HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2017 December 08.

Published in final edited form as:

J Am Stat Assoc. 2017 ; 112(519): 1261–1273. doi:10.1080/01621459.2016.1208615.

Extrinsic local regression on manifold-valued data

Lizhen Lin,

Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX

Brian St Thomas,

Department of Statistical Science, Duke University, Durham, NC

Hongtu Zhu, and

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX and University of North Carolina, Chapel Hill, NC

David B. Dunson

Department of Statistical Science, Duke University, Durham, NC

Abstract

We propose an extrinsic regression framework for modeling data with manifold valued responses and Euclidean predictors. Regression with manifold responses has wide applications in shape analysis, neuroscience, medical imaging and many other areas. Our approach embeds the manifold where the responses lie onto a higher dimensional Euclidean space, obtains a local regression estimate in that space, and then projects this estimate back onto the image of the manifold. Outside the regression setting both intrinsic and extrinsic approaches have been proposed for modeling i.i.d manifold-valued data. However, to our knowledge our work is the first to take an extrinsic approach to the regression problem. The proposed extrinsic regression framework is general, computationally efficient and theoretically appealing. Asymptotic distributions and convergence rates of the extrinsic regression estimates are derived and a large class of examples are considered indicating the wide applicability of our approach.

Keywords

Convergence rate; Differentiable manifold; Geometry; Local regression; Object data; Shape statistics

1 Introduction

Although the main focus in statistics has been on data belonging to Euclidean spaces, it is common for data to have support on non-Euclidean geometric spaces. Perhaps the simplest example is to directional data, which lie on *circles or spheres*. Directional statistics dates back to R.A. Fisher's seminal paper (Fisher, 1953) on analyzing the directions of the earth's magnetic poles, with key later developments by Watson (1983), Mardia and Jupp (2000), Fisher et al. (1987) among others. Technological advances in science and engineering have led to the routine collection of more complex geometric data. For example, diffusion tensor

imaging (DTI) obtains local information on the directions of neural activity through 3 × 3 *positive definite matrices* at each voxel (Alexander et al., 2007). In machine vision, a digital image can be represented by a set of k-landmarks, the collection of which form *landmark based shape spaces* (Kendall, 1984). In engineering and machine learning, images are often preprocessed or reduced to a collection of *subspaces*, with each data point (an image) in the sample data represented by a subspace. One may also encounter data that are stored as *orthonormal frames*, *surfaces*, *curves*, and *networks*.

Statistical analysis of data sets whose basic elements are geometric objects requires a precise mathematical characterization of the underlying space and inference is dependent on the geometry of the space. In many cases (e.g., space of positive definite matrices, spheres, shape spaces, etc), the underlying space corresponds to a *manifold*. Manifolds are general topological spaces equipped with a differentiable/smooth structure which induces a geometry that does not in general adhere to the usual Euclidean geometry. Therefore, new statistical theory and models have to be developed for statistical inference of manifold-valued data. There have been some developments on inferences based on i.i.d (independent and identically distributed) observations on a known manifold. Such approaches are mainly based on obtaining statistical estimators for appropriate notions of location and spread on the manifold. For example, one could base inference on the center of a distribution on the Fréchet mean, with the asymptotic distribution of sample estimates obtained (Bhattacharya and Patrangenaru, 2003, 2005; Bhattacharya and Lin, 2016). There has also been some consideration of nonparametric density estimation on manifolds (Bhattacharya and Dunson, 2010; Lin et al., 2016; Pelletier, 2005). Bhattacharya and Bhattacharya (2012) provides a recent overview of such developments.

There has also been a growing interest in modeling the relationship between a manifold-valued response Y and Euclidean predictors X. For example, many studies are devoted to investigating how brain shape changes with age, demographic factors, IQ and other variables. It is essential to take into account the underlying geometry of the manifold for proper inference. Approaches that ignore the geometry of the data can potentially lead to highly misleading predictions and inferences. Some geometric approaches have been developed in the literature. For example, Fletcher (2011) develops a geodesic regression model on Riemannian manifolds, which can be viewed as a counterpart of linear regression on manifolds, and subsequent work of Hinkle et al. (2012) generalizes polynomial regression model to the manifold. These parametric and semi-parametric models are elegant, but may lack sufficient flexibility in certain applications. Shi et al. (2009) proposes a semi-parametric intrinsic regression model on manifolds, and Davis et al. (2007) generalizes an intrinsic kernel regression method on the Riemannian manifold, considering applications in modeling changes in brain shape over time. Yuan et al. (2012) develops an intrinsic local polynomial model on the space of symmetric positive definite matrices, which has applications in diffusion tensor imaging. A drawback of intrinsic models is the heavy computational burden incurred by minimizing a complex objective function along geodesics, typically requiring evaluation of an expensive gradient in an iterated algorithm. The objective functions often have multiple modes, leading to large sensitivity to start points. Further, existence and uniqueness of the population regression function holds only under

relatively restrictive support conditions. Therefore, usual descent algorithms used in estimation are not guaranteed to converge to a global optima.

With the motivation of developing general purpose computationally efficient, theoretically sound and practically useful regression modeling frameworks for manifold-valued response data, we propose a nonparametric extrinsic regression model by first embedding the manifold where the response resides onto some higher-dimensional Euclidean spaces. We use equivariant embeddings, which preserve a great deal of geometry for the images. A local regression estimate (such as a local polynomial estimate) of the regression function is obtained after embedding, which is then projected back onto the image of the manifold. Outside the regression setting, both intrinsic and extrinsic approaches have been proposed for modeling of manifold-valued data and for mathematically studying the properties of manifolds. However, to our knowledge, our work is the first in taking an extrinsic approach in the regression modeling context. Our approach is general, has elegant asymptotic theory and outperforms intrinsic models in terms of computation efficiency. In addition, there is essentially no difference in inference with the examples considered.

The article is organized as follows. Section 2 introduces the extrinsic regression framework. In Section 3, we explore the full utilities of our method through applications to three examples in which the response resides on different manifolds. A simulation study is carried out for data on the sphere (example 1) applying both intrinsic and extrinsic models. The results indicate the overall superiority of our extrinsic method in terms of computational complexity and time compared to that of intrinsic methods. The extrinsic models are also applied to planar shape manifolds in example 2, with applications considered to simulated data and to modeling the brain shape of the Corpus Callosum from an ADHD (Attention Deficit/Hyperactivity Disorder) study. In example 3, our method is applied to data on the Grassmannian considering both simulated and real data. Section 4 is devoted to studying the asymptotic properties of our estimators in terms of asymptotic distribution and convergence rate.

2 Extrinsic local regression on manifolds

Let $Y \in M$ be the response variable in a regression model where (M, ρ) is a general metric space with distance metric ρ. Let $X \in \mathbb{R}^m$ be the covariate or predictor variable which can be random or fixed. Given data (x_i, y_i) $(i = 1, \ldots, n)$, the goal is to model a regression relationship between Y and X. The typical regression framework with $y_i = F(x_i) + \epsilon_i$ is not appropriate here as expressions like $y_i - F(x_i)$ are not well-defined due to the fact that the space M (e.g., a manifold) where the response variable lies is in general not a vector space. Let $P(x, y)$ be the joint distribution of (X, Y) and $P(x)$ be the marginal distribution of X with marginal density $f_X(x)$. Denote $P(y|x)$ as the conditional distribution of Y given X with conditional density $p(y|x)$. One can define the population regression function or map $F(x)$ (if it exists) as

$$F(x) = \underset{q \in M}{\operatorname{argmin}} \int_M \rho^2(q, y)\, P(dy|x), \quad (1)$$

where ρ is a distance metric on M.

Let M be a d-dimensional differentiable or smooth manifold. A manifold M is a topological space that locally behaves like a Euclidean space. In order to equip M with a metric space structure, one can employ a Riemannian structure, with ρ taken to be the geodesic distance, which defines an *intrinsic regression function*. Alternatively, one can embed the manifold onto some higher dimensional Euclidean space via an embedding map J and use the Euclidean distance $\|\cdot\|$ instead. The latter model is referred to as an *extrinsic regression model*. One of the potential hurdles for carrying out intrinsic analysis is that uniqueness of the population regression function in (1) (with ρ taken to be the geodesic distance) can be hard to verify. Le and Barden (2014) establish several interesting results for the regression framework and provide broader conditions for verifying the uniqueness of the population regression function. Intrinsic models can be computationally expensive, since minimizing their complex objective functions typically require a gradient descent type algorithm. In general, this requires fine tuning at each step, which results in an excessive computational burden. Further, these gradient descent algorithms are not always guaranteed to converge to a global minimum or only converge under very restrictive conditions. In contrast, the uniqueness of the population regression holds under very general conditions for extrinsic models. Extrinsic models are extremely easy to evaluate and are orders of magnitude faster than intrinsic models.

Let $J: M \to E^D$ be an embedding of M onto some higher dimensional ($D \geq d$) Euclidean space E^D and denote the image of the embedding as $\tilde{M} = J(M)$. By the definition of embedding, the differential of J is a map between the tangent space of M at q and the tangent space of E^D at $J(q)$; that is, $d_q J: T_q M \to T_{J(q)} E^D$ is an injective map and J is a homeomorphism of M onto its image $\tilde{M}$. Here $T_q M$ is the tangent space of M at q and $T_{J(q)} E^D$ is the tangent space of E^D at $J(q)$. Let $\|\cdot\|$ be the Euclidean norm. In an extrinsic model, the true extrinsic regression function is defined as

$$F(x) = \underset{q \in M}{\operatorname{argmin}} \int_M \|J(q) - J(y)\|^2 P(dy|x) = \underset{q \in M}{\operatorname{argmin}} \int_{\tilde{M}} \|J(q) - z\|^2 \tilde{P}(dz|x) \tag{2}$$

where $\tilde{P}(\cdot \mid x) = P(\cdot \mid x) \circ J^{-1}$ is the conditional probability measure on $J(M)$ given x induced by the conditional probability measure $P(\cdot \mid x)$ via the embedding J.

