Nonparametric Inference In Functional Data

Zuofeng Shang

Purdue University

Joint work with Guang Cheng from Purdue Univ.

An Example

Consider the functional linear model:

$$Y = \alpha + \int_0^1 X(t)\beta(t)dt + \epsilon,$$

where

- $\beta \in W_2^m(0,1)$, the Sobolev space of order m
- X is a random process
- ϵ is zero-mean error

General Aim

In this talk, we address the following questions in a unified framework:

- how to construct confidence interval for the regression mean $\mu = \alpha + \int_0^1 x(t)\beta(t)dt$?
- how to construct prediction interval for Y_{future}?
- how to test $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$?
-

General Aim

In this talk, we address the following questions in a unified framework:

- how to construct confidence interval for the regression mean $\mu = \alpha + \int_0^1 x(t)\beta(t)dt$?
- how to construct prediction interval for Y_{future}?
- how to test $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$?
-

General Aim

In this talk, we address the following questions in a unified framework:

- how to construct confidence interval for the regression mean $\mu = \alpha + \int_0^1 x(t)\beta(t)dt$?
- how to construct prediction interval for Y_{future}?
- how to test $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$?
-

General Aim

In this talk, we address the following questions in a unified framework:

- how to construct confidence interval for the regression mean $\mu = \alpha + \int_0^1 x(t)\beta(t)dt$?
- how to construct prediction interval for Y_{future}?
- how to test $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$?
-

Literature Review

- The existing methods for inference rely on functional principle component analysis (FPCA), which requires the covariance kernel and reproducing kernel to share common ordered eigenfunctions, i.e., perfectly aligned; Müller and Stadtmüller (2005), Cai and Hall (2006), Hall and Horowitz (2007), etc.
- There is a lack of unified treatment for various inference problems such as confidence/prediction interval construction, (adaptive) hypothesis testing, and functional contrast testing.

Literature Review

- The existing methods for inference rely on functional principle component analysis (FPCA), which requires the covariance kernel and reproducing kernel to share common ordered eigenfunctions, i.e., perfectly aligned; Müller and Stadtmüller (2005), Cai and Hall (2006), Hall and Horowitz (2007), etc.
- There is a lack of unified treatment for various inference problems such as confidence/prediction interval construction, (adaptive) hypothesis testing, and functional contrast testing.

Model and Assumptions:

- Model:

$$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)dt + \epsilon_i,$$

  where $(Y_1, X_1), \ldots, (Y_n, X_n)$ are *iid* samples and $E\{\epsilon_i\} = 0$, $E\{\epsilon_i^2\} = 1$
- Functional parameter: $\beta \in W_2^m(0,1)$, the m-order Sobolev space
- Covariance function: $C(s,t) = E\{X(s)X(t)\}$ satisfies

$$\int_0^1 C(s,t)\beta(s)ds = 0 \Leftrightarrow \beta = 0$$

Model and Assumptions:

- Model:
$$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)dt + \epsilon_i,$$
where $(Y_1, X_1), \ldots, (Y_n, X_n)$ are *iid* samples and $E\{\epsilon_i\} = 0$, $E\{\epsilon_i^2\} = 1$
- Functional parameter: $\beta \in W_2^m(0,1)$, the m-order Sobolev space
- Covariance function: $C(s,t) = E\{X(s)X(t)\}$ satisfies
$$\int_0^1 C(s,t)\beta(s)ds = 0 \Leftrightarrow \beta = 0$$

Model and Assumptions:

- Model:

$$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)dt + \epsilon_i,$$

where $(Y_1, X_1), \ldots, (Y_n, X_n)$ are *iid* samples and $E\{\epsilon_i\} = 0$, $E\{\epsilon_i^2\} = 1$

- Functional parameter: $\beta \in W_2^m(0,1)$, the m-order Sobolev space
- Covariance function: $C(s,t) = E\{X(s)X(t)\}$ satisfies

$$\int_0^1 C(s,t)\beta(s)ds = 0 \Leftrightarrow \beta = 0$$

FPCA Estimation

- Sample covariance function:

$$\widehat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))$$

- KarhunenLoéve decomposition:
  - $C(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t)$ with $\lambda_1 \geq \lambda_2 \geq \dots$
  - $\widehat{C}(s,t) = \sum_{k=1}^{\infty} \widehat{\lambda}_k \widehat{\psi}_k(s)\widehat{\psi}_k(t)$ with $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots$
- Estimate β by $\widehat{\beta} = \widehat{b}_1\widehat{\psi}_1 + \widehat{b}_2\widehat{\psi}_2 + \cdots + \widehat{b}_{k_n}\widehat{\psi}_{k_n}$, where $\widehat{b}_j$ are estimated basis coefficients.

