NUS National University of Singapore | IMS Institute for Mathematical Sciences

SCIENTIFIC REPORTS

Frontiers of Functional Data Analysis: Challenges and Opportunities in the Era of AI

19 Aug 2024–13 Sep 2024

Organizing Committee

Alexander Aue
University of California at Davis

Ying Chen
National University of Singapore

Zhenhua Lin
National University of Singapore

Qiwei Yao
London School of Economics

CONTENTS PAGE

ROLLED MODELS FOR MANIFOLD-VALUED FUNCTIONAL DATA

KARTHIK BHARATH

Classification AMS 2020: 62R30; 62M20; 62H12

Keywords: Fréchet mean; Gaussian process; Parallel transport

1. INTRODUCTION

Imagine that a curve γ on the unit sphere $\mathbb{S}^2$ drawn in wet ink is rolled along a plane without slipping or twisting so as to trace out a curve $\gamma^{\downarrow}$ in the tangent space of the initial point $\gamma(0)$, identified with $\mathbb{R}^2$. The geometric operation engenders a local isometry between the two curves so that they determine each other uniquely upto isometries, and the operation may be extended to arbitrary d-dimensional connected complete manifolds M. Mathematically, in the intrinsic picture, the Euclidean curve $\gamma^{\downarrow} : [0,1] \to T_{\gamma(0)}M$ on the tangent space of the starting point $\gamma(0)$ is known as the *unrolling* of γ, and is determined by the initial value differential equation

$$\dot{\gamma}^{\downarrow}(t) = P^{\gamma}_{0\leftarrow t}\dot{\gamma}(t), \quad \gamma^{\downarrow}(0) = 0,$$

where $\dot{\gamma}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t) \in T_{\gamma(t)}M$, and $P^{\gamma}_{0\leftarrow t} : T_{\gamma(t)}M \to T_{\gamma(0)}M$ is the parallel transport map along the curve γ, a linear isometry. Choice of coordinates in a tangent space is arbitrary, a frame is hence needed to represent a tangent vector in standard coordinates of $\mathbb{R}^d$. Given an orthonormal frame $U : T_{\gamma(0)}M \to \mathbb{R}^d$, $U\gamma^{\downarrow}(t)$ is then a curve in $\mathbb{R}^d$. In fact, $\gamma^{\downarrow}$ may be defined on the tangent space T_bM of an arbitrary point $b \in M$ outside the cut locus of $\gamma(0)$ by modifying the differential equation as

$$\dot{\gamma}^{\downarrow}(t) = P^{c}_{0\leftarrow 1}P^{\gamma}_{0\leftarrow t}\dot{\gamma}(t), \quad \gamma^{\downarrow}(0) = \exp_b^{-1}(\gamma(0)), \tag{1}$$

where $\exp^{-1}$ is the inverse of Riemannian exponential map $\exp : TM \to M$, and c is the geodesic between b and $\gamma(0)$. Absence of slipping is characterised by use of the parallel transport along γ; twisting is relevant when $\mathbb{S}^2$ is viewed as an embedded submanifold of $\mathbb{R}^3$, where the parallel transport in the above equation is replaced by a curve in $SE(3)$, the isometry group of $\mathbb{R}^3$ [1].

Operationally, from (1), for a fixed point $b \in M$ equipped with a frame U for its tangent space, we can define four maps to: (i) unroll a curve γ in M to obtain a curve $\gamma^{\downarrow}$ in $\mathbb{R}^d$; (ii) peform the reverse operation of *rolling* an $\mathbb{R}^d$-valued curve $\gamma^{\downarrow}$ to obtain γ on M; (iii) use the exponential map at $\gamma(t)$, to *wrap* a curve z in $\mathbb{R}^d$ with respect to γ in M by parallel transporting along curves c and γ the deviation $Uz - \gamma^{\downarrow}$ from the mean; (iv) *unwrap* a curve x on M with respect γ on M by reversing the wrapping operation. Under some conditions, the wrapping and unwrapping maps are also inverses of each other. Rolling/unrolling operations have been used in statistics for curve-fitting using splines, first on $\mathbb{S}^2$ [2] and more recently on general manifolds [3].

The goal of this work is to use the four maps to: define generative statistical models for functional data assuming values in M by pushing forward under the rolling and

wrapping maps a parametric stochastic process model $\{\mathbb{P}_\theta, \theta \in \Theta\}$ for random curves in $\mathbb{R}^d$; given discretely observed M-valued data $\{x_i(t_j)\}$ on a common time grid, estimate θ using unrolling and unwrapping maps. In particular, we will focus on the case where $\{\mathbb{P}_\theta, \theta \in \Theta\}$ corresponds to Gaussian measures parametrized by a mean and positive definite covariance function. Figure 1 provides an illustration.

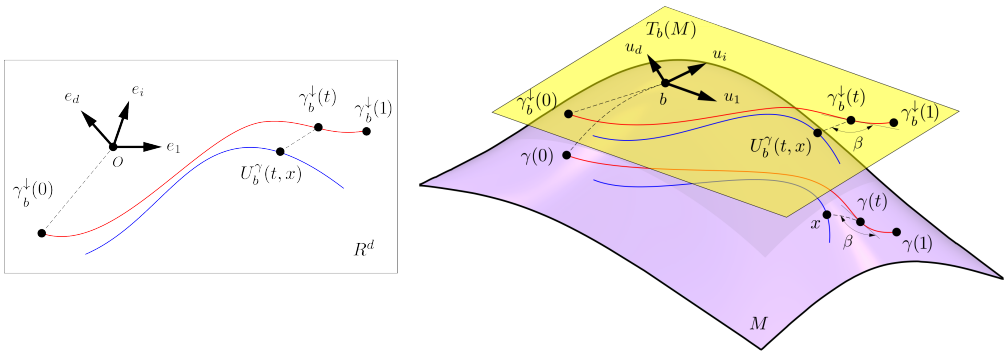


FIGURE 1. Realisation from a Gaussian process in $\mathbb{R}^d$ is mapped to a random curve on M. Red is the mean of the Gaussian procecess, with respect to which the rolling is performed, and blue is the realization from the Gaussian prcoess that deviates from the mean curve. The line segment connecting points between the blue and red curves at arbitrary t, and the corresponding angle denoted β, indicate distances and angles preserved by the (un)rolling.

2. THEORETICAL RESULTS

2.1. **Fréchet mean and rolled mean.** For a random curve $x : [0,1] \to M$ the *population Fréchet mean curve* is defined as the Fréchet mean of x, defined pointwise in t as the minimizer of $s \mapsto E\{\rho^2(x(t), s)\}$, where ρ is the intrinsic distance on M; its sample version is defined by taking expectation with respect to the empirical measure on a sample of curves. The Fréchet mean curve coincides with the rolling of the mean of the $\mathbb{R}^d$-valued process, not necessarily Gaussian, under some conditions.

Theorem 2.1. *For every t, the rolled mean is the Fréchet mean of $x(t)$ if any of the following conditions are true:*

(i) *Every point in M has an empty cut locus;*
(ii) *M is a symmetric manifold, the distribution of $x(t)$ has even symmetry about $\gamma(t)$, and has a unique Fréchet mean that lies outside the cut locus of $\gamma(t)$.*

Condition (ii) concerns the interplay between notions of symmetry of a manifold and that of a probability measure on it. A function $g : M \to \mathbb{R}$ on a symmetric manifold M is said to be symmetric if $g(p) = g(\sigma_p(p))$ for every $p \in M$, where $\sigma_p : M \to M$ is a geodesic-reversing isometry. A distribution ν on M is said to possess *even symmetry* about $p \in M$ if $\nu = (\exp_p)_\# \lambda$, the pushforward under the exponential map at p of a mean-zero distribution λ on T_pM with Lebesgue density f, when T_pM is identified with $\mathbb{R}^d$, satisfies $f(v) = f(-v)$ for every $v \in T_pM$.

