Statistical Regression Analysis of Shape Data

Mengmeng Guo, Jingyong Su
Department of Mathematics and Statistics, Texas Tech University

Abstract

We develop a multivariate regression model when responses or predictors are on nonlinear manifolds, rather than on Euclidean spaces. The nonlinear constraint makes the problem challenging and needs to be studied carefully. By performing principal component analysis (PCA) on tangent space of manifolds at mean, we use principal directions instead in the model. Then, the ordinary regression tools can be utilized. We apply the framework to shape data (ozone hole contours). Specially, we adopt the square-root velocity representation and parametrization-invariant metric proposed by Srivastava et al. (2011) . Experimental results have shown that we can not only perform efficient regression analysis on the non-Euclidean data, but also achieve high prediction accuracy by the constructed model.

Introduction

❑ Regression analysis of Euclidean data has been studied for decades. The methodologies and tools have been well developed.

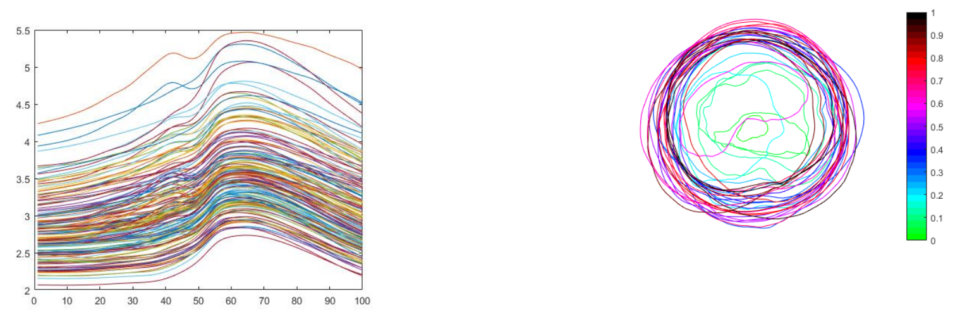
The multiple linear regression model has the form

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + e_i$$

for $i \in \{1, \dots, n\}$ where

- $y_i \in \mathbb{R}$ is the real-valued response for the i-th observation
- $b_0 \in \mathbb{R}$ is the regression intercept
- $b_j \in \mathbb{R}$ is the j-th predictor's regression slope
- $x_{ij} \in \mathbb{R}$ is the j-th predictor for the i-th observation
- $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term

❑ The explosive growth of non-Euclidean data such as functions, curves, surfaces, images and trajectories of those above attracted tremendous attention.



Functional data (left) and curves (right). The left figure displays the absorbance spectrum functions of 215 meat sample from Tecator dataset. The right displays the ozone hole contours of September from 1982 to 2016.

❑ Traditional MLR model will not work because

- The usual Euclidean calculus will not apply
- Conventional methods and statistics may not apply

Our Idea

1. Describe data on manifold using representation proposed by Srivastava, Klassen et al.
2. Perform PCA in tangent space at mean
3. Use principal component scores for regression instead

Our contribution

❑ We develop a multivariate regression model when response variable or predictors are on nonlinear manifolds, rather than on Euclidean space.

❑ We remove phase variability by performing registration to preserve representative statistical summary and capture the variability of data.

Methodology

❑ Mathematics representation of shape and functional data: An absolutely continuous , n-dimensional parameterized curve β, such that β(t) :[0,1] → R^n.

- Diffeomorphisms:

Γ= {$\boldsymbol{\gamma}$: $[\mathbf{0},\mathbf{1}] \to [\mathbf{0},\mathbf{1}]$, $\boldsymbol{\gamma}$(0)=0, $\boldsymbol{\gamma}$(1)=1, $\boldsymbol{\gamma}$ is a diffeomorphism}
$\boldsymbol{\gamma}$ ∈ Γ , Γ be set of all positive to be the set of all positive diffeomorphisms from [0,1] to itself.