FPCA Estimation

- Sample covariance function:

$$\widehat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))$$

- KarhunenLoéve decomposition:
  - $C(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t)$ with $\lambda_1 \geq \lambda_2 \geq \dots$
  - $\widehat{C}(s,t) = \sum_{k=1}^{\infty} \widehat{\lambda}_k \widehat{\psi}_k(s)\widehat{\psi}_k(t)$ with $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots$
- Estimate β by $\widehat{\beta} = \widehat{b}_1\widehat{\psi}_1 + \widehat{b}_2\widehat{\psi}_2 + \cdots + \widehat{b}_{k_n}\widehat{\psi}_{k_n}$, where $\widehat{b}_j$ are estimated basis coefficients.

FPCA Estimation

- Sample covariance function:

$$\widehat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))$$

- KarhunenLoéve decomposition:
  - $C(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t)$ with $\lambda_1 \geq \lambda_2 \geq \dots$
  - $\widehat{C}(s,t) = \sum_{k=1}^{\infty} \widehat{\lambda}_k \widehat{\psi}_k(s)\widehat{\psi}_k(t)$ with $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots$
- Estimate β by $\widehat{\beta} = \widehat{b}_1\widehat{\psi}_1 + \widehat{b}_2\widehat{\psi}_2 + \cdots + \widehat{b}_{k_n}\widehat{\psi}_{k_n}$, where $\widehat{b}_j$ are estimated basis coefficients.

Penalized Estimation:

$$(\widehat{\alpha}, \widehat{\beta}) = \arg \min_{\alpha \in \mathbb{R}, \beta \in W_2^m(0,1)} \ell_{n,\lambda}(\alpha, \beta),$$

where

$$\ell_{n,\lambda}(\alpha, \beta) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \alpha - \int_0^1 X_i(t)\beta(t)dt)^2 + \frac{\lambda}{2} \int_0^1 |\beta^{(m)}(t)|^2 dt.$$

Advantage of Penalized Estimation

- No perfect alignment assumption
- Provides a unified framework for inference
- Easy to make nonparametric inference within regularization framework
- Estimation performance is better

Advantage of Penalized Estimation

- No perfect alignment assumption
- Provides a unified framework for inference
- Easy to make nonparametric inference within regularization framework
- Estimation performance is better

Advantage of Penalized Estimation

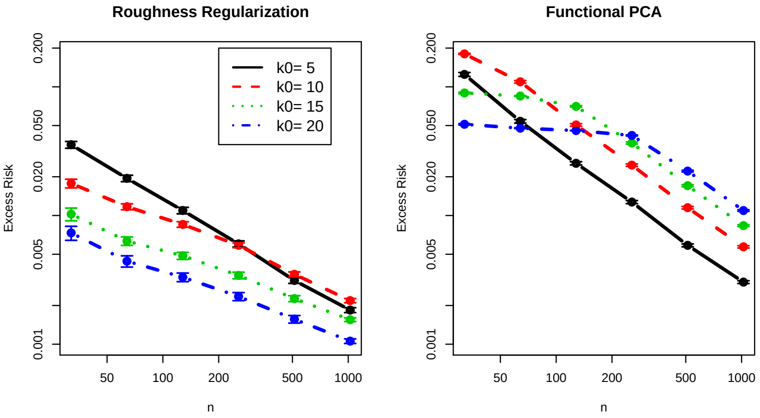
- No perfect alignment assumption
- Provides a unified framework for inference
- Easy to make nonparametric inference within regularization framework
- Estimation performance is better

Advantage of Penalized Estimation

- No perfect alignment assumption
- Provides a unified framework for inference
- Easy to make nonparametric inference within regularization framework
- Estimation performance is better

A Graphical Comparison of FPCA and Penalized Estimation: Cai and Yuan (2012)

k_0 controls the alignment between covariance and reproducing kernels. Larger value of k_0 yields more misalignment.



Assumption: Simultaneous Diagonalization

There exists functions φ_ν and nondecreasing sequences $\rho_\nu \asymp \nu^{2k}$ for some $k > 0$ such that for any $\nu, \mu \geq 1$,

$$\int_0^1 \int_0^1 C(s,t)\varphi_\nu(s)\varphi_\mu(t)dsdt = \delta_{\nu\mu},$$

and

$$\int_0^1 \varphi_\nu^{(m)}(t)\varphi_\mu^{(m)}(t)dt = \rho_\nu \delta_{\nu\mu}.$$

Furthermore, any $\beta \in W_2^m(0,1)$ satisfies $\beta = \sum_\nu b_\nu \varphi_\nu$ for some real sequence b_ν.