2.2. **Rolled Gaussian process on M.** Consider a Gaussian process $t \mapsto z(t) \in \mathbb{R}^d$ with mean function $t \mapsto m(t) \in \mathbb{R}^d$ and covariance kernel $(s,t) \mapsto k(s,t) \in \mathrm{Sym}_{>0}(d)$, where $\mathrm{Sym}_{>0}(d)$ is the cone of positive definite matrices within the vector space of $d \times d$ real symmetric matrices. The wrapping map may be used, with respect to the rolled mean, to transform z to a stochastic process on M, which we refer to as a *rolled Gaussian process*.

Definition 2.2. *Let $z \sim GP(m,k)$, and choose $b \in M$ and frame U of T_bM. With $\tilde{m}_b = Um$ as a curve in T_bM, let $\gamma := \tilde{m}_b^{\uparrow}$ be the rolling of $\tilde{m}$. The process $x = y_b^{\uparrow\gamma}$ obtained by wrapping of $y_b := Uz$ with respect to γ is a rolled Gaussian process, denoted $x \sim \mathcal{RGP}(m,k;b,U)$.*

The point b and a frame U for T_bM are arbitrary, but inconsequential for modelling.

Proposition 2.3. *Starting from point $b \in M$ with frame U for T_bM, let $x \sim \mathcal{RGP}(m,K;b,U)$. If one starts instead from $b' \in M$ with basis U' for $T_{b'}M$, then there exists unique m' and K' such that the rolled Gaussian process $x' \sim \mathcal{RGP}(m',K';b',U')$ is equal in distribution to x.*

For practical purposes, it is convenient to consider a parametric model for the Gaussian process z with respect to a particular basis. Let $\{\phi_s : [0,1] \to \mathbb{R}\}$ be a B-spline basis [4]. Let the mean $m(t) = M_w\phi(t)$, and assume a separable covariance $K(t,t') = \phi(t)^\top V_w \phi(t') U_w$, where $\phi(t) = \{\phi_1(t),\ldots,\phi_k(t)\} \in \mathbb{R}^k$ is a vector and $M_w \in \mathbb{R}^{d\times k}$, $U_w \in \mathrm{Sym}_{>0}(d)$ and $V_w \in \mathrm{Sym}_{>0}(k)$ are matrices that parameterise the model. The convenience of this particular choice is that the curve z can be written

$$z(t) = \sum_{s=1}^{k} w_s \phi_s(t), \tag{2}$$

where $W = (w_1,\ldots,w_k) \sim \mathcal{MN}(M_w,U_w,V_w)$, the matrix normal distribution with mean matrix, M_w, row covariance, U_w, and column covariance, V_w.

2.3. **Estimation.** If $Z \in \mathbb{R}^{d\times r}$ is obtained by observing z in (2) at times $t_1,\ldots,t_r$ then $Z \sim \mathcal{MN}\left(M_w\Phi, U_w, \Phi^\top V_w \Phi\right)$, where $\Phi = \{\phi(t_1),\ldots,\phi(t_r)\} \in \mathbb{R}^{k\times r}$; the distribution of X, the corresponding discretisation of the rolled Gaussian process x is denoted as $X \sim \mathcal{RMN}(M_w,U_w,V_w;b,U)$, and represents the model for the discretely observed sample of curves $\{x_i(t_j)\}$.

Exploiting the relationship between the rolled mean and the Fréchet mean curves in Theorem 2.1, the estimator $\hat{M}_w$ of the mean parameter M_w is defined as follows. Let $H(\hat{\Gamma}) \in \mathbb{R}^{d\times r}$ be the unrolling of the discretised sample Fréchet mean curve onto T_bM followed by a transformation to standard coordinates using the frame U. Define $\hat{M}_w = H(\hat{\Gamma})\Phi^-$, where Φ^- is the right inverse of Φ.

Theorem 2.4. *Let $x \sim \mathcal{RGP}(m,k;b,U)$. Assume that the Fréchet mean curve of x exists and is unique, and suppose that the sample Fréchet mean curve converges in probability to it, as $n \to \infty$, in the C^1 topology. Then, under any of the conditions in Theorem 2.1, as $n \to \infty$, $\hat{M}_w$ converges in probability to M_w.*

The C^1 topology for convergence is needed to ensure that the sequence of parallel transport maps along the sample Fréchet mean curve converge to their limit along the the population Fréchet mean curve. Estimators of covariances U_w and V_w are defined using $\hat{M}_w$ [5], but it is unclear if they are consistent.

3. Robotics application with curves on $SO(3)$

The data are time-indexed orientations of the end-effector of a Franka robot arm as it was guided $n = 60$ times to perform a task to deposit the contents of a dustpan into a bin at $r = 100$ time points. 3D orientations are represented by elements of the rotation group $SO(3)$, which under the unsigned unit quaternion representation, can be identified with $\mathbb{S}^3$ modulo the antipodal map. Fixing the sign of each data point then identifies $SO(3)$ with a hemisphere of $\mathbb{S}^3$.

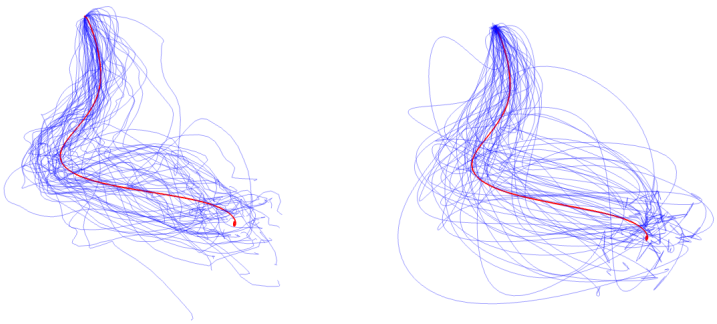


FIGURE 2. Left: Unwrapped SO(3) curves (blue), and unrolled fitted mean (red); Right: simulations from the fitted Gaussian process model;

Left panel of Figure 2 shows the unwrapped curves (blue) in $\mathbb{R}^3$ and the unrolled mean, $H(\hat{\Gamma})$ (red) based on $\hat{M}_w$. The curves have a common starting point at $t = 0$, shown near the top-left in this plot; variability seems to increase with t, especially following a kink point that corresponds to the dustpan being turned to empty its contents. As a visual appraisal of the fitted model, right panel of Figure 2 shows $n = 60$ realisations from $\mathcal{RMN}(\hat{M}_w, \hat{U}_w, \hat{V}_w; b, U)$, using the same unwrapping coordinates and projection as in left panel of Figure 2. These simulated curves are smoother than the real data, which is a consequence of the basis used, but they have similar heteroscedastic variation.

References

[1] R. W. Sharpe (2000). *Differential geometry: Cartan's generalization of Klein's Erlangen program*, Springer.

[2] P. E. Jupp and J. T. Kent (1987). Fitting smooth paths to spherical data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 36, 34-36.

[3] K. R. Kim, I. L. Dryden, H. Le, and K. E. Severn (2021). Smoothing splines on Riemannian manifolds, with applications to 3D shape space *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83, 108-132.

[4] C. DeBoor (2002). *A practical guide to splines*, Springer.

[5] P. Dutilleul (1999). The MLE algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, 64, 105-123.

School of Mathematical Sciences, University of Nottingham
Email address: karthik.bharath@nottingham.ac.uk

AN UNDERSTANDING OF PRINCIPAL DIFFERENTIAL ANALYSIS

GILES HOOKER AND EDWARD GUNNING

Classification AMS 2020: 62G08, 62H12, 62H25, 62M09

Keywords: principal differential analysis, Gaussian process, dimension reduction, ordinary differential equation

1. INTRODUCTION

Classically, functional data analysis focusses on the study of data comprised of univariate functions measured on some continuous domain: $X_1(t), \ldots, X_n(t)$, although this conception has been frequently extended into a more general class of *object data* [3]. One of the distinguishing features of this framework is access to derivatives $D^k X(t)$; presenting both questions about selecting between derivatives as covariates [1, 2] as well as relationships between derivatives.