We now proceed to propose an estimator for $F(x)$. Let $K: \mathbb{R}^m \to \mathbb{R}$ be a multivariate kernel function such that $\int_{\mathbb{R}^m} K(x)dx = 1$ and $\int_{\mathbb{R}^m} xK(x)dx = 0$. One can take K to be a product of m one-dimensional kernel functions for example. Let $H = \mathrm{Diag}(h_1, \ldots, h_m)$ with $h_i > 0$ ($i = 1, \ldots, m$) be the bandwidth vector and $|H| = h_1 \ldots h_m$. Let $K_H(x) = \frac{1}{|H|} K(H^{-1}x)$ and

$$\hat{F}(x) = \underset{y \in E^D}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{K_H(x_i - x) \|y - J(y_i)\|^2}{\sum_{i=1}^{n} K_H(x_i - x)} = \sum_{i=1}^{n} \frac{J(y_i) K_H(x_i - x)}{\sum_{i=1}^{n} K_H(x_i - x)}, \tag{3}$$

which is basically a weighted average of points $J(y_1), \ldots, J(y_n)$. We are now ready to define the *extrinsic kernel estimate of the regression function* $F(x)$ as

$$\hat{F}_E(x) = J^{-1}(\mathscr{P}(\hat{F}(x))) = J^{-1}(\operatorname*{argmin}_{q \in \tilde{M}} \|q - \hat{F}(x)\|), \quad (4)$$

where $\mathcal{P}$ denotes the projection map onto the image $\tilde{M}$. Basically, our estimation procedure consists of two steps. In step one, it calculates a local regression estimate on the Euclidean space after embedding. In step two, the estimate obtained in step one is projected back onto the image of the manifold. Although we assume the projection is unique, uniqueness needs to be verified for each manifold and embedding. In general, we require that the image $\tilde{M}$ is closed in the Euclidean space. These conditions tend to be straightforward to show, as is illustrated for the examples considered in Section 3.

Note that, alternatively, we can obtain some robust estimator under our proposed framework by first proposing a robust estimator of $\tilde{F}_E(x)$. This can be done by replacing the terms with $\|\cdot\|^2$ in equation (3) with a term using $\|\cdot\|$.

A kernel estimate is obtained first in (3) before projection. However, the framework can be easily generalized using higher order local polynomial regression estimates (of degree p) (Fan and Gijbels, 1996). For example, one can have a *local linear estimator* (Fan, 1993) for $\hat{F}(x)$ before projection. That is, for any x, let

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname*{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} \|J(y_i) - \beta_0 - \beta_1^t (x_i - x)\|^2 K_H(x_i - x) \Big). \quad (5)$$

Then, we have

$$\hat{F}(x) = \hat{\beta}_0(x), \quad (6)$$

$$\hat{F}_E(x) = J^{-1}(\mathscr{P}(\hat{F}(x)) = J^{-1}(\operatorname*{argmin}_{q \in \tilde{M}} \|q - \hat{F}(x)\|). \quad (7)$$

The properties of the estimator $\hat{F}_E(x)$ where $\hat{F}(x)$ is given by the general pth local polynomial estimator of $J(y_i), \ldots, J(y_n)$ are explored in Theorem 4.4.

Remark 1

The embedding J used in the extrinsic regression model is in general not unique. It is desirable to have an embedding that preserves as much geometry as possible. An *equivariant embedding* preserves a substantial amount of geometry. Let G be some large Lie group

acting on M. We say that J is an equivariant embedding if we can find a group homomorphism $\phi: G \rightarrow GL(D, \mathbb{R})$ from G to the general linear group $GL(D, \mathbb{R})$ of degree D such that

$$J(gq) = \phi(g) J(q)$$

for any $g \in G$ and $q \in M$. The intuition behind equivariant embedding is that the image of M under the group action of the Lie group G is preserved by the group action of $\phi(G)$ on the image, thus preserving many geometric features. Note that the choice of embedding is not unique and in some cases constructing an equivariant embedding can be a non-trivial task, but in most of the cases a natural embedding arises and such embeddings can often be verified as equivariant.

Note that our work addresses different problems from that of Cheng and Wu (2013), which provides an elegant framework for high dimensional data analysis and manifold learning by first performing local linear regression on a tangent plane estimate of a lower-dimensional manifold where the high-dimensional data concentrate.

3 Examples and applications

The proposed extrinsic regression framework is very general and has appealing asymptotic properties as will be shown in Section 4. To illustrate the wide applicability of our approach and validate its finite sample performance, we carry out a study by applying our method to various examples with the response taking values in different well-known manifolds. For each of the examples considered, we provide details on the embeddings, verify such embeddings are equivariant, and give explicit expressions for the projections to obtain the final estimate in each case. In example 1, we simulate data from a 2-dimensional sphere and compare the estimates from our extrinsic regression model with that of an intrinsic model. The result indicates that the extrinsic models clearly outperform the intrinsic models by orders of magnitude in terms of computational complexity and time. In example 2, we first study a simulated example where a comparison study shows even greater computational gain for the extrinsic model over the intrinsic one compared with the sphere case. We then consider a data example with response a planar shape, in which the brain shape of the subjects are represented by landmarks on the boundary. Example 3 provides details of the estimator when the responses take values on a Stiefel or Grassmann manifold. The method is illustrated with a synthetic data set and a solar flare data set, both of which have subspace responses of possibly mixed dimension and covariates, which are the corresponding time points.

We will not consider an example with DTI responses below due to page concerns, but the extrinsic model can be applied in a similar fashion by using the log matrix map as the embedding. Yuan et al. (2013) considers varying coefficient model in which the log matrix map is also used.

Example 1

Statistics on the 2-dimensional sphere S^2, often called *directional statistics*, has a long history (Fisher, 1953; Watson, 1983; Mardia and Jupp, 2000; Fisher et al., 1987). Marzio et al. (2014) considers a smoothing model for regression with both predictors and responses on spheres. Recently, Wang and Lerman (2015) applied a nonparametric Bayesian regression model to an example with response on the circle S^1. We first work out the details for the extrinsic regression method with the responses lying on a d-dimensional sphere S^d, then illustrate the model with simulation data $\{(x_i, y_i), i = 1, \ldots, n\}$, where $y_i \in S^2$, the 2-dimensional sphere.

Note that S^d is a submanifold of $\mathbb{R}^{d+1}$; therefore, the inclusion map $\iota : S^d \rightarrow \mathbb{R}^{d+1}$, where $\iota(y) = y$ serves as a natural embedding onto $\mathbb{R}^{d+1}$. It is easy to check that the embedding is an equivariant embedding. The intuition behind this embedding is that it preserves a lot of the symmetry of the sphere. Given $J(y_1), \ldots, J(y_n)$ with the embedding $J = \iota$, one first obtains $\hat{F}(x)$ as given in (3). Its projection onto the image $\tilde{M}$ is given by

$$\hat{F}_E(x) = \hat{F}(x) / \|\hat{F}(x)\|, \text{when} \hat{F}(x) \neq 0.$$

In the following, we consider a simulation study for a regression model with responses on the 2-dimensional unit sphere. The objective of this simulation study is to illustrate the application of the proposed extrinsic regression framework to data with sphere-valued response and to demonstrate the computational advantages over the intrinsic methods via a comparison study. To simulate the data, first consider a common and useful distribution, the von Mises-Fisher distribution (Fisher, 1953) on the unit sphere, which has the following density:

$$p_{MF}(y; \mu, \kappa) \propto \exp\left(\kappa \mu^T y\right),$$

where κ is a concentration parameter with μ a location parameter. We simulate the data (the y values) from the unit sphere by letting the mean function be covariate-dependent. In particular, for this example, we will use data generated by the following model

$$\begin{gathered} \beta \sim N_3(0, I), \quad x_i^1 \sim N(0,1), \quad x_i^2 \sim N(0,1), \quad x_i^3 = x_i^1 * x_i^2, \quad i = 1, \ldots, n \\ y_i \sim MF(\mu_i, \kappa), \quad \mu_i = \frac{\beta \circ x_i}{|\beta \circ x_i|}, \quad \kappa \text{some fixed known value,} \end{gathered} \tag{9}$$

where $\beta \bigcirc x$ is the Hadamard product $(\beta_1 x^1, \ldots, \beta_m x^m)$.

As an example of what the data looks like, we generate one thousand ($n = 1000$) observations from the above model with $\kappa = 10$ so that realizations are near their expected value. Figure 1 shows this example in which 100 predictions from the extrinsic model are plotted against their true values using 900 training points. To select the bandwidth h we use

10-fold cross-validation with h ranging from [.1, .2, …, 1.9, 2] and choose the value that gives minimum average mean square error. Residuals for the mean square error are measured using the intrinsic distance, or great circle distance, on the sphere.

To illustrate the utility and advantages of extrinsic regression models, we compare our method to an *intrinsic kernel regression model* that uses intrinsic distance of the sphere to minimize the objective function. Computations on the sphere are in general not as intensive compared to more complicated manifolds such as shape spaces, etc, but it still requires an iterative algorithm, such as gradient descent, for the intrinsic model in order to obtain a kernel regression estimate. The following simulation results demonstrate extrinsic kernel regression gives at least as accurate estimates as intrinsic kernel regression but in much less computation time even for S^2.