- Representation of Curves (the square-root velocity field(SRVF)). Let β(t) :[0,1] → R^n, ∀ t is defined by a function q: [0,1] → R^n as

$$\mathrm{q}(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}} \in R^n$$

- Advantage of SRVF representation:

1. An elastic metric reduces to the L^2 metric
2. Isometry under re-parametrization and rotation for L^2 metric

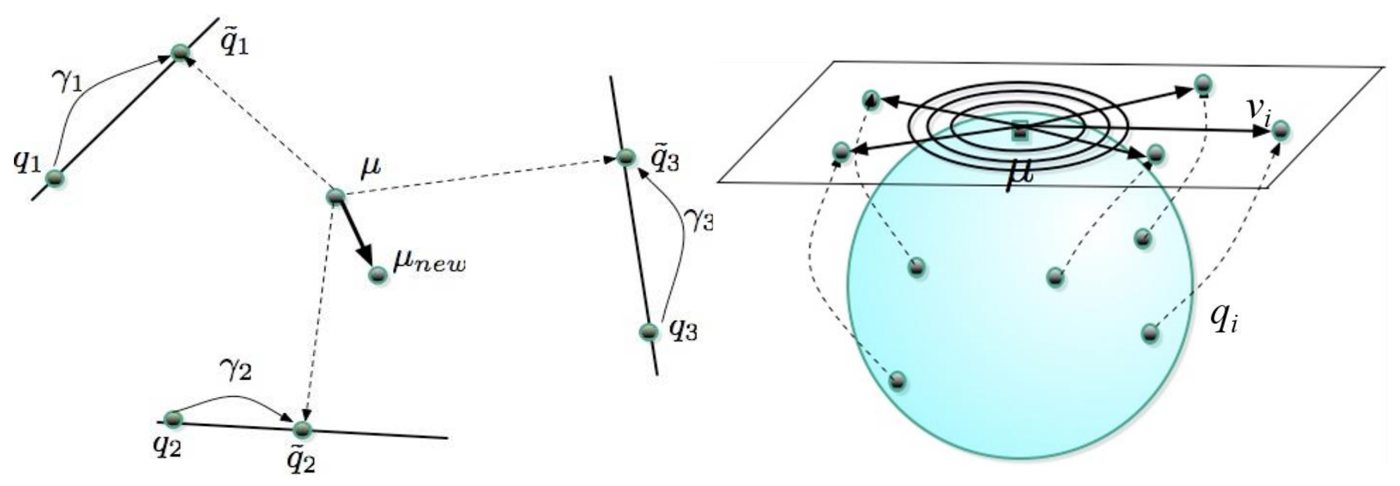
$$\|q_1 - q_2\| = \|O(q_1,\boldsymbol{\gamma}) - O(q_2,\boldsymbol{\gamma})\|$$

3 . The proper metric leads to computation of mean and covariance

4. The tangent space at mean becomes a vector space, where conventional statistics and methods will apply

Statistical Summary: Karcher mean

$$\mu = argmin_{[q]\in S} \sum_{i=1}^{n} d_s([q],[q_i])^2$$



Principal Component Analysis(PCA):

1. Represent smooth curves $y_1(t), \dots, y_n(t)$ with SRVF's $q_1(t), \dots, q_n(t)$
2. Find Karcher mean of all q_i's μ
3. Tangent space T_μ(S) of S at μ, $\tilde{q}_i \in [q]$ has shortest distance to μ, by mapping curve $\tilde{q}_i \to v_i$=$log_\mu(\widetilde{q}_i)$ with inverse exponential map, we can define covariance matrix K= $\frac{1}{n-1}\sum_{i=1}^{n} v_i v_i^t$
3. With PCA, using SVD K =$U \sum U$, estimating the most variation of shooting vectors v_i's, denoted by V_1, V_2. Dominating direction of v_i's at μ can be written as $V = \boldsymbol{c_1} V_1 + \boldsymbol{c_2} V_2$, $\boldsymbol{c_1}$=< v_i , V_1 >, $\boldsymbol{c_2}$=< v_i , V_2 >