Construction of CI

Let $\mu_0 = \alpha + \int_0^1 x_0(t)\beta(t)dt$ be the regression mean at $X = x_0$. The 95% confidence interval for μ_0 is

$$CI : \widehat{\mu}_0 \pm 1.96\sigma_n/\sqrt{n},$$

where $\widehat{\mu}_0 = \widehat{\alpha} + \int_0^1 x_0(t)\widehat{\beta}(t)dt$, $\sigma_n^2 = 1 + \sum_\nu \frac{x_\nu^2}{1+\lambda\rho_\nu}$, $x_\nu = \int_0^1 x_0(t)\varphi_\nu(t)dt$.

Construction of PI

Let Y_0 be future response generated from $Y_0 = \mu_0 + \epsilon$, then the 95% prediction interval for Y_0 is

$$PI : \widehat{\mu}_0 \pm 1.96\sqrt{1 + \sigma_n^2/n}.$$

Theoretical Validity

Theorem

If ϵ is sub-exponential, the true function β_0 is suitably smooth, and λ is properly tuned, e.g., $\lambda \asymp n^{-k/(2k+1)}$. Then as $n \to \infty$,

$$P(\mu_0 \in CI) \to 0.95, \text{ and } P(Y_0 \in PI) \to 0.95.$$

Penalized Likelihood Ratio Test

Testing hypotheses $H_0 : \alpha = \alpha_0, \beta = \beta_0$ versus $H_1 : H_0$ is not true. Define the penalized likelihood ratio test (PLRT)

$$PLRT_n = \ell_{n,\lambda}(\alpha_0, \beta_0) - \ell_{n,\lambda}(\widehat{\alpha}, \widehat{\beta}),$$

where $(\widehat{\alpha}, \widehat{\beta})$ is the penalized MLE.

Wilks Phenomenon

Wilks phenomenon means that the null limit distribution of the likelihood ratio is free of any nuisance parameters and design distribution.

Theorem

Suppose H_0 holds and $E\{\epsilon^4\} < \infty$, and λ is suitably tuned, e.g., $\lambda \asymp n^{-4k/(4k+1)}$. Then

$$2n\sigma^2 \cdot PLRT_n \overset{d}{\approx} \chi^2_{u_n},$$

where

$$\sigma^2 = \frac{\int_0^\infty (1+x^{2k})^{-1}dx}{\int_0^\infty (1+x^{2k})^{-2}dx}, \quad u_n = \frac{1}{c\lambda^{\frac{1}{2k}}} \frac{(\int_0^1 (1+x^{2k})^{-1}dx)^2}{(\int_0^1 (1+x^{2k})^{-2}dx)},$$

c is constant free of α_0, β_0, distribution of X.

Minimax Property of PLRT

Suppose we want to test $H_0 : \beta = 0$, but the following local alternative hypothesis is true:

$$H_{1n} : \beta = \beta_n,$$

where β_n satisfies $\|\beta_n\|_{L^2} \geq cn^{-2k/(4k+1)}$.

Theorem

For arbitrary $\varepsilon > 0$, there exist c such that for any $n \geq 1$:

$$\inf_{\beta_n \in W_2^m(0,1) : \|\beta_n\|_{L^2} \geq cn^{-2k/(4k+1)}} P_{\beta_n}(\textit{reject } H_0) \geq 1 - \varepsilon.$$

An Example: Standard Brownian Motion

When $m = 2$ (cubic spline) and X is Brownian motion with covariance function

$$C(s,t) = \min\{s,t\}, \ \ s,t \in (0,1),$$

we have $\sigma^2 \approx 1.08$ and $u_n \approx 0.31\lambda^{-1/6}$. Therefore,

$$2n(1.08) \cdot PLRT_n \overset{d}{\approx} \chi^2_{u_n}.$$

An Adaptive Testing Procedure Based on Likelihood Ratio

If the smoothness degrees of both X and β are unknown, how well can we do? We will propose a testing procedure adaptive to these smoothness degrees and show that our procedure achieves the minimax rate of testing.

Let $PLRT(k)$ be the penalized likelihood ratio test associated with k, and

$$\tau_k = \frac{PLRT(k) - E\{PLRT(k)\}}{\sqrt{Var(PLRT(k))}}, \; k = 1, 2, \ldots, k_n.$$

Define

$$AT = B_n(\max_{1 \le k \le k_n} \tau_k - B_n),$$

where B_n satisfies $2\pi B_n^2 \exp(B_n^2) = k_n^2$.