The intrinsic kernel regression estimate minimizes objective function $f(y) = \sum_{i=1}^{n} w_i d^2(y, y_i)$, where y and y_i are points on the sphere S^2, W_i are determined by the Gaussian kernel function, and $d(\cdot, \cdot)$ in this case is the great circle distance. Then the gradient of f on the sphere is given by

$$\nabla f(y) = \sum_{i=1}^{n} w_i 2d(y, y_i) \frac{\log_y(y_i)}{d(y, y_i)} = \sum_{i=1} 2w_i \frac{\arccos\left(y^T y_i\right)}{\sqrt{1 - (y^T y_i)^2}} \left(y_i - \left(y^T y_i\right) y\right).$$

where $log_y(y_i)$ is the log map or the inverse exponential map on the sphere. Estimates for y can be obtained through a gradient descent algorithm with step size δ and error threshold ϵ. We applied the intrinsic and extrinsic models to the same set of data using the Gaussian kernel function.

Twenty different data sets of 2000 observations were generated from the above sphere regression model with von-Mises Fisher concentration parameter $\kappa = \{1, 2, \ldots, 20\}$. Of the 2000 observations, 50 were used to check the accuracy of the extrinsic and intrinsic estimates. To see the effect of training sample size on the quality of the estimates, the estimates were also made on subsets of the 1950 training observations, starting with 2 observations and increasing to all 1950 observations. The same training observations were always used for both models. In both models, the bandwidth was chosen through 10-fold cross validation. The intrinsic kernel regression was fit with step size $\delta = .01$ and error threshold $\epsilon = .001$. The performance of the two methods is compared in terms of MSE and predictive MSE. The MSE is calculated using the great circle distance between predicted values and the true expected value, while predictive MSE is calculated using the great circle distance between the predicted values and the realized values. The performance results using 50 hold out observations can be seen in Figure 2.

Predictive MSE does not converge to 0 because the generating distribution has a high variance; however, as the concentration increases, the predictive MSE does approach 0. The extrinsic and intrinsic kernel regressions perform similarly with large sample sizes. The extrinsic kernel regression drops in predictive MSE faster than the intrinsic model, which

may stem from only having the kernel bandwidth as a tuning parameter which can be selected more easily than choosing the bandwidth, step-size, and error thresholds even through cross-validation.

A significant advantage of the extrinsic kernel regression is the speed of computation. Both methods were implemented in C++ using Rcpp (Eddelbuettel and Fran¸cois, 2011), and resulted in up to a 60× improvement in speed in making a single prediction using all of the training observations. For speed comparisons, a single prediction was made given the same number of test observations, and the time to produce the estimate was recorded. Each of these trials was done five times, and we compare the mean time to producing the estimate in Figure 3.

Note that the same kernel weights are computed in both algorithms, so the difference is attributable to the gradient descent versus extrinsic optimization procedures. Since the speed comparisons were done for computing a single prediction and the difference is due almost entirely to the gradient descent steps, making multiple predictions results in an even more favorable comparison for the extrinsic model. This experiment shows that the extrinsic kernel regression applied to sphere data performs at least as well on prediction and can be computed significantly faster.

Example 2

We now consider an example with responses on *planar shapes*. Planar shapes are one of the most important classes of landmark based shapes spaces. Such spaces were defined by Kendall (1977, 1984) with pioneering work by Bookstein (1978) motivated from applications on biological shapes. We now describe the geometry of the space which will be used in obtaining regression estimates for our model. Roughly speaking, the planar shape consists of a collection of k-landmarks modulo the action of Euclidean motions such as translations, scalings and rotations. Let $z = (z_1, \ldots, z_k)$ with $z_1, \ldots, z_k \in \mathbb{R}^2$ be a set of k landmarks, and $< z >= (\bar{z}, \ldots, \bar{z})$ where $\overline{z}=\sum_{i=1}^{k} z_i/k$. We first center and normalize z to get $u=\frac{z-<z>}{\|z-<z>\|}$. u can be viewed as an element on some high-dimensional sphere S^{2k-3}, which is called the *pre-shape*. The *planar shape* $\sum_2^k$ can now be represented as the quotient of the pre-shape under the action of the rotation group $SO(2)$, or the 2 by 2 special orthogonal group. Therefore, a point on the planar shape can be identified as the orbit or equivalent of z which we denote by $\sigma(z)$. Viewing z as elements in the complex plane, one can embed $\sum_2^k$ onto the $S(k, \mathbb{C})$, the space of $k \times k$ complex Hermitian matrices via the Veronese-Whitney embedding (see e.g. Bhattacharya and Bhattacharya (2012)):

$$J(\sigma(z)) = uu* = ((u_i \overline{u}_j))_{1 \leq, i, j \leq k}, \quad (10)$$

where u^* is the conjugate transpose of u and $\bar{u}_j$ is the conjugate of u_j. One can verify the embedding is equivariant (see Kendall (1984)) by taking the Lie group G to be the special unitary group $SU(k) = \{A \in GL(k, \mathbb{C}), AA^* = I, det(A) = I\}$.

We now describe the projection after $\hat{F}(x)$ is given by (3), where $J(y_i)$ ($i = 1, \ldots, n$) are obtained using the embedding given in (10). Letting $\upsilon(x)$ be the eigenvector corresponding to largest eigenvalue of $\hat{F}(x)$, by a careful calculation, one can show that the projection of $\hat{F}(x)$ is given by

$$\mathscr{P}_{J(M)}\left(\hat{F}(x)\right) = \upsilon(x)\,\upsilon(x)^*.$$

Therefore, the extrinsic kernel regression estimate is given by

$$\hat{F}_E(x)\,J^{-1}\left(\upsilon(x)\,\upsilon(x)^*\right). \quad (11)$$

Comparison to Intrinsic model on synthetic data set—We compare the extrinsic model to an intrinsic model on synthetic planar shape data to understand if the great computational benefits observed for sphere data extend to other manifolds. Intuitively, as the Log map on a manifold grows in complexity, we would expect that the gains from using the extrinsic method would also grow, since we can avoid iteratively computing the Log map.

Planar shape data were generated using a scheme for polar coordinates. First, we generate m-dimensional covariates for the observation that will be linked to the responding shape. For each of the k landmarks that are in the data set, we generate an intercept for that landmark by getting one angle in $[0, 2\pi]$ and one radius in $\mathbb{R}^+$. Together, these specify an intercept shape. We add random noise, centered at a function of the covariates, to the angle and radius, potentially using different functions. This procedure generates K ($k = 1, \ldots, K$) landmarks linked to m-dimensional ($j = 1, \ldots, m$) covariates for N ($n = 1, \ldots, N$) observations.

$$\begin{array}{cc}
\phi_{0k} \sim Unif\,(0, 2\pi) & \text{Generate intercept angles} \\
r_{0k} \sim N\left(r, \sigma_{r_0}^2\right) & \text{Generate intercept radii} \\
x_{n,j} \sim Ga\,(a, b) & \text{Generate covariate} \\
r_{n,k} \sim N\left(r_{0k} + \beta_1 x_{n,1}, \sigma_r^2\right) & \text{Generate shape radii} \\
\phi'_{n,k} \sim N\left(\phi_{0k} + \beta_2 x_{n,2} + \beta_3 x_{n,3}, \sigma_\phi^2\right) & \text{Generate shape angles} \\
\phi_{n,k} \text{mod} \phi'_{n,k}\,(\text{mod} 2\pi) & \text{Standardize angles} \\
z_{n,k} = r_{n,k}\left(\cos\left(\phi_{n,k}\right) + i \sin\left(\phi_{n,k}\right)\right) & \text{Convert to complex form for the landmark.}
\end{array}$$

Here $y_n = (z_{n,1}, \ldots, z_{n,K})$ is the nth response on the planar shape for covariate $x_n = (x_{n,1}, x_{n,2}, x_{n,3})$. In our case, we simplified testing by letting $\sigma_r = \sigma_\phi$ for values in $\{0.1, 0.2, \ldots, 2\}$. See Figure 4 for an example of planar shapes resulting from this procedure with a low level of noise ($\sigma = 0.1$).

For the intrinsic model, we can use the same method as before in the sphere example with a gradient descent type algorithm for obtaining the estimate. We replace the log map on the sphere with a log map for the planar shape. If $p_{\tilde{y}}(y) = \tilde{y}\langle\tilde{y}, y\rangle/\|\tilde{y}\|^2$ is the projection of y onto $\tilde{y}$, the Log map between two points $\tilde{y}$, y on the planar shape is defined as

$$\mathrm{Log}_{\tilde{y}}(y) = \frac{\theta\left(y - p_{\tilde{y}}(y)\right)}{\|y - p_{\tilde{y}}(y)\|}, \text{where}\,\theta = \arccos\frac{|\langle\tilde{y}, y\rangle|}{\|\tilde{y}\|\|y\|}.$$

We simulated 2000 observations with 3 covariates and 20 landmarks from our synthetic data procedure, and held 50 out as a validation set for measuring the predictive error. The kernel bandwidth was chosen for each model using 10-fold cross validation on the full training set. We measured the training error and predictive error for training sample sizes starting at 100, increasing to 1950 by steps of 25. When predicting the holdout sample of 50, we tracked the computation time to make the estimate. The results are shown in Figure 5.