Principal Differential Analysis (PDA) was proposed in [4] as a means of modeling these relationships. In the classical PDA formulation, an m-th order ODE of the form:

$$D^m X(t) = \beta_0(t)X(t) + \beta_1(t)DX(t) + \cdots + \beta_{m-1}(t)D^{m-1}X(t) + \eta(t) \tag{1.1}$$

is proposed as a model for data. This takes the form of a concurrent linear model described in [5] in which $D^m(X)$ is a response that depends on lower-order derivatives through time-varying functions $\beta(t)$. However, (1.1) also takes the form of a time-varying linear ordinary differential equation (ODE), and PDA leverages this framework in a number of ways.

In particular, PDA is proposed as providing two quantities:

(1) Data reduction: solutions to (1.1) can be expressed as

$$X(t) = \Phi(t, 0)\mathbf{x}_0 + \int \Phi(t, s)\eta(s)ds \tag{1.2}$$

in which $\Phi(s, t)$ is an $(m-1) \times (m-1)$ *transition* matrix at each (s, t) and x_0 is the vector $(X(0), DX(0), \ldots, D^{m-1}X(0))$. Here we regard $\Phi(t, 0)$ as providing a basis expansion in which to represent $X(t)$, providing a data-reduction method akin to functional principal components analysis.

(2) Representation of behaviour. For time-invariant ODE's in which the $\beta(t)$ are constant with $\eta(t) = 0$. Solutions to (1.1) can be expressed in terms of

$$X(t) = \sum_{k=1}^{m-1} c_k e^{b_k t} \left(\cos(d_k t) + i \sin(d_k t)\right) \tag{1.3}$$

in which (b_k, c_k, d_k) are obtained from the eigen decomposition of a matrix representation of the dynamics. Under this framework, the instantaneous behaviour of $X(t)$ can be interpreted by considering the decomposition in (1.3) at the current values of β.

We revisit the PDA model, re-interpreting it as a description of a data-generating process in which $\eta(t)$ is a random error process specific to each observations. This view yields a number of a consequences, most specifically a bias in the classical PDA estimates which we correct with an iterative procedure.

2. Consequences of a Generative Model

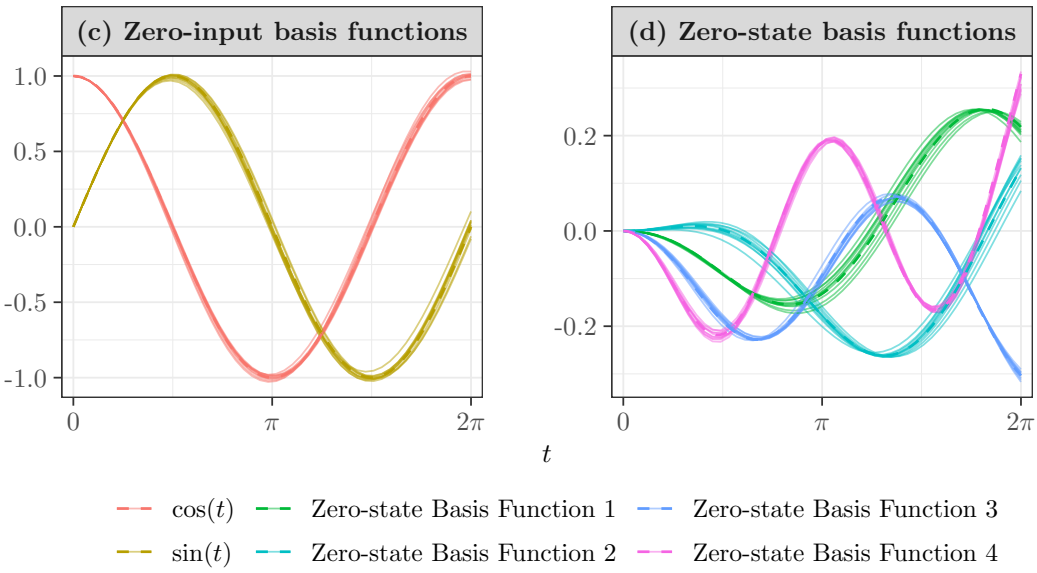
This presentation regards (1.1) as a generating model, we assume that $\eta(t) \sim (0, \Sigma)$ is a mean zero random error process with covariance $\Sigma(s,t)$. We first observe that this framework augments the dimension reduction approach in [4] with a second term derived from (1.2):

$$\text{cov}(\tilde{\mathbf{x}}(s), \tilde{\mathbf{x}}(t)) = \Phi(s,0)\mathbf{\Sigma}_0\Phi(t,0)^\top + \int_0^s \int_0^t \Phi(s,u)\Sigma(u,v)\Phi(t,v)^\top \mathrm{d}v\mathrm{d}u$$

yielding a new decomposition. We illustrate this below based on simulating data from simple harmonic motion forced by Gaussian process noise:

$$D^2X(t) = -X(t) + \eta(t)$$

illustrated below in which the left hand plot provides simulation estimates of variation associated with initial conditions $\mathbf{x}_0$ and the right-hand plot variance associated with the random process $\eta(\cdot)$.



A second consequence involves biases in the estimated $\hat{\beta}(t)$. The original PDA formulation minimizes the integrated sum of squared errors from (1.1):

$$\sum \int \left(D^m X_i(t) - \beta_0 X_i(t) - \ldots - \beta_{m-1}(t) D^{m-1} X_i(t)\right)^2 dt$$

which can be solved by a linear regression at each time point t. Writing $Z(t)$ as the matrix containing rows $(X(t), DX(t), \ldots, D^{m-1}X(t))$ we obtain an estimate

$$\hat{\beta}(t) = (Z(t)^T Z(T))^{-1} Z(t)^T D^{m-1} X(t) \tag{2.1}$$

although [4] represented each $\beta_j(t)$ via a basis expansion allowing the use of smoothing penalties.

In classical linear regression (2.1) substituting $D^{m-1}X(t) = Z(t)\beta(t) + \eta(t)$ results in unbiassed estimates. Here, however, we observe from (1.2) that $EZ(t)\eta(t) = \int_{s=0}^{t} \Phi(s,t)_{m-1,\cdot}\Sigma(s,t)ds$ and an approximation to bias in $\hat{\beta}$:

$$E\hat{\beta}(t) = \beta(t) + E(Z(t)^T Z(T))^{-1} EZ(t) \int_{s=0}^{t} \Phi(s,t)_{m-1,\cdot}\Sigma(s,t)ds.$$

Within the bias term, both $\Phi(s,t)$ and Σ require estimates for $\beta(t)$, resulting in the following iterative

(1) **Initialize** Begin with ordinary least squares (OLS) estimates for the parameters $\beta_0(t), \ldots, \beta_{m-1}(t)$ by minimizing the ISSE and obtain residuals $\hat{\eta}(t)$ and covariance

$$\hat{\Sigma}(s,t) = \frac{1}{n-m}\sum \hat{\eta}_i(s)\hat{\eta}_i(t)$$

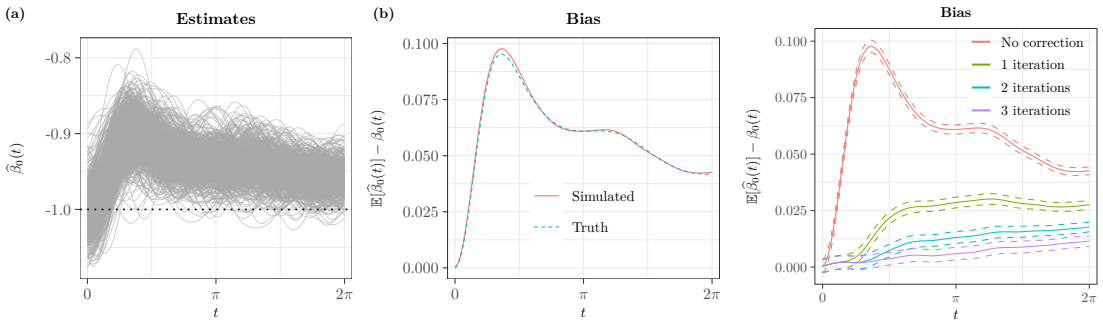
along with the transition matrices $\Phi(s,t)$.

