** 409–422.

Prentice, M. J. (1989). Spherical regression on matched pairs of orientation statistics. *Journal of the Royal Statistical Society: Series B* 241–248.

Qi, H. and Sun, D. (2006). A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM Journal on Matrix Analysis and Applications* **28** 360–385.

Sheline, Y. I. and Raichle, M. E. (2013). Resting state functional connectivity in preclinical Alzheimer's disease. *Biological Psychiatry* **74** 340–347.

Shi, X., Styner, M., Lieberman, J., Ibrahim, J. G., Lin, W. and Zhu, H. (2009). Intrinsic regression models for manifold-valued data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009* 192–199. Springer.

Steinke, F. and Hein, M. (2009). Non-parametric regression between manifolds. In *Ad-*

vances in Neural Information Processing Systems 1561–1568.

Steinke, F., Hein, M. and Schölkopf, B. (2010). Nonparametric regression between general Riemannian manifolds. *SIAM Journal on Imaging Sciences* **3** 527–563.

Su, J., Dryden, I. L., Klassen, E., Le, H. and Srivastava, A. (2012). Fitting smoothing splines to time-indexed, noisy points on nonlinear manifolds. *Image and Vision Computing* **30** 428–442.

Takatsu, A. (2011). Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics* **48** 1005–1026.

Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Wang, H., Marron, J. et al. (2007). Object oriented data analysis: Sets of trees. *Annals of Statistics* **35** 1849–1873.

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

Ziezold, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians* 591–602. Springer.

Address of Alexander Petersen
Statistics and Applied Probability
University of California
Santa Barbara, CA 93106-3110
E-mail: petersen@pstat.ucsb.edu

Address of Hans-Georg Müller
Department of Statistics
Mathematical Sciences Building 4118
399 Crocker Lane
University of California, Davis
One Shields Avenue
Davis, CA 95616
E-mail: hgmueller@ucdavis.edu