The results are consistent with what we expected from both theory and what we observed from the sphere example. The performance in terms of root mean squared error, which is measured intrinsically on the shape space for both models, is similar. However, the computation time is drastically reduced for the extrinsic model, with the extrinsic model being hundreds of times faster than the intrinsic model.

We also noticed in this example that the intrinsic model was much more sensitive to the choice of bandwidth. When inspecting the RMSE results of each validation test, the extrinsic RMSE results could vary from 1 - 2, while the intrinsic RMSE could vary from 1 - 15 over the same bandwidth range. Because the choice of intrinsic bandwidth is so important, this might explain why the intrinsic model seems to slightly over fit on lower training sample sizes, leading to the slightly worse predictive RMSE and slightly better training RMSE.

Corpus Callosum (CC) data set—We study ADHD-200 dataset [1] in which the shape contour of the brain Corpus Callosum is recorded for each subject along with variables such as gender, age, and ADHD diagnosis. 50 landmarks were placed outlining the CC shape for 647 patients for the ADHD-200 dataset. The age of the patients range from 7 to 21 years old, with 404 typically developing children and 243 individuals diagnosed with some form of ADHD. The original data set differentiates between types of ADHD diagnoses, and we simplify the problem of choosing a kernel by using a binary response for an ADHD diagnosis.

According to the findings in Huang et al. (2015), there is not a significant effect of gender on the area of different segments of the CC; however diagnosis and the interaction between diagnosis and age were found to be statistically significant ($p < .01$). With knowledge of these results, we performed the extrinsic kernel regression method for the CC planar shape response using diagnosis, x^1, and age, x^2, as covariates. Therefore, one is interested in the

[1] http://fcon_1000.projects.nitrc.org/indi/adhd200/

regression analysis of y (the planar shape) as a function of age and diagnostic status. The choice of kernel between two sets of covariates $x_1 = \left(x_1^1, x_1^2\right)$ and $x_2 = \left(x_2^1, x_2^2\right)$ is

$$K_H(x_1, x_2) = \begin{cases} \exp\left(-\frac{\left(x_1^2 - x_2^2\right)^2}{h}\right)/h^2 & \text{if} x_1^1 \equiv x_2^1 \\ 0 & \text{if} x_1^1 \neq x_2^1. \end{cases}$$

The motivation for using this kernel is that one wishes to essentially run local smoothing within each diagnostic group given the significant diagnostic variable. We visualize how the CC shape develops over time by making predictions at different time points. We show predictions for ages 9, 12, 16, and 19 year old children of ADHD diagnosis or typical development. The results can be seen in Figure 6.

What we can observe from the two plots is that the CC shapes for the 8 year olds seem to be close, but by age 12 the shapes have diverged substantially, with shrinking of the CC being apparent in later years in development. This quality of the CC shapes between ADHD and normal development is consistent with results found in the literature (Huang et al., 2015).

In previous studies, ADHD diagnoses were clustered using the shape information to predict the diagnosis class, and the centroid of the cluster is the predicted shape for that class (Huang et al., 2015). Our method adds to this analysis from a regression perspective and predicts the CC shape as a function of age and diagnosis. Our method also has the benefit of evaluating quickly, making selection of the bandwidth for the kernel through cross-validation feasible.

Example 3

We now consider another two important manifolds, the Stiefel manifolds and Grassman manifolds (Grassmannians). The Stiefel manifold, $V_k(\mathbb{R}^m)$, is the collection of k orthonormal frames in $\mathbb{R}^m$, which consists of k ordered unit vectors in $\mathbb{R}^m$ that are orthonormal to each other. That is, $V_k(\mathbb{R}^m) = \{X \in S(m, k), XX^T = I_m\}$. The Stiefel manifold includes the m dimensional sphere S^m as a special case with k=1 and $O(m)$ the orthogonal group when $k = m$. The Stiefel manifold is a compact manifold of dimension $km - k - k(k-1)/2$ and it is a submanifold of $\mathbb{R}^{km}$. The inclusion map onto $\mathbb{R}^{km}$ can be further shown to be an equivariant embedding. Applications of Stiefel manifold are present in earth sciences, medicine, astronomy, meteorology and biology. Examples of data on the Stiefel manifold include the orbit of the comets and the vector cardiogram. As stated in Chikuse (2003), the vector cardiogram is in general considered as an oriented closed-space curve generated by a point moving in time, and each point on the curve represents the resultant electrical activity of the heart at that instant. A vector cardiogram (the orientation) is represented by two orthonormal unit vectors in $\mathbb{R}^3$, thus a point in $V_{2,3}$. Similarly, the orientations of the orbits of the comets given by the direction of the perihelion and the directed unit normal vector to the orbit can also be represented by elements in $V_{2,3}$.

Considering the extrinsic regression method for Stifel manifold-valued response data, we first obtain $\hat{F}(x)$, and the next step is to obtain the projection of $\hat{F}(x)$ onto $\tilde{M} = J(M)$. We first make an *orthogonal decomposition* of $\hat{F}(x)$ by letting $\hat{F}(x) = U(x)S(x)$, where $U(x) \in V_{k,m}$, which can be viewed as the orientation of $\hat{F}(x)$ and $S(x)$ is positive semi-definite, which has the same rank as $\hat{F}(x)$. Then the projection of $\hat{F}(x)$ (or projection set) is given by

$$\mathscr{P}_{\tilde{M}}\left(\hat{F}(x)\right) = \left\{U(x) \in V_{k,m} : \hat{F}(x) = U(x)\left(\hat{F}(x)^T \hat{F}(x)\right)^{1/2}\right\}.$$

See Theorem 10.2 in Bhattacharya and Bhattacharya (2012) for a proof of this result. The projection is unique, i.e., the above set is a singleton if and only if $\hat{F}(x)$ is of full rank.

The Grassmann manifold or the Grassmannian $Gr_k(\mathbb{R}^m)$ is the space of all the subspaces of a fixed dimension k whose basis elements are k orthonormal unit vectors in $\mathbb{R}^m$, which is closely related to the Stiefel manifold $V_{k,m}$. The key difference between a point on the Grassmannian and a point on the Stiefel manifold is that the ordering of the k orthonormal vectors in $\mathbb{R}^m$ does not matter for the former. The Grassmannian can be viewed as the quotient space of the Stiefel manifold modulo $O(k)$, the k by k orthogonal group. That is, $Gr_k(\mathbb{R}^m) = V_k(\mathbb{R}^m)/O(k)$. A point on the Stiefel manifold can be viewed as a representative of the orbits for the Grassmannian. The equivariant embedding for $Gr_k(\mathbb{R}^m)$ also exists (Chikuse, 2003). Let $X \in V_{k,m}$ be a representative element of any equivalent class in $Gr_k(\mathbb{R}^m)$. So a point in the Grassmannian can be represented by the orbit $\sigma(X) = XR$ where $R \in O(k)$. Then an embedding can be given by

$$J(\sigma(X)) = XX^T.$$

The collection of XX^T forms a subspace of $\mathbb{R}^{m2}$. We can verify that J is an equivariant embedding under the group action of $G = O(m)$.

There are many applications of Grassmann manifolds, in which the subspaces are the basic element in signal processing, machine learning and so on. We consider a regression model with subspace valued response. Given the estimate $\hat{F}(x)$, the next step is to derive the projection of $\hat{F}(x)$ onto $\tilde{M} = J(M)$. Since all XX^T form a subspace, one can use the following procedure to calculate the projection map of $\hat{F}(x)$ to the Grassmann manifold by finding an orthonormal basis for the image. This algorithm is a special case of the projection via Conway embedding (St. Thomas et al., 2014).

1. Find the eigendecomposition $\hat{F}(x) = Q\Lambda Q^{-1}$
2. Take the k eigenvectors corresponding to the top k eigenvalues in Λ as an orthonormal basis for $\hat{F}_E(x)$, $Q_{[1:k,]}$.

We consider two illustrative examples, one synthetic and one from a series of images from a solar flare, for extrinsic kernel regression with subspace valued response variables. In the examples, we allow the responses to be subspaces of *different dimensions*. The technique is

unique compared to other subspace regression techniques because the extrinsic distance offers a well defined and principled distance between responses of different dimension. This avoids the need to constrain the responses to be a fixed dimension or hard coding a heuristic distance between subspaces of different dimension into the distance function.

We consider a synthetic example in which the predictors are the time points and the responses are points on the Grassmann manifold. We draw orthonormal bases from the Matrix von Mises-Fisher distribution as their representation. We generate N draws from the following process with concentration parameter κ, in which the first n_1 draws are of dimension 4 and the last n_2 draws are of dimension 5,

for $1 \leq t \leq N$ **do**
  Draw $X \sim MN(0, I_m, I_5)$
  $\mu_{[,1]} := t + X_{[,1]}$, $\mu_{[,2]} := t - X_{[,2]}$, $\mu_{[,3]} := t^2 + X_{[,3]}$, $\mu_{[,4]} := tX_{[,4]}$
  if $t > n_1$ **then**
    $\mu_{[,5]} := t + tX_{[,5]}$
  end if
  $Y_t := \mathrm{vMF}(\kappa M)$
end for

Here the only covariate associated with Y_t is t. With a concentration of $\kappa = 1$, and $n_1 = n_2 = 50$, we generate much noisier data than before, and are able to correctly predict the dimension of the subspace at each time point. The predicted dimension at each time point and the residuals are plotted in Figure 7.

The key advantage of this method is not requiring any constraints on the dimension of the input or output subspaces. This is important in some examples, such as the solar flare example we will illustrate. The solar flare data consists of a large quantity of images in a series that is difficult to analyze. By dividing the images into smaller sets, and summarizing each set of images as a subspace, we reduce the amount of data and processing power required to analyze when a solar flare may have activated. In some cases, because of sporadic activity, we are not guaranteed that the dimension of the subspace is the same, leading to substantial problems in implementing intrinsic methods.