(2) **Iterative Correction:** Apply the bias-reduction formula iteratively:

$$\hat{\beta}_j^{(k+1)}(t) = \hat{\beta}_j^{(k)}(t) - E(Z(t)^T Z(T))^{-1} EZ(t) \int_{s=0}^{t} \Phi(s,t)_{m-1,\cdot}\Sigma(s,t)ds \tag{2.2}$$

After each iteration, the residuals $\hat{\eta}_i^{(k)}(t)$ are updated along with $\hat{\Sigma}$ and Φ until convergence.

We illustrate this with the same simple harmonic motion example below in which we first provide an estimate of the bias for the coefficient of $X(t)$ and estimates after three steps of bias reduction



3. Conclusion

This work reframes PDA as a statistical model that accounts for both deterministic ODE dynamics and stochastic disturbances. This has consequences for both the estimation of functional coefficients and presenting a variance decomposition from the resulting model. We will further illustrate the application of these methods to provide linear approximations to non-linear ODE's and some consequences for approaches to registration.

REFERENCES

[1] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York, 2006.

[2] Giles Hooker and Han Lin Shang. Selecting the derivative of a functional covariate in scalar-on-function regression. *Statistics and Computing*, 32(3):35, 2022.

[3] J Steve Marron and Andrés M Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753, 2014.

[4] James O. Ramsay. Principal Differential Analysis: Data Reduction by Differential Operators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(3):495–508, 1996. Publisher: [Royal Statistical Society, Wiley].

[5] James O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition, 2005.

DEPARTMENT OF STATISTICS AND DATA SCIENCE, UNIVERSITY OF PENNSYLVANIA
Email address: ghooker@wharton.upenn.edu

NEURAL TANGENT KERNEL IN IMPLIED VOLATILITY FORECASTING: A NONLINEAR FUNCTIONAL AUTOREGRESSION APPROACH

HANNAH L. H. LAI

Classification AMS 2020: 46T99, 68T01

Keywords: Nonlinear Functional Autoregression; Neural Tangent Kernel; Implied Volatility Forecasting.

We denote by $\mathcal{H} = L^2(\mathcal{I})$ the Hilbert space consisting of all square-integrable surfaces defined on a compact set $\mathcal{I} \subset \mathbb{R}^q$ and equipped with the inner product $\langle f, g\rangle_{\mathcal{H}} = \int_{\mathcal{I}} f(u)g(u)\,du$, for any $f, g \in L^2(\mathcal{I})$. Define the squared L^2 norm of a function by $\|f\|_{\mathcal{H}} = \langle f, f\rangle_{\mathcal{H}}$.

Let $\{Y_i\}_{i=1}^n$ be a series of n random surfaces that take values on $\mathcal{H}_Y = L^2(\mathcal{I}_Y)$. Associated with each Y_i, there is a regressor surface $X_i \in \mathcal{H}_X = L^2(\mathcal{I}_X)$. We consider functions with finite second moment, i.e., $\mathbb{E}[||Y_i||^2_{\mathcal{H}_Y}] < \infty$ and $\mathbb{E}[\|X_i\|^2_{\mathcal{H}_X}] < \infty$. For simplicity, we assume that Y_i and X_i are centered functions, i.e., $\mu_X(v) = \mathbb{E}[X_i(v)] = 0, \forall v \in \mathcal{I}_X$ and $\mu_Y(u) = \mathbb{E}[Y_i(u)] = 0, \forall u \in \mathcal{I}_Y$. Let P_X and P_Y denote the distributions of X and Y, and $P_{Y|X} : \mathcal{H}_X \times \mathcal{H}_Y \to \mathbb{R}$ the conditional distribution of Y given X. If $L_2(P_X)$ represents the class of all measurable functions of X with $\mathbb{E}[f^2(X)] < \infty$ under P_X, then $L_2(P_Y)$ is similarly defined for Y. Our goal is to capture the potential nonlinear dependence between Y_i and X_i through a function $g : \mathcal{H}_X \to \mathcal{H}_Y$

$$Y_i = g(X_i) + \epsilon_i, \tag{0.1}$$

where ϵ_i is a noise function with $\mathbb{E}[\epsilon_i(u)] = 0, \forall u \in \mathcal{I}_Y$ and $\mathbb{E}[\|\epsilon_i\|^2_{\mathcal{H}_Y}] < \infty$. In our study, X_i is a vector of lagged surfaces $Y_{i-1}, Y_{i-2}, \ldots$ or their linear combination. Hence, the model (0.1) is a nonlinear functional autoregression model (NFAR).

We project Y_i onto a set of orthonormal basis functions $\varphi = (\varphi_1, \varphi_2 \ldots)^T$ with $\varphi_j \in \mathcal{H}_Y$

$$Y_i = \sum_{j=1}^{\infty} y_{ij}\varphi_j, \quad \text{with } y_{ij} = \langle Y_i, \varphi_j\rangle_{\mathcal{H}_Y}, \tag{0.2}$$

with $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ...)^T \in \mathcal{H}_{\boldsymbol{y}} \subseteq \mathbb{R}^\infty$ the projection coefficients of Y_i onto the basis functions φ, satisfying $\mathbb{E}[y_{ij}y_{rv}] = 0$ for $j \neq v$, $j, v \in \mathbb{N}_+$ and any $i, r \in \{1, \ldots, n\}$. Similarly, we project X_i onto a sequence of orthogonal basis functions $\psi = (\psi_{1,}, \psi_2, ...)^T$ with $\psi_j \in \mathcal{H}_X$

$$X_i = \sum_{j=1}^{\infty} x_{ij}\psi_j, \quad \text{with } x_{ij} = \langle X_i, \psi_j\rangle_{\mathcal{H}_X}, \tag{0.3}$$

with $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ...)^T \in \mathcal{H}_{\boldsymbol{x}} \subseteq \mathbb{R}^\infty$ the projection coefficients of X_i onto the basis functions ψ, satisfying $\mathbb{E}[x_{ij}x_{rv}] = 0$ for $j \neq v$, $j, v \in \mathbb{N}_+$ and any $i, r \in \{1, \ldots, n\}$. Transitioning from functions to vectors, we define $f : \mathcal{H}_x \to \mathcal{H}_y$

$$\boldsymbol{y}_i = f(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i, \tag{0.4}$$

where $\boldsymbol{\epsilon}_i$ is a noise vector with $\mathbb{E}[\boldsymbol{\epsilon}_{ij}] = 0$ and $\mathbb{E}[\|\boldsymbol{\epsilon}_i\|^2] < \infty$. Although vectors offer a more compact representation of functions, they still exist within an infinite-dimensional framework unless additional restrictions are assumed to hold. This inherent complexity makes the empirical estimation of Equation (0.4) challenging when working with finite sample sizes. To address this issue, we employ classical sieve methods leading to finite-dimensional vector spaces. [1]

To elucidate the nonlinear relation between X_i and Y_i in Equation (0.1), we introduce another Hilbert space of functions generated by a positive-definite kernel $K : \mathcal{H}_X \times \mathcal{H}_X \to \mathbb{R}$ defined on the inner product of $\mathcal{H}_X$ through a function $\rho : \mathbb{R}^3 \to \mathbb{R}^+$, such that

$$K(X_i, X_j) = \rho(\langle X_i, X_i\rangle_{\mathcal{H}_X}, \langle X_i, X_j\rangle_{\mathcal{H}_X}, \langle X_j, X_j\rangle_{\mathcal{H}_X}), \tag{0.5}$$

for any $X_i, X_j \in \mathcal{H}_X$. The function-on-function regression problem in Equation (0.1) can be reformulated as a functional kernel regression, in which the task is to find $B_0 \in \mathcal{B}(\mathcal{H}_Y, \mathfrak{M}_X)$ such that

$$B_0 = \underset{B\in\mathcal{B}(\mathcal{H}_Y,\mathfrak{M}_X)}{\arg\min}\ \mathbb{E}[\|Y_i - B^*K(., X_i)\|^2_{\mathcal{H}_Y}]. \tag{0.6}$$

The solution for the kernel functional regression can be found in [2]. We define a new kernel $k : \mathcal{H}_{\boldsymbol{x}} \times \mathcal{H}_{\boldsymbol{x}} \to \mathbb{R}$ such that for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{H}_{\boldsymbol{x}}$

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \rho(\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \langle \boldsymbol{x}_j, \boldsymbol{x}_j\rangle). \tag{0.7}$$

Lemma 0.1 (Isomorphism between Reproducing Kernel Hilbert Spaces). *Under Equations* (0.3) *and* (0.7)*, it holds that*

$$\begin{aligned} k(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \langle k(., \boldsymbol{x}_i), k(., \boldsymbol{x}_j)\rangle \\ &= \langle K(., X_i), K(., X_j)\rangle_{\mathfrak{M}_X} = K(X_i, X_j). \end{aligned} \tag{0.8}$$

Then the RKHS $\mathfrak{M}_X$ nested on $\mathcal{H}_X$ is isometrically isomorphic to the RKHS $\mathfrak{M}_{\boldsymbol{x}}$ nested on $\mathcal{H}_{\boldsymbol{x}}$.