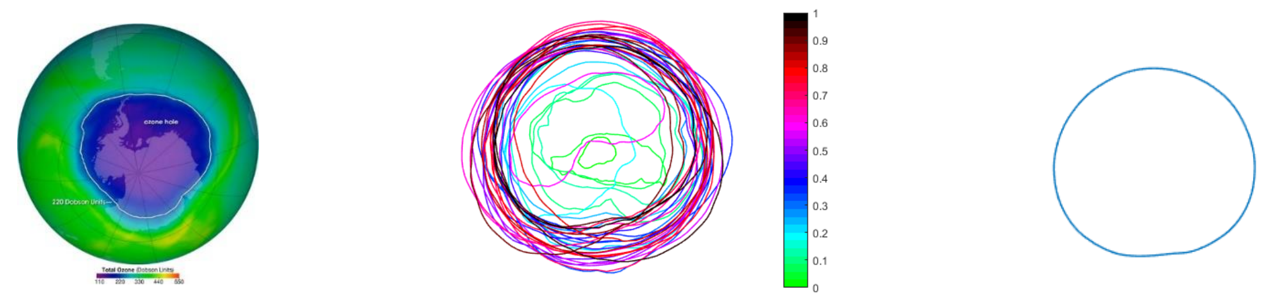
Application

❑ Study of Ozone Hole Contours

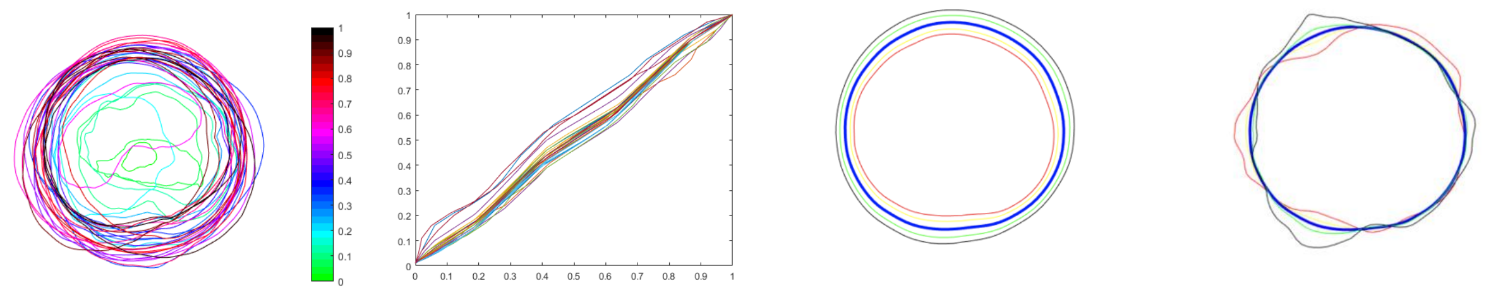
- Ozone is a gas made up of three oxygen atoms O_3. It exists naturally in small amounts in the stratosphere.
- The ozone hole is ozone depletion around earth's south pole area.
- Influencing factor: EESC(Equivalent effective stratospheric chlorine), Antarctic Zonal Wind Speed, Temperature, Heat flux, Solar Radiation, CO_2

Experimental Results

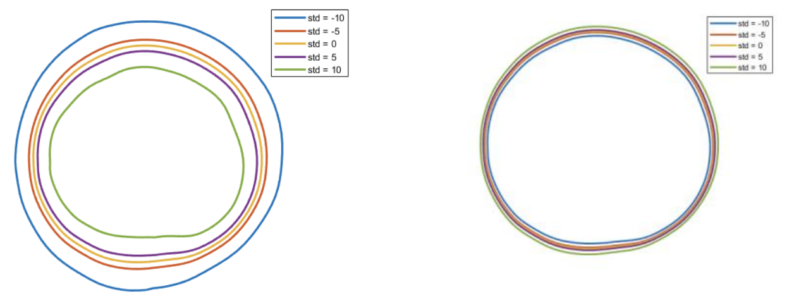
- ozone hole contours and mean contour



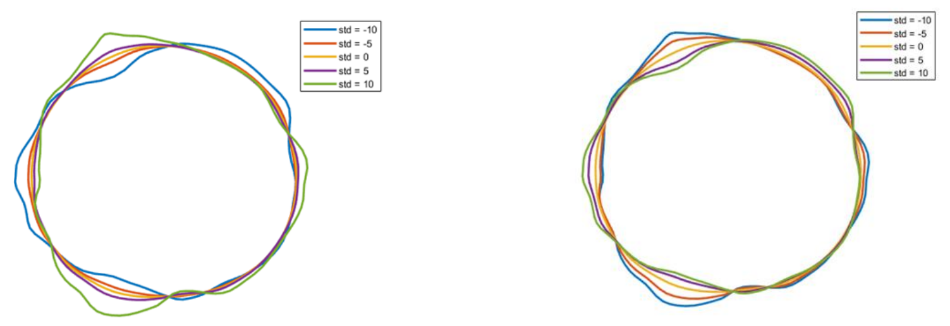
- Registered contour, warping function and modes by PCs



- MLR on first principal component score:
- The p-value is 3.987e-7, R^2=0.8633
- EESC explains 62.3% and zonal wind explains 21.4%



- MLR on second principal component score:
- The p-value is 3.987e-7, R^2=0.9149
- CO_2 explains 28.93%, zonal wind explains 27.84%, solar radiation explains 11.90%



Conclusion

- We studied the regression problem where response variables are functions and curves.
- The representation by Srivastava et al. is utilized and PCA at tangent space of mean is performed.
- The principal component scores are used instead for MLR models.
- This framework can be applied to functional data or high dimensional curves.