We apply this method to a series of images from a solar flare in Hall and Willett (2015). The data contains 300 snapshots of 232 × 292 pixel data, which were collected from the Solar Data Observatory. For each set of ten images, we vectorize the pixel data and concatenate the vectors to obtain a matrix for subspace estimation. The left singular vectors give the subspace spanning the images, with the dimension chosen by the top d singular values explaining 90% of the total variation. The extrinsic kernel regression procedure is then applied to the 30 periods and their images are recovered treating the kernel estimate of the subspace as the new left singular vectors. The original data and the recovered estimates at the given time point (measured in snapshots) can be seen in Figure 8.

When there is no solar flare, the subspace describing each image set is fairly static, and the kernel regression can be trained to be quite smooth. When looking at the residuals of the

kernel regression, it becomes very obvious when the solar flare activity begins and ends. The residuals $\|Y - \hat{Y}\|_2$ of each image and the estimated image are shown in Figure 9.

For all the examples considered above for which equavariant embeddings are available, extrinsic approaches are in general advantageous over the intrinsic models. But there are complex manifolds such as higher-dimensional shape spaces for which good embedding are hard to construct. For these cases, we expect intrinsic models to perform better than extrinsic ones.

4 Asymptotic properties of the extrinsic regression model

In this section, we investigate the large sample properties of our extrinsic regression estimates. We assume the marginal density $f_X(x)$ is differentiable and the absolute value of any of the partial derivatives of $f_X(x)$ of order two are bounded by some constant C. In our proof, we assume our kernel function K takes a product form. That is, $K(x) = K_1(x^1) \ldots K_m(x^m)$ where $x = (x^1, \ldots, x^m)$ and $K_1, \ldots, K_m$ are one dimensional symmetric kernels such that $\int_{\mathbb{R}} K_i(u)du = 1$, $\int_{\mathbb{R}} uK_i(u)du = 0$ and $\int_{\mathbb{R}} u^2K_i(u)du < \infty$ for $i = 1, \ldots, m$. The results can be generalized to kernels with arbitrary form and with H given by a more general positive definite matrix instead of a diagonal matrix. Theorem 4.1 derives the asymptotic distribution of the extrinsic regression estimate $\hat{F}_E(x)$ for any x.

Theorem 4.1

Let $\mu(x) = E(\tilde{P}(dy|x))$, which is the conditional mean regression function of $\tilde{P}$ and assume $\mu(x)$ is differentiable. Assume $n|H| \to \infty$. Denote $x = (x^1, \ldots, x^m)$. Let

$\tilde{\mu}(x) = \mu(x) + \frac{Z(x)}{f_X(x)}$, where the ith component $Z_i(x)$ ($i = 1, \ldots, D$) of $Z(x)$ is given by

$$Z_i(x) = h_1^2 \left(\frac{\partial f}{\partial x^1} \frac{\partial \mu_i}{\partial x^1} + \frac{1}{2} f_X(x) \left(\frac{\partial^2 \mu_i}{\partial (x^1)^2} + \ldots + \frac{\partial^2 \mu_i}{\partial x^m x^1} \right) \right) \int v_1^2 K_1(v_1)\, dv_1 + \ldots + h_m^2 \left(\frac{\partial f}{\partial x^m} \frac{\partial \mu_i}{\partial x^m} + \frac{1}{2} f_X(x) \left(\frac{\partial^2 \mu_i}{\partial x^1 x^m} + \ldots + \frac{\partial^2 \mu_i}{\partial (x^m)^2} \right) \right). \tag{12}$$

Assume the projection $\mathcal{P}$ of $\tilde{\mu}(x)$ onto $\tilde{M} = J(M)$ is unique and $\mathcal{P}$ is continuously differentiable in a neighborhood of $\tilde{\mu}(x)$. Then the following holds assuming $P(dy \mid x) \circ J^{-1}$ has finite second moments:

$$\sqrt{n|H|}\, d_{\tilde{\mu}(x)} \mathscr{P} \left(\hat{F}(x) - \tilde{\mu}(x) \right) \xrightarrow{L} N\left(0, \tilde{\Sigma}(x)\right), \tag{13}$$

where $d_{\tilde{\mu}(x)} \mathcal{P}$ is the differential from $T_{\tilde{\mu}(x)}\mathbb{R}^D$ to $T_{\mathcal{P}(\tilde{\mu}(x))}\tilde{M}$ of the projection map $\mathcal{P}$ at

$\tilde{\mu}(x) = \mu(x) + \frac{Z(x)}{f_X(x)}$. Here $\tilde{\Sigma}(x) = B^T\bar{\Sigma}(x)B$, where B is the $D \times d$ matrix of the differential

$d_{\tilde{\mu}(x)}\mathcal{P}$ with respect to given orthonormal bases of $T_{\tilde{\mu}(x)}\mathbb{R}^D$ and $T_{\mathcal{P}(\tilde{\mu}(x))}\tilde{M}$, and the (j, k)th entry of $\bar{\Sigma}(x)$ is given by (14) with

$$\overline{\sum}_{jk}=\frac{\sigma\left(J_j\left(y\right), J_k\left(y\right)\right)\int K(\upsilon)^2 d\upsilon}{f_X\left(x\right)}, \quad (14)$$

where $\sigma(J_j(y), J_k(y)) = \mathrm{Cov}(J_j, J_k)$, and J_j is the jth element of $J(y)$. Here $\xrightarrow{L}$ indicates convergence in distribution.

Corollary 4.2 is on the mean integrated squared error of the estimates.

Corollary 4.2

Assuming the same conditions of Theorem 4.1 and the covariate space is bounded, the mean integrated squared error of $\hat{F}_E(x)$ is of the order $O(n^{-4/(m+4)})$, with the choice of h_i's ($i = 1, \ldots, m$) to be of the same order, that is, of $O(n^{-1/(m+4)})$.

Remark 2—Note that in nonparametric regression with both predictors (m-dimensional) and responses in the Euclidean space, the optimal order of the mean integrated squared error is $O(n^{-4/(m+4)})$ under the assumption that the true regression function has bounded second derivative. Our method achieves the same rates. However, whether such rates are minimax in the context of manifold valued response is not known.

Theorem 4.3 shows some results on uniform convergence rates of the estimator.

Theorem 4.3

Assume the covariate space $x \in \chi \subset \mathbb{R}^m$ is compact and $\mathcal{P}$ has a continuous first derivative. Then

$$\sup_{x\in\chi}\|d_{\tilde{\mu}(x)}\mathscr{P}\left(\hat{F}\left(x\right)-E\left(\hat{F}\left(x\right)\right)\right)\|=O_p\left(\log^{1/2}n/\sqrt{n|H}\right). \quad (15)$$

As pointed out in Section 2, it is ideal in many cases to fit a higher order (say pth order) local polynomial model in estimating $\mu(x)$ before projecting back onto the image of the manifold. Such estimates are more appealing especially when $F(x)$ is more curved over a neighborhood of x. One can show that similar results as those of Theorem 4.1 hold, though with much more involved argument.

We now give details of such estimators and their asymptotic distributions are derived in Theorem 4.4. Recall $F(x) = E(P(dy \mid x))$ and $\mu(x) = E(\tilde{P}(dy \mid x))$ and $J(y_1), \ldots, J(y_n)$ are the points on $\tilde{M} = J(M)$ after embedding J. We first obtain an estimate $\hat{F}(x)$ of $\mu(x)$ using pth order local polynomial estimation. The intermediate estimate $\tilde{F}(x)$ is then projected back to $\tilde{M}$ to obtain the final estimate of $F(x)$. The general framework is given as follows:

$$\left\{\hat{\beta}_k^j(x)\right\}_{0\le|k|\le p,1\le j\le D} = \underset{\left\{\beta_k^j(x)\right\}_{0\le|k|\le p,1\le j\le D}}{\operatorname{argmin}} \sum_{i=1}^{n}\left(\left\|J(y_i)-\left(\sum_{0\le|k|\le p}\beta_k^1(x)(x_i-x)^{|k|},\ldots,\sum_{0\le|k|\le p}\beta_k^D(x)(x_i-x)^{|k|}\right)^T\right\|^2\times K_H(x_i-x)\right). \quad (16)$$

Some of the notation used in (16) are given as follows:

$$\boldsymbol{k}=(\boldsymbol{k}_1,\ldots,\boldsymbol{k}_m),\quad |\boldsymbol{k}|=\sum_{l=1}^{m}\boldsymbol{k}_l,\quad |\boldsymbol{k}|\in\{0,\ldots,p\},$$
$$\boldsymbol{k}!=k_1!\times\ldots\times k_m!,\quad x^k=(x^1)^{k_1}\times\ldots\times(x^m)^{k_m}\quad \sum_{0\le|k|\le p}=\sum_{j=0}^{p}\sum_{k_1=0}^{j}\cdots\sum_{\substack{k_m=0\\|k|=k_1+\ldots+k_m=j}}^{j}.$$

When $\boldsymbol{k}=\boldsymbol{0}$, $\left(\hat{\beta}_0^1,\ldots\hat{\beta}_0^D\right)^T$ corresponds to the kernel estimator, which is the same as the estimator given in (3). When $p=1$, $\left(\hat{\beta}_{\boldsymbol{k}=0}^1,\ldots\hat{\beta}_{\boldsymbol{k}=0}^D\right)^T$ coincides with the estimator $\hat{\boldsymbol{\beta}}_0$ in (5).