Theorem 0.2 (Vector-to-vector regression). *Given the decomposition of Y_i in Equation* (0.2) *and X_i in Equations* (0.3)*, under some technical Assumptions and Lemma 0.1, for a positive definite kernel k defined by Equation* (0.7)*, if there is a covariance matrix $\Sigma_{\boldsymbol{xx}}$ of $k(., \boldsymbol{x})$ that is diagonal, then the function-to-function regression model in Equation* (0.6) *may be represented equivalently by*

$$\beta_0 = \underset{\beta\in\mathcal{B}(\mathcal{H}_{\boldsymbol{y}},\mathfrak{M}_{\boldsymbol{x}})}{\arg\min}\ \mathbb{E}[\|\boldsymbol{y}_i - \beta^*k(., \boldsymbol{x}_i)\|^2], \tag{0.9}$$

with solution $\beta_0 = \Sigma^\dagger_{\boldsymbol{xx}}\Sigma_{\boldsymbol{xy}}$. This leads to

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_i|\boldsymbol{x}_i] &= \beta_0^*k(., \boldsymbol{x}_i) \\ &= \Sigma_{\boldsymbol{yx}}\Sigma^\dagger_{\boldsymbol{xx}}k(., \boldsymbol{x}_i) \\ &= \mathbb{E}[\{(\Sigma^\dagger_{\boldsymbol{xx}}k(., \boldsymbol{x}_i))(\boldsymbol{x})\}\boldsymbol{y}]. \end{aligned} \tag{0.10}$$

[1] Sieve methods involve truncating the regression for the full set of projection coefficients while striving to minimize any loss of information.

In our work, we utilize the Neural Tangent Kernel (NTK) of [1], a flexible kernel class that uses neural networks to capture complex nonlinear dependencies in data effectively. The NTK describes how neural networks behave under first-order gradient descent training and is calculated as the inner product of the network's weight gradients. Our empirical analysis includes over 6 million European calls and put options from the S&P 500 Index, covering January 2009 to December 2021. The results confirm the superior forecasting accuracy of the fNTK across different time horizons. When applied to short delta-neutral straddle trading, the fNTK achieves a Sharpe ratio ranging from 1.30 to 1.83 on a weekly to monthly basis, translating to 90% to 675% relative improvement in portfolio returns compared to forecasts based on functional Random Walk model.

References

[1] Jacot, A., F. Gabriel, and C. Hongler . Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[2] Sang, P. and B. L. Nonlinear Function-on-Function Regression by RKHS. *arXiv preprint arXiv:2207.08211*, 2022.

Email address: hlhlai@u.nus.edu

ROBUST MODEL AVERAGING PREDICTION

JIALIANG LI

Classification AMS 2020:

Keywords: Longitudinal data analysis, Microbiome data analysis, Model averaging, Rank regression, Robust estimation, Sure screening

Model averaging is an attractive ensemble technique to construct fast and accurate prediction. Despite of having been widely practiced in cross-sectional data analysis, its application to longitudinal data is rather limited so far. We consider model averaging for longitudinal response when the number of covariates is ultrahigh. To this end, we propose a novel two-stage procedure in which variable screening is first conducted and then followed by model averaging. In both stages, a robust rank-based estimation function is introduced to cope with potential outliers and heavy-tailed error distributions, while the longitudinal correlation is modeled by a modified Cholesky decomposition method and properly incorporated to achieve efficiency. Asymptotic properties of our proposed methods are rigorously established, including screening consistency and convergence of the model averaging estimates. Extensive simulation studies demonstrate that our method outperforms existing competitors, resulting in significant improvements in screening and prediction performance. Finally, we apply our proposed framework to analyze a human microbiome dataset, showing the capability of our procedure in resolving robust prediction using massive metabolites.

DEPARTMENT OF STATISTICS & DATA SCIENCE, NATIONAL UNIVERSITY OF SINGAPORE, SINGAPORE
Email address: jialiang@nus.edu.sg

Functional Principal Component Analysis For Distribution-Valued Processes

Hang Zhou and Hans-Georg Müller

Keywords: Distributional Data Analysis, Functional Data Analysis, Longitudinal Data Analysis, Sparse Designs, Stochastic Process, Wasserstein Metric

Functional data are samples of realizations of square integrable scalar or vector-valued functions that have been extensively studied (Hsing & Eubank 2015, Wang et al. 2016). The restriction to the realm of Euclidean space-valued functions that also encompasses Hilbert-space valued functional data, i.e., function-valued stochastic processes (Chen et al. 2017), is an essential feature of functional data, but proves too restrictive as new complex non-Euclidean data types are emerging. A previous very general model for the case of a metric-space valued process for which one observes a sample of realizations (Dubey & Müller 2020) includes distribution-valued processes as a special case. The general framework developed in this previous approach utilizes a notion of metric covariance and includes a certain kind of functional principal component analysis for general metric space-valued processes by using Fréchet integrals (Petersen & Müller 2016) and is limited to the case of fully observed metric space-valued functional data, where it is assumed that $X_i(t)$ is known for all t in the time domain and cannot be extended to the case of sparsely sampled processes. We present models and analysis tools for a specific yet important class of random object-valued stochastic processes: those where the time-indexed objects are univariate distributions. Distribution-valued stochastic processes are encountered in various complex applications. We start with an i.i.d. sample of realizations of such processes. The statistical modeling of distribution-valued processes is an essential yet still missing tool for the emerging field of distributional data analysis (Petersen et al. 2022), while various modeling approaches for distributional regression and distributional time series have been studied recently (Kokoszka et al. 2019, Ghodrati & Panaretos 2022, Chen et al. 2023, Zhu & Müller 2023).

We aim for intrinsic modeling of distributions rather than extrinsic approaches. An issue that is of additional practical relevance and theoretical interest is that available observations typically are not available continuously in time but only at discrete time points. These considerations motivate a comprehensive intrinsic model for distribution-valued processes where the processes may be fully or only partially observed. Throughout we work with the 2-Wasserstein metric $d_{W,2}$ and optimal transports, which move distributions along geodesics. The challenge of intrinsic modeling is that the Wasserstein space of distributions does not have a linear or vector space structure. This challenge can be addressed by making use of rudimentary algebraic operations on the space of optimal transports (Zhu & Müller 2023). From the outset we aim to deal with centered processes. Since no subtraction exists in the Wasserstein space, the centering of distribution-valued processes is achieved by substituting transport processes for distributional processes: For each time argument the distributions that constitute the values of a distributional process at a fixed time t are replaced by

transports from the barycenter (Fréchet mean) of the process at t to the distribution that corresponds to the value of the process at time t. These transports are well defined if one adopts the Wasserstein metric. Their Fréchet mean is the identity transport, i.e., these transports are centered.