Size of the Test

A valid test should achieve the correct size.

Theorem

Under $H_0 : \beta = 0$, if $k_n \asymp (\log n)^{d_0}$, for some constant $d_0 \in (0, 1/2)$, then for any $\gamma \in (0, 1)$,

$$P(AT \leq c_\gamma) \to 1 - \gamma, \quad \text{as } n \to \infty,$$

where $c_\gamma = -\log(-\log(1 - \gamma))$.

Adaptive Minimax Rate

Suppose k^* is the true value of k. Let

$$\delta(n, k^*) = n^{-2k^*/(4k^*+1)}(\log\log n)^{k^*/(4k^*+1)}.$$

Theorem

Suppose $k_n \asymp (\log n)^{d_0}$, for some constant $d_0 \in (0, 1/2)$. Then, for any $\varepsilon \in (0, 1)$, there exists $c > 0$ s.t. for any $n \geq 1$,

$$\inf_{\|\beta\|_{L^2} \geq c\delta(n,k^*)} P_\beta(\textit{reject } H_0) \geq 1 - \varepsilon.$$

Simulation Setup

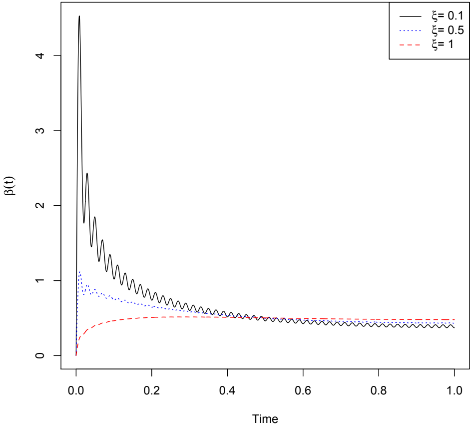
- $X(t) = \sum_{j=1}^{100} \sqrt{\lambda_j}\eta_j V_j(t)$, where

$$\lambda_j = (j-0.5)^{-2}\pi^{-2}, \quad V_j(t) = \sqrt{2}\sin((j-0.5)\pi t),$$

$$\eta_1, \ldots, \eta_{100} \overset{iid}{\sim} N(0,1).$$

- The test function is $\beta_0^{B,\xi} = \frac{B}{\sqrt{\sum_{k=1}^{\infty} k^{-2\xi-1}}} \sum_{j=1}^{100} j^{-\xi-0.5} V_j(t)$, where $B = 0, 0.1, 1$ and $\xi = 0.1, 0.5, 1$.
- Draw n *iid* samples from $Y = \int_0^1 X(t)\beta_0(t)dt + N(0,1)$ for $n = 100, 500$.

Figure: Plots of $\beta_0(t)$ when $B = 1$



Coverage Proportion of Confidence Interval

Table: $100\times$ coverage proportion (average length) of CI when $B = \xi = 1$

$n = 100$	$n = 500$
95.11(0.56)	94.99(0.39)

Size Comparison with Hilgert, Mas and Verzelen (2013)

Hilgert, Mas and Verzelen (2013) proposed an FPCA-based testing procedure which is adaptive to the truncation parameter k_n. We compare our approaches with theirs, denoted HMV.

Table: 100×size when $B = 0$

	$n = 100$	$n = 500$
HMV	4.97	5.26
PLRT	5.45	5.19
AT	5.13	5.04

Power Comparison with Hilgert, Mas and Verzelen (2013)

Table: 100×power when $n = 100$

	Test	$B = 0.1$	$B = 1$
$\xi = 0.1$	HMV	5.80	81.78
	AT	6.12	81.56
	PLRT	20.00	84.20
$\xi = 1$	HMV	7.07	99.84
	AT	9.47	99.98
	PLRT	23.95	99.98

Power Comparison with Hilgert, Mas and Verzelen (2013)

Table: 100×power when $n = 500$

	Test	$B = 0.1$	$B = 1$
$\xi = 0.1$	HMV	8.48	100
	AT	9.57	100
	PLRT	21.27	100
$\xi = 1$	HMV	16.13	100
	AT	26.51	100
	PLRT	34.08	100

Summary

- We propose applicable procedures for inference in functional data analysis
- Our approaches do not require perfect alignment
- Our approaches are asymptotic valid, i.e., desired size and coverage probability
- The PLRT and adaptive testing procedures are more powerful than existing ones.
- Extensions to general cases not reported here:
  - quasi-likelihood framework
  - composite hypotheses
  - adaptive testing in non-Gaussian error

Thank you for your attention!