Finally, we have

$$\hat{F}(x)=\hat{\beta}_0(x)=\left(\hat{\beta}_{\boldsymbol{k}=0}^1,\ldots,\hat{\beta}_{\boldsymbol{k}=0}^D\right)^T, \quad (17)$$

$$\hat{F}_E(x)=J^{-1}\left(\mathscr{P}\left(\hat{F}(x)\right)\right)=J^{-1}\left(\underset{q\in\tilde{M}}{\operatorname{argmin}}\|q-\hat{F}(x)\|\right). \quad (18)$$

Theorem 4.4 derives the asymptotic distribution of $\hat{F}_E(x)$, with $\hat{F}(x)$ obtained using pth order polynomial local regression of $J(y_1)$, …, $J(y_n)$ given in (17).

Theorem 4.4

Let $\hat{F}_E(x)$ be given in (18). Assume the $(p+2)$th moment of the kernel function $K(x)$ exists and $\mu(x)$ is $(p+2)$th order differentiable in a neighborhood of $x=(x^1,\ldots,x^m)$. Assume the projection $\mathcal{P}$ of $\mu(x)$ onto $\tilde{M}=J(M)$ is unique and $\mathcal{P}$ is continuously differentiable in a neighborhood of $\tilde{\mu}(x)$, where $\tilde{\mu}(x)=\mu(x)+\text{Bias}(x)$, with Bias($x$) given in equations (20) and (21) of the web supplementary. If $P(dy\mid x)\circ J^{-1}$ has finite second moments, then we have:

$$\sqrt{n|H|}d_{\tilde{\mu}(x)}\mathscr{P}\left(\hat{F}(x)-\tilde{\mu}(x)\right)\xrightarrow{L}N\left(0,\tilde{\sum}(x)\right), \quad (19)$$

where $d_{\tilde{\mu}(x)}\mathcal{P}$ is the differential from $T_{\tilde{\mu}(x)}\mathbb{R}^D$ to $T_{\mathcal{P}\tilde{\mu}(x)}\tilde{M}$ of the projection map $\mathcal{P}$ at $\tilde{\mu}(x)$. Here $\Sigma(x) = B^T\bar{\Sigma}(x)B$, where B is the $D \times d$ matrix of the differential $d_{\tilde{\mu}(x)}\mathcal{P}$ with respect to given orthonormal basis of tangent space $T_{\tilde{\mu}(x)}\mathbb{R}^D$ and tangent space $T_{\mathcal{P}\tilde{\mu}(x)}\tilde{M}$ and the jkth entry of $\bar{\Sigma}(x)$ is given by (14). Here $\xrightarrow{L}$ indicates convergence in distribution.

Remark 3—Note that the order of the bias term Bias(x) differs when p is even and when p is odd (see the web supplementary for more details).

Remark 4—Our theoretical results are characterized in terms of the integrated mean squared error and the asymptotic distribution of the regression estimate. Wang and Lerman (2015) uses a Bayesian nonparametric model which provides a posterior distribution on the regression function, and the theoretical results are quantified in terms of posterior contraction rates. Bayesian inference for regression on manifold is in general difficult due to the inherent difficulty in specifying a valid likelihood. Further, full Bayesian inference requires developing MCMC algorithms for sampling the posterior distribution which can be highly non-trivial and also computationally extensive.

5 Conclusion

We have proposed an extrinsic regression framework for modeling data with manifold valued responses and shown desirable asymptotic properties of the resulting estimators. We applied this framework to a variety of applications, such as responses restricted to the sphere, shape spaces, and linear subspaces. The principle motivating this framework is that kernel regression and Riemannian geometry both rely on locally Euclidean structures. This property allows us to construct inexpensive estimators without loss of predictive accuracy as demonstrated by the asymptotic behavior of the mean integrated square error, and also the empirical results. Empirical results even suggest that the extrinsic estimators may perform better due to their reduced complexity and ease of optimizing tuning parameters such as kernel bandwidth. Future work may also use this principle to guide sampling methodology when trying to sample parameters from a manifold or optimizing an EM-algorithm, where it may be computationally or mathematically difficult to restrict intermediate steps to the manifold.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We sincerely thank the Associate Editor and two reviewers for their constructive comments on an earlier version of the paper. Dr. Lin's research was partially funded by NSF IIS1546331. Dr. Zhu's work was partially supported by NIH grants MH086633 and 1UL1TR001111, NSF grants SES-1357666 and DMS-1407655, and a grant from Cancer Prevention Research Institute of Texas.

References

Alexander A, Lee J, Lazar M, Field A. Diffusion tensor imaging of the brain. Neurotherapeutics. 2007; 4(3):316–329. [PubMed: 17599699]

Bhattacharya, A., Bhattacharya, R. Nonparametric Inference on Manifolds: With Applications to Shape Spaces IMS Monograph #2. Cambridge University Press; 2012.

Bhattacharya A, Dunson DB. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. Biometrika. 2010; 97(4):851–865. [PubMed: 22822255]

Bhattacharya R, Lin L. Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. The Proceedings of the American Mathematical Society. 2016 to appear.

Bhattacharya RN, Patrangenaru V. Large sample theory of intrinsic and extrinsic sample means on manifolds. Ann Statist. 2003; 31:1–29.

Bhattacharya RN, Patrangenaru V. Large sample theory of intrinsic and extrinsic sample means on manifolds-ii. Ann Statist. 2005; 33:1225–1259.

Bookstein, F. The Measurement of Biological Shape and Shape Change Lecture Notes in Biomathematics. Springer; Berlin: 1978.

Cheng M, Wu H. Local linear regression on manifolds and its geometric interpretation. Journal of the American Statistical Association. 2013; 108(504):1421–1434.

Chikuse, Y. Statistics on Special Manifolds. Springer; New York: 2003.

Davis B, Fletcher P, Bullitt E, Joshi S. Population shape regression from random design data. ICCV 2007 IEEE 11th International Conference on. 2007:1–7.

Eddelbuettel D, Fran¸cois R. Rcpp: Seamless R and C++ integration. Journal of Statistical Software. 2011; 40(8):1–18.

Fan J. Local linear regression smoothers and their minimax efficiencies. Ann Statist. 1993; 21(1):196–216.

Fan, J., Gijbels, I. Local Polynomial Modelling and Its Applications. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis; 1996.

Fisher, N., Lewis, T., Embleton, B. Statistical Analysis of Spherical Data. Cambridge Uni. Press; Cambridge: 1987.

Fisher R. Dispersion on a sphere. Proc Roy Soc London Ser A. 1953; 217:295–305.

Fletcher T. Geodesic Regression on Riemannian Manifolds. MICCAI Workshop on Mathematical Foundations of Computational Anatomy (MFCA). 2011:75–86.

Hall E, Willett R. Online convex optimization in dynamic environments. Selected Topics in Signal Processing, IEEE Journal of. 2015; 9(4):647–662.

Hinkle, J., Muralidharan, P., Fletcher, P., Joshi, S. Polynomial regression on riemannian manifolds. In: Fitzgibbon, A.Lazebnik, S.Perona, P.Sato, Y., Schmid, C., editors. Computer Vision ECCV 2012, volume 7574 of Lecture Notes in Computer Science. Springer; Berlin Heidelberg: 2012. p. 1-14.

Huang C, Styner M, Zhu H. Penalized mixtures of offset-normal shape factor analyzers with application in clustering high-dimensional shape data. J Amer Statist Assoc. 2015 to appear.

Kendall DG. The diffusion of shape. Adv Appl Probab. 1977; 9:428–430.

Kendall DG. Shape manifolds, procrustean metrics, and complex projective spaces. Bull of the London Math Soc. 1984; 16:81–121.

Le H, Barden D. On the measure of the cut locus of a Fréchet mean. Bulletin of the London Mathematical Society. 2014; 46(4):698–708.

Lin L, Rao V, Dunson DB. Bayesian nonparametric inference on the Stiefel manifold. Statistics Sinica. 2016 to appear., Arxive 1311.0907.

Mardia, K., Jupp, P. Directional Statistics. Wiley; New York: 2000.

Marzio MD, Panzera A, Taylor CC. Nonparametric regression for spherical data. Journal of the American Statistical Association. 2014; 109(506):748–763.

Pelletier B. Kernel density estimation on Riemannian manifolds. Statistics and Probability Letters. 2005; 73(3):297–304.

Shi X, Styner M, Lieberman J, Ibrahim JG, Lin W, Zhu H. Intrinsic regression models for manifold-valued data. Med Image Comput Comput Assist Interv. 2009; 12(2):192–199. [PubMed: 20426112]

St Thomas B, Lin L, Lim LH, Mukherjee S. Learning subspaces of different dimension. ArXiv e-prints. 2014 1404.6841.

Wang X, Lerman G. Nonparametric Bayesian regression on manifolds via Brownian motion. ArXiv e-prints. 2015

Watson, GS. Statistics on Spheres. Vol. 6. University Arkansas Lecture Notes in the Mathematical Sciences, Wiley; New York: 1983.