For our study of stochastic transport processes we introduce representations

$$T(t) = g(Z(t)) \odot T_0,$$

where $Z(t)$ is a $\mathbb{R}$-valued random process, g is a bijective function that maps $\mathbb{R}$ to $(-1, 1)$ and T_0 is a single random transport that is a summary characteristic for each realization of the transport process. Here $\odot$ is a multiplication operation by which a transport is multiplied with a scalar (Zhu & Müller 2023). By construction, $g(Z(t)) \odot T_0$ lies on the extended geodesic that passes through T_0. We develop a predictor for each individual $T_i(t)$ based on observations obtained at discrete time points and establish asymptotic convergence rates for the components of the model for both densely and sparsely sampled distributional processes. These are novel even for classical real-valued functional data.

Let $\mathcal{W}$ be the set of finite second moment probability measures on the closed interval $\mathcal{S} \subset \mathbb{R}$,

$$\mathcal{W} = \left\{ \mu \in \mathcal{P}(\mathcal{S}) : \int_{\mathcal{S}} |x|^2 \mathrm{d}\mu(x) < \infty \right\}, \tag{1}$$

where $\mathcal{P}(\mathcal{S})$ is the set of all probability measures on $\mathcal{S}$. The p-Wasserstein distance $d_{W,p}(\cdot, \cdot)$ between two measures $\mu, \nu \in \mathcal{W}$ is

$$d_{W,p}(\mu, \nu) := \inf \left\{ \left(\int_{\mathcal{S}^2} |x_1 - x_2|^p \mathrm{d}\Gamma(x_1, x_2) \right)^{1/p} : \Gamma \in \Gamma(\mu, \nu) \right\} \quad \text{for } p > 0, \tag{2}$$

where $\Gamma(\mu, \nu)$ is the set of joint probability measures on $\mathcal{S}^2$ with μ and ν as marginal measures. The Wasserstein space $(\mathcal{W}, d_{W,p})$ is a separable and complete metric space (Ambrosio et al. 2008, Villani et al. 2009). Here we assume $\mathcal{S} = [0, 1]$ without loss of generality to simply the notation. Given two probability measures $\mu, \nu \in \mathcal{W}$, the optimal transport from μ to ν is the map $T : \mathcal{S} \to \mathcal{S}$ that minimizes the transport cost,

$$\underset{T \in \mathcal{T}}{\arg\inf} \left\{ \left(\int_{\mathcal{S}} |T(u) - u|^p \mathrm{d}\mu(u) \right)^{1/p}, \text{ such that } T\#\mu = \nu \right\}, \tag{3}$$

where $\mathcal{T} = \{T : \mathcal{S} \mapsto \mathcal{S} | T(0) = 0, T(1) = 1, T \text{ is non-decreasing}\}$ is the transport space and $T\#\mu$ is the push-forward measure of μ, defined as $(T\#\mu)(A) = \mu\{x \in \mathcal{S} \mid T(x) \in A\}$ for all A in the Borel algebra of $\mathcal{S}$. This optimization problem, also known as the Monge problem, is a relaxation of the Kantorovich problem (2). If μ is absolutely continuous with respect to the Lebesgue measure, then problems (2) and (3) are equivalent and have a unique solution $T(u) = F_\nu^{-1} \circ F_\mu(u)$ for $p = 2$, where F_μ and F_ν^{-1} are the cumulative distribution and quantile functions of μ and ν, respectively (Gangbo & McCann 1996).

For a distribution-valued process $X(t)$ with random distributions on domain $\mathcal{S}$ where $t \in \mathcal{D}$ for a closed interval in $\mathbb{R}$, the cross-sectional Fréchet mean of $X(t)$ at each t is

$$\mu_{\oplus,2}(t) = \text{argmin}_{\omega \in \mathcal{W}} \mathbb{E} d^2_{W,2}(X(t), \omega).$$

We then define the (optimal) transport process $T(\cdot)$, where $T(t)$ represents the optimal transport from $\mu_{\oplus,2}(t)$ to $X(t)$, $\mu_{\oplus,2}(t)$ serves as the mean, and the transport $T(t)$ from $\mu_{\oplus,2}(t)$ to $X(t)$ quantifies the difference between $X(t)$ and $\mu_{\oplus,2}(t)$ for each $t \in \mathcal{D}$ under the Wasserstein metric. It is thus advantageous to use the transport space $\mathcal{T}$.

A scalar multiplication operation in the transport space (Zhu & Müller 2023),

$$\alpha \odot T(u) := \begin{cases} u + \alpha\{T(u) - u\}, & 0 < \alpha \leq 1 \\ u, & \alpha = 0 \\ u + \alpha\left\{u - T^{-1}(u)\right\}, & -1 \leq \alpha < 0 \end{cases}$$

induces a geodesic on $\mathcal{T}$ from $\text{Unif}_{\mathcal{S}}$ to T, denoted by $u \odot T$ for all $u \in [-1, 1]$. We introduce a binary relation $\sim$ on $\mathcal{T}$, defined as $T_1 \sim T_2$ if and only if there exists $a \in [0, 1]$ such that $T_1 = a \odot T_2$ or $T_2 = a \odot T_1$ and demonstrate that $\sim$ is an equivalence relation on $\mathcal{T}$.

In analogy to the decomposition of Euclidean-valued functional data into a mean function and a stochastic part, we assume that the centered transport processes $T(t)$ can be decomposed into a scalar random function $U(t)$ that serves as a scalar multiplier in the transport space and a characteristic overall transport T_0,

$$T(t) = U(t) \odot T_0, \text{ for all } t \in \mathcal{D}, \tag{4}$$

where T_0 is a random element in $\mathcal{T}$ associated with each realization of the transport process. The scalar multiplier function is a stochastic process that takes values in $(-1, 1)$ and is derived from an unconstrained process Z through a transformation g as follows,

$$U(t) = g(Z(t)),\ Z(t) \in \mathbb{R},\ \mathbb{E}[Z(t)] = 0,\ g : \mathbb{R} \mapsto (-1, 1),\ g \text{ is bijective, for all } t \in \mathcal{D}. \tag{5}$$

The mean zero stochastic process $Z(t)$ in conjunction with the bijective map $g : \mathbb{R} \mapsto (-1, 1)$ further characterizes the transport process T, where $T(t)$ resides in $\{T : T \in [T_0]_\sim\} \cup \{T : T \in [T_0^{-1}]_\sim\}$, which includes the geodesic from T_0^{-1} to T_0.

For some situations it is appropriate and advantageous to further assume that the process Z is a Gaussian process, a property that can be harnessed to obtain methods for the important case where the distribution-valued trajectories are only observed on a discrete grid of time points that might be sparse. that the stochastic transport process (4) is well-defined.

Further details can be found in the preprint: Zhou, H. and Müller, H.G., 2023. Optimal transport representations and functional principal components for distribution-valued processes. arXiv preprint arXiv:2310.20088.

References

Ambrosio, L., Gigli, N. & Savaré, G. (2008), *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*, Springer Science & Business Media.

Chen, K., Delicado, P. & Müller, H.-G. (2017), 'Modeling function-valued stochastic processes, with applications to fertility dynamics', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **79**, 177–196.

Chen, Y., Lin, Z. & Müller, H.-G. (2023), 'Wasserstein regression', *J. Amer. Statist. Assoc.* **118**(542), 869–882.

Dubey, P. & Müller, H.-G. (2020), 'Functional models for time-varying random objects', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**(2), 275–327.

Gangbo, W. & McCann, R. J. (1996), 'The geometry of optimal transportation', *Acta Math.* **177**, 113–161.

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

Hsing, T. & Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, John Wiley & Sons.

Kokoszka, P., Miao, H., Petersen, A. & Shang, H. L. (2019), 'Forecasting of density functions with an application to cross-sectional and intraday returns', *Int. J. Forecast.* **35**(4), 1304–1317.

Petersen, A. & Müller, H.-G. (2016), 'Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data', *Biometrika* **103**(1), 103–120.

Petersen, A., Zhang, C. & Kokoszka, P. (2022), 'Modeling probability density functions as data objects', *Econom. Stat.* **21**, 159–178.

Villani, C. et al. (2009), *Optimal Transport: Old and New*, Springer.

Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016), 'Functional data analysis', *Annu. Rev. Stat. Appl.* **3**, 257–295.