Yuan Y, Zhu H, Lin W, Marron JS. Local polynomial regression for symmetric positive definite matrices. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012; 74(4): 697–719. [PubMed: 23008683]

Yuan Y, Zhu H, Styner M, Gilmore JH, Marron JS. Varying coefficient model for modeling diffusion tensors along white matter tracts. Ann Appl Stat. 2013; 7(1):102–125. [PubMed: 24533040]

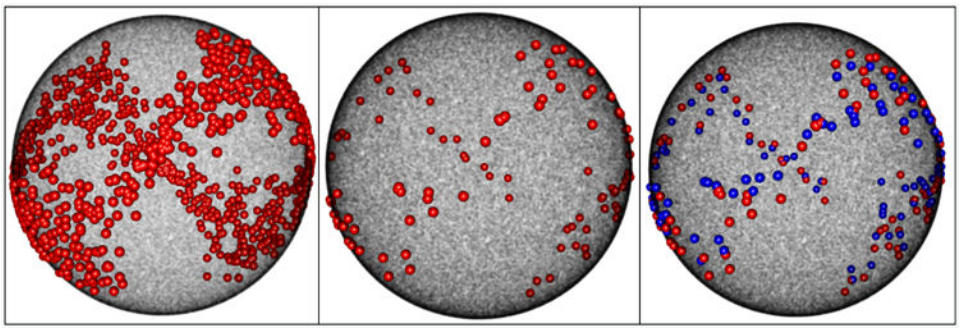

Figure 1.

Left The training values on the sphere. **Middle** The held out values to be predicted through extrinsic regression. **Right** The extrinsic predictions (blue) plotted against the true values (red).

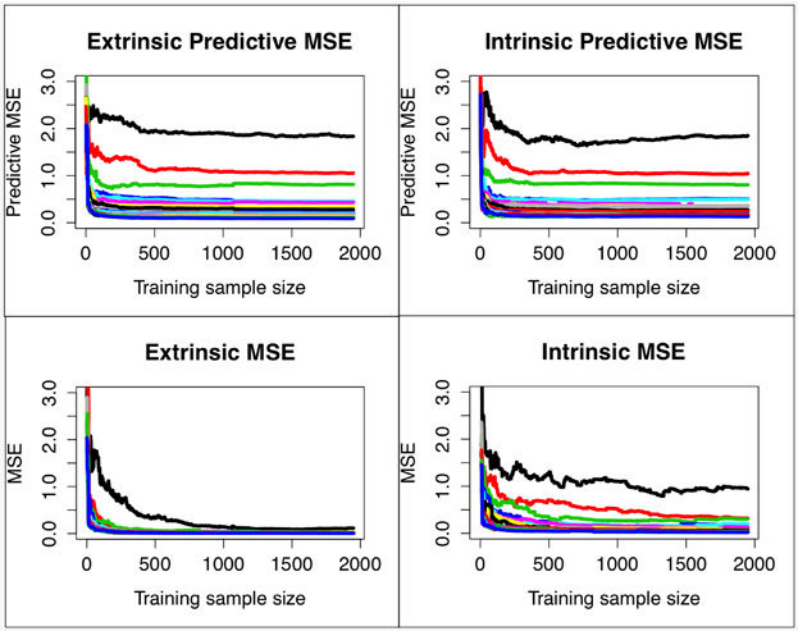

Figure 2.

The performance of extrinsic and intrinsic regression models on 50 test observations from sphere regression models with concentration parameters from 1 to 20. Each color corresponds to a concentration parameter. The extrinsic and intrinsic models have similar performance in predictive MSE with low concentration parameters. However in terms of MSE, the extrinsic model appears to perform better with lower sample sizes even with lower concentration parameters.

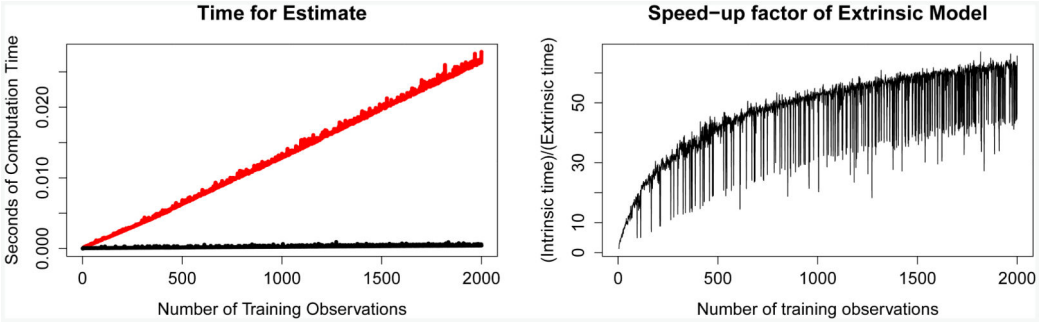

Figure 3.

Speed comparisons between the extrinsic and intrinsic kernel regressions as a function of the number of training observations. The average seconds to produce an estimate for a single test observation are plotted in red for the intrinsic model, and black for the extrinsic model. The multiple between the speed for the intrinsic and extrinsic estimates plotted are also plotted for reference.

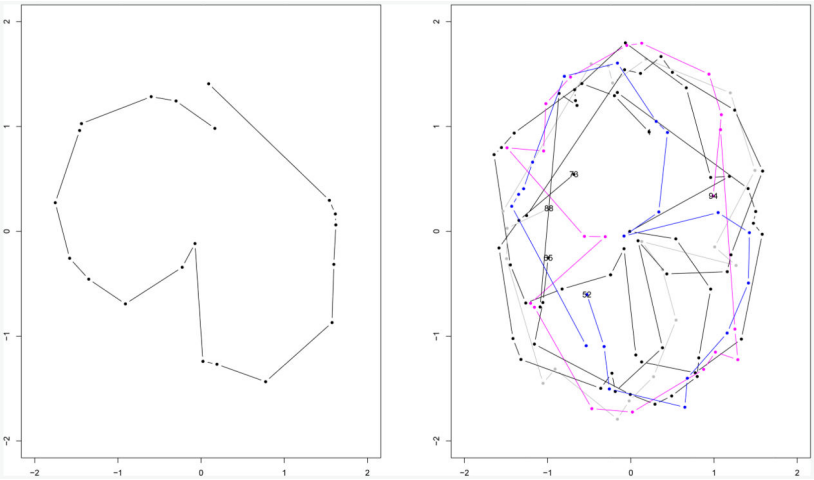

Figure 4.

Examples of synthetic planar shapes with 20 landmarks generated using $\sigma_r = \sigma_\phi = 0.1$. The variation in shape is driven by the covariates linked to each shape and the idiosyncratic error.

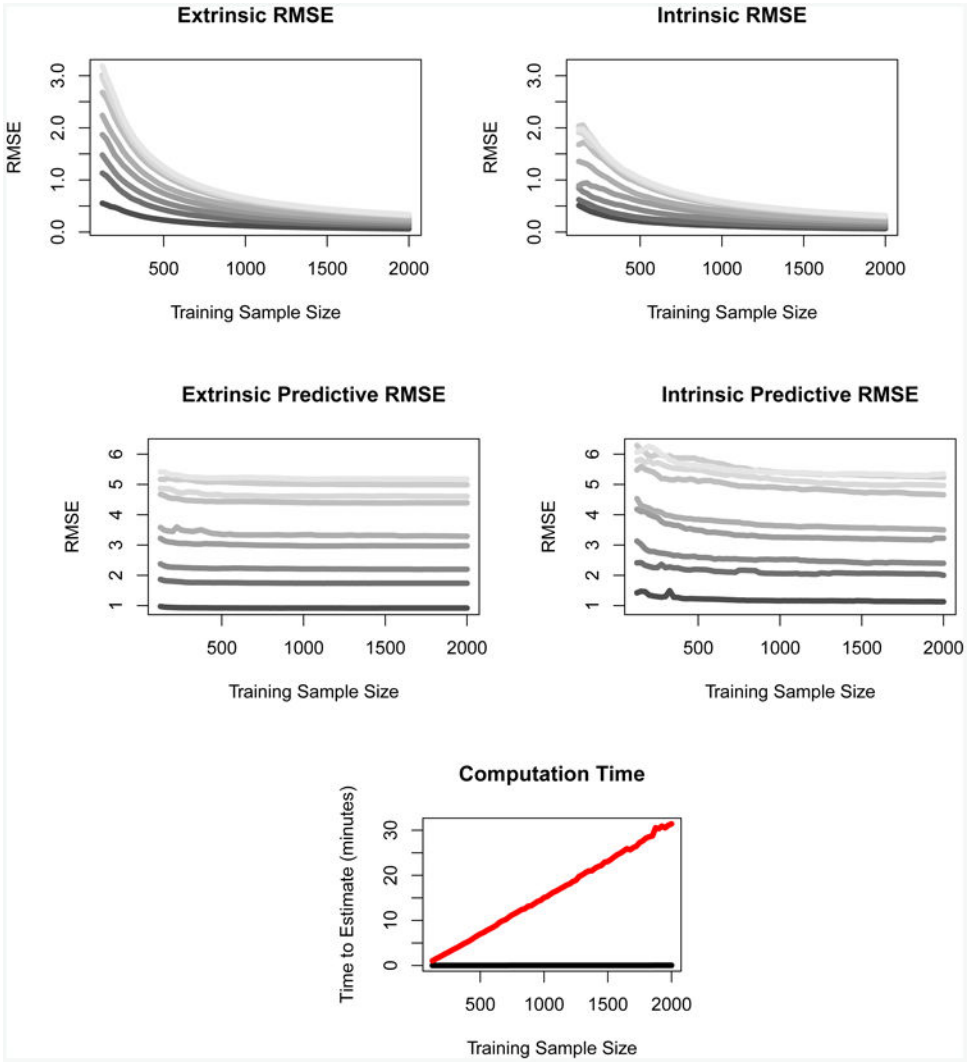

Figure 5.

Results from training Intrinsic and Extrinsic models on synthetic planar shape data. Each line in the RMSE plots correspond to synthetic data generated from the same variance level in {0.1, …, 2.0}. For computation time, the red line is the Intrinsic model, the black line is the Extrinsic model. Each point is the average computation time over all the variance levels tested. Like in the sphere model, performance is similar in terms of RMSE. The most noticeable difference between the two is the computation time (in minutes) for the intrinsic model to make estimates.

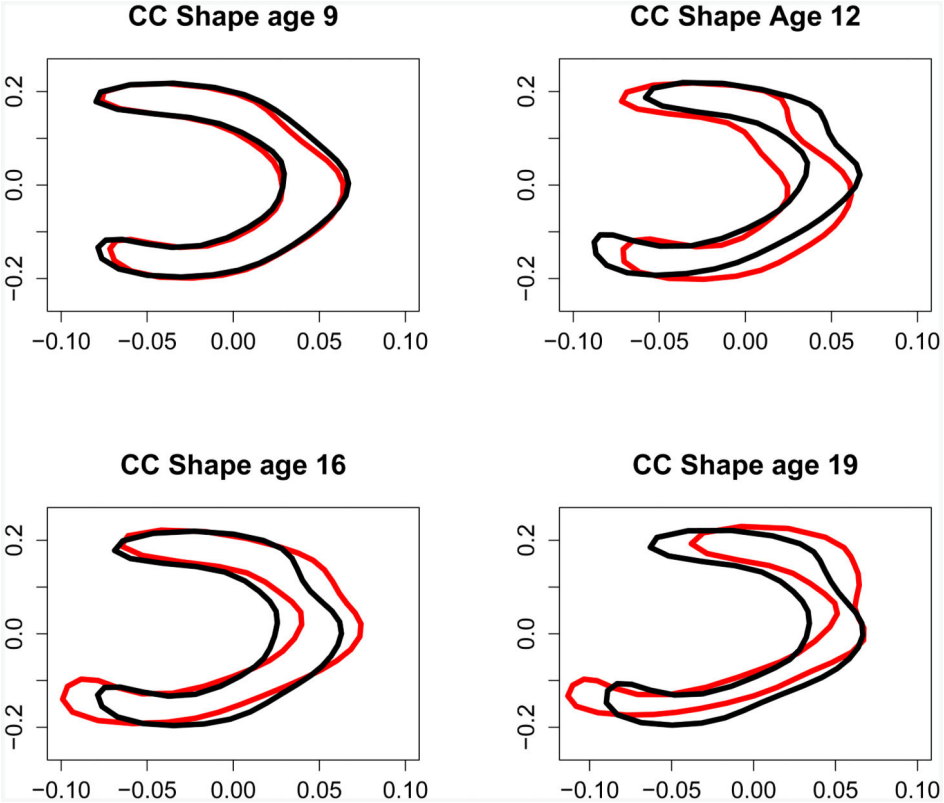

Figure 6.

Predicted CC shape for children ages 9, 12, 16, and 19. The black shape corresponds to typically developing children, while the red shape corresponds to children diagnosed with ADHD. Kernel regression allows us to visualize how CC shape changes through development.

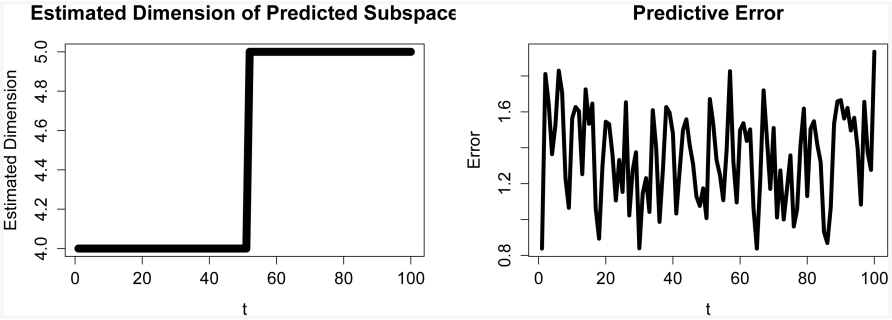

Figure 7.

The estimated dimension and residual for the extrinsic kernel regression estimate at each time point t from data generated from the specified model. The regression estimate is accurate on the dimension of the subspace and prediction residuals are consistent with a concentration parameter $\kappa = 1$.

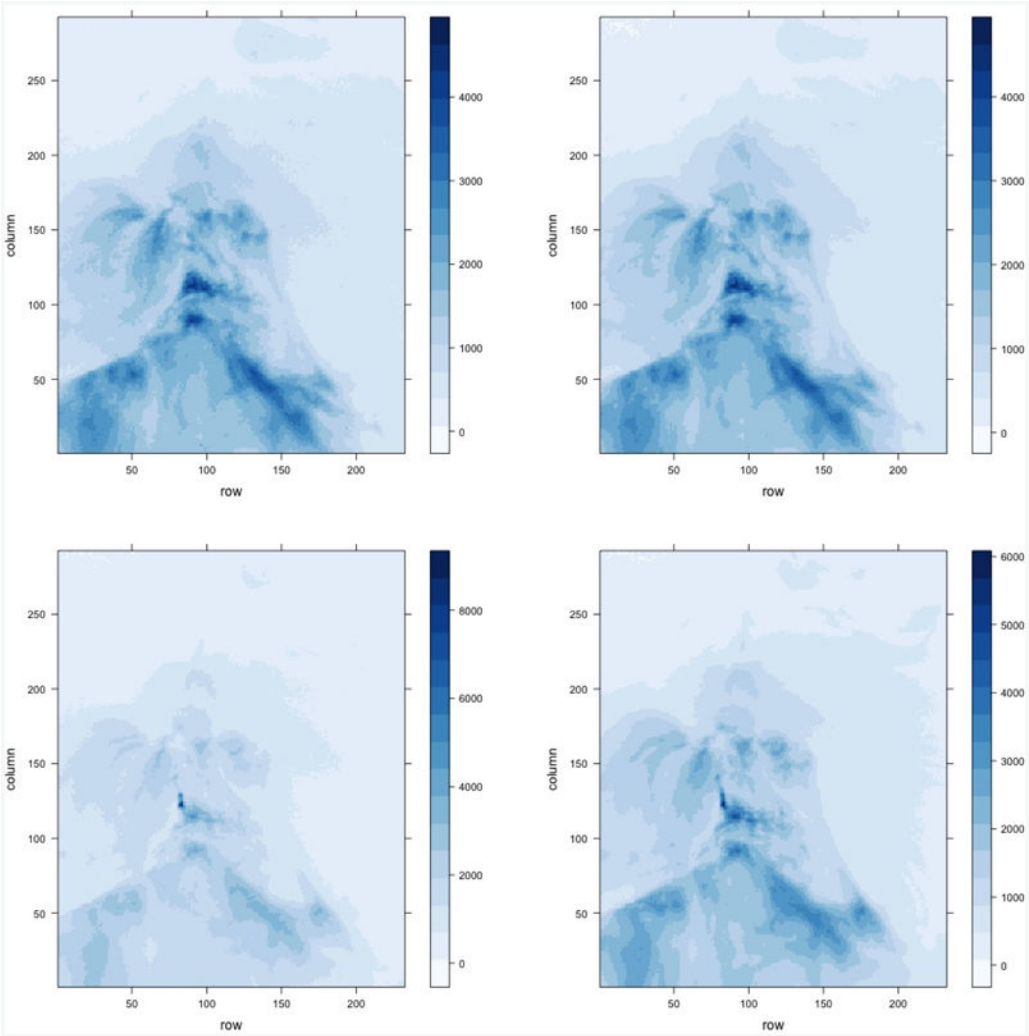

Figure 8.

Pixel representation of the data (left column) and the extrinsic kernel estimates (right column) for the 100th frame of the solar flare video (top row) when the flare is not active, and then the 218th frame of the video (bottom row) when the flare is at it's peak intensity.

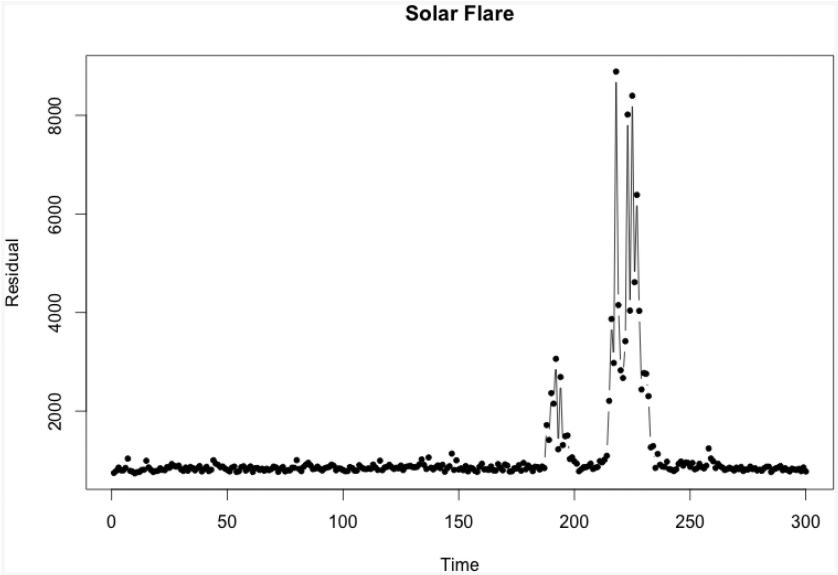

Figure 9.

The residuals of the solar flare images and the extrinsic kernel estimates over time. The spikes indicate solar flare activity.