Zhu, C. & Müller, H.-G. (2023), 'Autoregressive optimal transport models', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **85**(3), 1012–1033.

HIGH-DIMENSIONAL HILBERT-SCHMIDT LINEAR REGRESSION FOR HILBERT MANIFOLD VARIABLES

BYEONG UK PARK

Classification AMS 2020: 62R30, 62J07.

Keywords: non-Euclidean data, high-dimensional regression, Hilbert-Schmidt operators, spectral decomposition, penalization.

We present a unified framework for high-dimensional linear regression with non-Euclidean data where both the response variable and covariates take values in general Riemannian Hilbert manifolds. In our modeling, the response and covariates are allowed to orginate from distinct spaces and are interconnected by Hilbert-Schmidt operators. The methodology is developed under a general penalization scheme incorporating various non-convex penalty functions, thereby accommodating scenarios where the number of covariates grows exponentially with the sample size. Leveraging modern statistical theory for data residing on Hilbert manifolds, we establish the oracle property and derive error bounds for the proposed estimators. The practical validity of the proposed method is demonstrated via numerical simulation and real data applications.

Specifically, let Y and X_j for $1 \leq j \leq p$ be random variables taking values in Riemannian Hilbert manifolds $\mathcal{M}_Y$ and $\mathcal{M}_j$ respectively. Let $D_j := \dim(\mathcal{M}_j)$ be the dimensions of $\mathcal{M}_j$. We allow the case where $\dim(\mathcal{M}_Y) = \infty$. Let Log_y^Y and $\mathrm{Log}_{x_j}^j$ be Riemannian logarithmic maps at y and x_j in $\mathcal{M}_Y$ and $\mathcal{M}_j$, respectively. Also, let μ_Y and μ_j, respectively, denote the Fréchet means of Y and X_j. We consider the following Hilbert-Schmidt linear model:

$$\mathrm{Log}_{\mu_Y}^Y Y = \sum_{j=1}^{p} \mathfrak{B}_j(\mathrm{Log}_{\mu_j}^j X_j) + \varepsilon, \tag{0.1}$$

where $\mathfrak{B}_j$ are Hilbert-Schmidt operators and ε is a random error. We assume a high-dimensional setting where p diverges as the sample size n grows. In this setting, we impose a sparsity condition on the Hilbet-Schmidt operators $\mathfrak{B}_j$, meaning that the number of nonzero operators is relatively small. Our primary goal is to estimate the operators $\mathfrak{B}_j$ and recover the index set $\mathcal{S} := \{1 \leq j \leq p : \mathfrak{B}_j \neq 0\}$.

The estimation procedure is based on the spectral decomposition for X_j. Let $\widehat{\mu}_Y$ and $\widehat{\mu}_j$ be the empirical Fréchet means corresponding to μ_Y and μ_j, respectively. Then, the empirical covariance operators are given by $\hat{\mathcal{C}}_j := n^{-1} \sum_{i=1}^{n} \mathrm{Log}_{\widehat{\mu}_j}^j X_{ij} \otimes \mathrm{Log}_{\widehat{\mu}_j}^j X_{ij}$. Each $\hat{\mathcal{C}}_j$ admits a spectral decomposition $\hat{\mathcal{C}}_j = \sum_{k=1}^{D_j} \hat{\omega}_{jk}(\hat{e}_{jk} \otimes \hat{e}_{jk})$ with eigenvalues $\hat{\omega}_{jk}$ and the corresponding orthonormal basis $\{\hat{e}_{jk} : 1 \leq k \leq D_j\}$. From the spectral decomposition we get the estimated kth scores $\hat{\xi}_{i,jk}$ of $\mathrm{Log}_{\widehat{\mu}_j}^j X_{ij}$. We introduce (truncation) parameters K_j that diverge to infinity as the sample size increases for infinite-dimensional $\mathcal{M}_j$, and

are equal to D_j for finite-dimensional $\mathcal{M}_j$. Then, under the model (0.1) it holds that

$$\mathrm{Log}^Y_{\widehat{\mu}_Y} Y_i \approx \sum_{j=1}^{p} \sum_{k=1}^{K_j} \hat{\xi}_{i,jk} \cdot \mathfrak{B}^*_j(\hat{e}_{jk}) + \mathcal{P}^Y_{\mu_Y,\widehat{\mu}_Y}(\varepsilon_i)$$

where $\mathfrak{B}^*_j := \mathcal{P}^Y_{\mu_Y,\widehat{\mu}_Y} \circ \mathfrak{B}_j \circ \mathcal{P}^j_{\widehat{\mu}_j,\mu_j}$ and $\mathcal{P}^Y_{\mu_Y,\widehat{\mu}_Y}$ ($\mathcal{P}^j_{\widehat{\mu}_j,\mu_j}$) is the parallel transport that maps the tangent space $T_{\mu_Y}\mathcal{M}_Y$ ($T_{\widehat{\mu}_j}\mathcal{M}_j$) of $\mathcal{M}_Y$ ($\mathcal{M}_j$) at μ_Y ($\widehat{\mu}_j$) to the tangent space $T_{\widehat{\mu}_Y}\mathcal{M}_Y$ ($T_{\mu_j}\mathcal{M}_j$) at $\widehat{\mu}_Y$ (μ_j).

We actually estimate $\mathfrak{B}^*_j$ instead of $\mathfrak{B}_j$. Let $\beta^*_{jk} := \hat{\omega}^{1/2}_{jk}\mathfrak{B}^*_j(\hat{e}_{jk})$. We consider a general class of penalty functions ρ_λ, which encompasses the LASSO, SCAD and MCP penalty functions. With $\lambda_j := \sqrt{K_j} \cdot \lambda$ for a universal penalty parameter $\lambda > 0$, we formulate a penalized objective function $\mathcal{L}_n$ defined by

$$\mathcal{L}_n(\boldsymbol{\beta}) := \frac{1}{2n} \sum_{i=1}^{n} \Big\| \mathrm{Log}^Y_{\widehat{\mu}_Y} Y_i - \sum_{j=1}^{p} \sum_{k=1}^{K_j} \hat{\xi}_{i,jk} \hat{\omega}^{-1/2}_{jk} \cdot \beta_{jk} \Big\|^2 + \sum_{j=1}^{p} \rho_{\lambda_j}(\|\boldsymbol{\beta}_j\|),$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK_j})^\top \in (T_{\widehat{\mu}_Y}\mathcal{M}_Y)^{K_j}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top$. We solve the following constrained minimization problem:

$$\hat{\boldsymbol{\beta}}^* := \arg\min \Big\{ \mathcal{L}_n(\boldsymbol{\beta}) : \boldsymbol{\beta} \in (T_{\widehat{\mu}_Y}\mathcal{M}_Y)^{K_+} \text{ with } \sum_{j=1}^{p} \sqrt{K_j} \|\boldsymbol{\beta}_j\| \le R \Big\} \tag{0.2}$$

for some regularization parameter $R \ge 0$. The Hilbert-Schmidt operators $\mathfrak{B}^*_j$ are then estimated by

$$\hat{\mathfrak{B}}^*_j := \sum_{k=1}^{K_j} \hat{\omega}^{-1/2}_{jk} \cdot (\hat{e}_{jk} \otimes \hat{\beta}^*_{jk}).$$

We study the statistical properties of the estimators $\hat{\boldsymbol{\beta}}^*$ and the corresponding $\hat{\mathfrak{B}}^*_j$. We first derive the rates of convergence of the eigenvalues $\hat{\omega}_{jk}$ and the corresponding orthonormal bases $\hat{e}_{jk}$ to their population counterparts that are uniform over the diverging number of covariates. For this, we elicit some concentration inequalities for the empirical Fréchet means $\widehat{\mu}_j$ using empirical process theory. Built on these results, we derive the estimation error bounds of a stationary solution $\hat{\boldsymbol{\beta}}^*$ of the constrained minimization problem (0.2) in various convergence modes. We also show that $\hat{\boldsymbol{\beta}}^*$ exhibits the oracle property with selection consistency. From these results for $\hat{\boldsymbol{\beta}}^*$ we establish the error bounds and the oracle property of the estimated Hilbert-Schmidt operators $\hat{\mathfrak{B}}^*_j$.

REFERENCES

[1] Changwon Choi, and Byeong U. Park (2024). High-dimensional Hilbert-Schmidt linear regression for Hilbert manifold variables. *Manuscript*.

DEPARTMENT OF STATISTICS, SEOUL NATIONAL UNIVERSITY
Email address: bupark@snu.ac.kr

ROBUST INVERSE REGRESSION FOR MULTIVARIATE ELLIPTICAL FUNCTIONAL DATA

EFTYCHIA SOLEA

Classification AMS 2020:

Keywords: dimension reduction, elliptical distribution, functional data, sliced inverse regression, spatial sign

Functional data have received significant attention as they frequently appear in modern applications, such as functional magnetic resonance imaging (fMRI) and natural language processing. The infinite-dimensional nature of functional data makes it necessary to use dimension reduction techniques. Most existing techniques, however, rely on the covariance operator, which can be affected by heavy-tailed data and unusual observations. Therefore, in this paper, we consider a *robust functional sliced inverse regression (R-FSIR) for multivariate elliptical functional data*. For that reason, we define the elliptical distribution for a vector of random functions, extending the existing definition of [1] to the multivariate setting. We introduce a new statistical linear operator, called the *conditional spatial sign Kendall's tau covariance operator*, which can be seen as an extension of the multivariate Kendall's tau to both the conditional and functional settings, and is capable to handle heavy-tailed functional data and outliers. We show that the conditional spatial sign Kendall's tau covariance operator has the same eigenfunctions with the conditional covariance operator, and hence we can formulate the generalized eigenvalue problem based on this new operator to achieve estimation robustness. We derive the convergence rates of the proposed estimators for both completely and partially observed data. In practice, we can only observe the functions at discrete time points, and the new theoretical results support practical estimation procedure. Finally, we demonstrate the finite sample performance of our estimator using simulation examples and a real dataset based on fMRI. We observe that R-FSIR and FSIR have comparable performance for the Gaussian distribution with no outliers. However, R-FSIR outperforms FSIR for heavy-tailed data. Specifically, the efficiency of R-FSIR remains reasonably high, whereas the efficiency of FSIR decreases considerably. This is especially evident when outliers are added to the data.

REFERENCES

[1] Boente, G., Barrera, M. S., & Tyler, D. E. (2014). A characterization of elliptical distributions and some optimality properties of principal components for functional data. Journal of Multivariate Analysis, 131, 254-264.

[2] Ferre, L., and Yao, A. F. (2003). Functional sliced inverse regression analysis. Statistics, 37(6), 475-488.

Eftychia Solea, Eliana Christou and Jun Song Robust Inverse Regression for Multivariate Elliptical Functional Data. *Statistica Sinica*, 2024.

School of Mathematical Sciences, Queen Mary University of London, London, E1 4NS, United Kingdom
E-mail address: e.solea@qmul.ac.uk

DEEP LEARNING FOR FUNCTIONAL AND SURVIVAL DATA

JANE-LING WANG

Classification AMS 2020: 62N02, 62G20.

Keywords: Neural network, functional data analysis, censored data, minimax theory, semi-parametric efficiency

Deep learning, aka deep neural networks, has enjoyed tremendous success in applications for all kinds of data. However, its application to functional data is limited and the theoretical foundation for why it works is still lacking. This talk explores the application of deep neural networks (DNN) to two types of data: functional data and censored survival data.

- **Functional Data**: The infinite dimensionality of functional data means standard learning algorithms can be applied only after appropriate dimension reduction, typically through basis expansions. Currently, these bases are chosen a priori without the information for the task at hand and thus may be suboptimal. We instead propose to adaptively learn these bases in an end-to-end fashion. We introduce a DNN that employs a new basis-layer whose hidden units are each basis functions themselves, implemented as a micro neural network. This architecture learns parsimonious dimension reduction to functional inputs that focuses only on information relevant to the target rather than irrelevant variation in the input function. Across numerous classification and regression tasks that involve functional data this method empirically outperforms other types of DNN.

- **Survival Data**: While DNN have demonstrated empirical success in applications for survival data, most of these methods are difficult to interpret and mathematical understanding of them is lacking. We study the partially linear Cox model, where the nonlinear component of the model is implemented using a deep neural network. The proposed approach is flexible and able to circumvent the curse of dimensionality, yet it facilitates interpretability of the effects of treatment covariates on survival. We establish asymptotic theory for maximum partial likelihood estimators and show that the nonparametric DNN estimator achieves the minimax optimal rate of convergence (up to a poly-logarithmic factor). Moreover, the corresponding parametric estimator for treatment covariate effects is $\sqrt{n}$-consistent, asymptotically normal, and attains semiparametric efficiency. Numerical experiments provide evidence of the advantages of the proposed method.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, DAVIS, CA 95616, USA
Email address: janelwang@ucdavis.edu

TWO-SAMPLE TESTS FOR EQUAL DISTRIBUTIONS IN SEPARABLE METRIC SPACES: A NEW DISTANCE-BASED APPROACH

JIN-TING ZHANG

Classification AMS 2020:62G10,62H10,62H15.

Keywords: two-sample test, equal distribution, distance-based test, high-dimensional data, functional data

1. Introduction

With the surge in advanced data collection methods, researchers are increasingly analyzing complex data objects within separable metric spaces across disciplines. Motivated by real-world datasets such as gene expression and economic indicators, we develop a robust distance-based test to evaluate distributional equality in such data. This paper addresses critical issues in existing methods, offering computational efficiency and accuracy for diverse applications; see [6] for more details.

2. Methodology

The proposed test is formulated using a distance-based statistic that leverages the negative definiteness property of metrics[2,3]. By deriving its asymptotic null distribution as a χ^2-type mixture, we introduce a rapid three-cumulant matching approximation [4] to bypass the computational cost of permutations that are widely adopted as in [2,3]. The test's asymptotic power and root-n consistency are established under local alternatives. These theoretic properties are not established for the methods developed in [2,3].

3. Results and Simulations

Extensive simulations demonstrate the test's superior size control and power across various settings, including high-dimensional and correlated data. Compared to MMD [1] and energy tests [2,3], our method exhibits consistent performance advantages, particularly in computational efficiency. Empirical validation using gene expression data confirms its ability to discern distributional differences effectively.

4. Applications

We apply the proposed test to two datasets: (1) high-dimensional gene expression data distinguishing normal and tumor colon tissues, and (2) functional data on the Gini index across countries. The first dataset, with its dimension much larger than its sample size, is available at http://genomics-pubs.princeton.edu/oncology/affydata/ and the second dataset is downloaded at https://data.worldbank.org/indicator/SI.POV.GINI. Results show the proposed test's robustness against data scaling and sensitivity to kernel

parameter choices in competing methods as developed in [1,2,3], highlighting its practicality in diverse contexts.

5. Conclusion

This study presents a versatile, efficient, and statistically robust approach for two-sample distribution testing in separable metric spaces. Future work includes extending this framework to multi-sample scenarios and exploring other functional data applications.

References

[1] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schlkopf, and A. Smola. A kernel two-sample test. Journal of Machine Learning Research, 13:723773, 2012.

[2] G. J. Szekely and M. L. Rizzo. Testing for equal distributions in high dimension. *Inter Stat, November(5)*, 2004.

[3] G. J. Szekely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):12491272, 2013.

[4] J.-T. Zhang. Approximate and asymptotic distributions of chi-squared-type mixtures with applications. *Journal of the American Statistical Association*, 100(469):273285, 2005.

[5] J.-T. Zhang, J. Guo, and B. Zhou. Testing equality of several distributions in separable metric spaces: A maximum mean discrepancy based approach. *Journal of Econometrics*, page 105286, 2022.

[6] J.-T. Zhang, M. Qian, and T. Zhu. Two-sample tests for equal distributions in separable metric spaces: a new distance-based approach. Unpublished manuscript, 2024.

Department of Statistics and Data Science, National University of Singapore
E-mail address: stazjt@nus.edu.sg