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
- \( \epsilon \) is zero-mean error
In this talk, we address the following questions \textbf{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 $Y_{\text{future}}$?
- how to test $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$?
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 \( H_1 : \beta \neq \beta_0 \)?
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.
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:** \(\beta \in W^m_2(0, 1)\), the \(m\)-order Sobolev space

- **Covariance function:** \(C(s, t) = E\{X(s)X(t)\}\) satisfies
  \[ \int_0^1 C(s, t)\beta(s)ds = 0 \iff \beta = 0 \]
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:** \(\beta \in W^m_2(0, 1)\), the \(m\)-order Sobolev space

- **Covariance function:** \(C(s, t) = E\{X(s)X(t)\}\) satisfies
  \[ \int_0^1 C(s, t)\beta(s)ds = 0 \iff \beta = 0 \]
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
  \[ \int_0^1 C(s, t)\beta(s)ds = 0 \iff \beta = 0 \]
FPCA Estimation

- Sample covariance function:
  \[
  \hat{C}(s, t) = \frac{1}{n} \sum_{i=1}^{n} (X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))
  \]

- Karhunen-Loéve decomposition:
  \[
  C(s, t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t) \quad \text{with} \quad \lambda_1 \geq \lambda_2 \geq \ldots
  \]
  \[
  \hat{C}(s, t) = \sum_{k=1}^{\infty} \hat{\lambda}_k \hat{\psi}_k(s)\hat{\psi}_k(t) \quad \text{with} \quad \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots
  \]
  Estimate \( \beta \) by \( \hat{\beta} = \hat{b}_1 \hat{\psi}_1 + \hat{b}_2 \hat{\psi}_2 + \ldots + \hat{b}_k \hat{\psi}_k + \ldots \), where \( \hat{b}_j \) are estimated basis coefficients.
FPCA Estimation

- Sample covariance function:

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

- Karhunen-Loéve decomposition:

  \[
  C(s, t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t) \text{ with } \lambda_1 \geq \lambda_2 \geq \ldots
  \]

  \[
  \hat{C}(s, t) = \sum_{k=1}^{\infty} \hat{\lambda}_k \hat{\psi}_k(s)\hat{\psi}_k(t) \text{ with } \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots
  \]

- Estimate \( \beta \) by \( \hat{\beta} = \hat{b}_1 \hat{\psi}_1 + \hat{b}_2 \hat{\psi}_2 + \cdots + \hat{b}_k \hat{\psi}_k \), where \( \hat{b}_j \) are estimated basis coefficients.
FPCA Estimation

- Sample covariance function:

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

- KarhunenLoève decomposition:

\[
C(s, t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t) \text{ with } \lambda_1 \geq \lambda_2 \geq \ldots
\]

\[
\hat{C}(s, t) = \sum_{k=1}^{\infty} \hat{\lambda}_k \hat{\psi}_k(s)\hat{\psi}_k(t) \text{ with } \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots
\]

- Estimate \( \beta \) by \( \hat{\beta} = \hat{b}_1 \hat{\psi}_1 + \hat{b}_2 \hat{\psi}_2 + \cdots + \hat{b}_n \hat{\psi}_n \), where \( \hat{b}_j \) are estimated basis coefficients.
Penalized Estimation:

\( (\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha \in \mathbb{R}, \beta \in W^m_2(0,1)} \ell_{n, \lambda}(\alpha, \beta), \)

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
$k_0$ controls the alignment between covariance and reproducing kernels. Larger value of $k_0$ yields more misalignment.
Assumption: Simultaneous Diagonalization

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

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

and

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

Furthermore, any $\beta \in W^m_2(0, 1)$ satisfies $\beta = \sum_\nu b_\nu \varphi_\nu$ for some real sequence $b_\nu$. 
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 $\mu_0$ is

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

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

$$PI : \hat{\mu}_0 \pm 1.96 \sqrt{1 + \frac{\sigma_n^2}{n}}.$$
Theoretical Validity

**Theorem**

If $\epsilon$ is sub-exponential, the true function $\beta_0$ is suitably smooth, and $\lambda$ is properly tuned, e.g., $\lambda \asymp n^{-k/(2k+1)}$. Then as $n \to \infty$,

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

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

where $(\hat{\alpha}, \hat{\beta})$ 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 $\lambda$ is suitably tuned, e.g., $\lambda \approx n^{-4k/(4k+1)}$. Then

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

where

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

$c$ is constant free of $\alpha_0, \beta_0$, distribution of $X$. 

**Wilks Phenomenon**

Wilks phenomenon means that the null limit distribution of the likelihood ratio is free of any nuisance parameters and design distribution.
Suppose we want to test $H_0 : \beta = 0$, but the following local alternative hypothesis is true:

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

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

**Theorem**

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

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

\[
2n(1.08) \cdot PLRT_n \overset{d}{\approx} \chi_{u_n}^2.
\]
If the smoothness degrees of both $X$ and $\beta$ 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
\[
AT = B_n(\max_{1 \leq k \leq k_n} \tau_k - B_n),
\]
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 $\gamma \in (0, 1)$,

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

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

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

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

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

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

- The test function is \( \beta_{0,\xi}^{B} = \frac{B}{\sqrt{\sum_{k=1}^{\infty} k^{-2\xi - 1}}} \sum_{j=1}^{100} j^{-\xi - 0.5} V_j(t) \), 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× coverage proportion (average length) of CI when $B = \xi = 1$

<table>
<thead>
<tr>
<th>n</th>
<th>100</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95.11(0.56)</td>
<td>94.99(0.39)</td>
</tr>
</tbody>
</table>
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 $B = 0$

<table>
<thead>
<tr>
<th></th>
<th>$n = 100$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMV</td>
<td>4.97</td>
<td>5.26</td>
</tr>
<tr>
<td>PLRT</td>
<td>5.45</td>
<td>5.19</td>
</tr>
<tr>
<td>AT</td>
<td>5.13</td>
<td>5.04</td>
</tr>
</tbody>
</table>
Power Comparison with Hilgert, Mas and Verzelen (2013)

Table: 100×power when $n = 100$

<table>
<thead>
<tr>
<th>Test</th>
<th>$B = 0.1$</th>
<th>$B = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi = 0.1$</td>
<td>HMV 5.80</td>
<td>81.78</td>
</tr>
<tr>
<td></td>
<td>AT 6.12</td>
<td>81.56</td>
</tr>
<tr>
<td></td>
<td>PLRT 20.00</td>
<td>84.20</td>
</tr>
<tr>
<td>$\xi = 1$</td>
<td>HMV 7.07</td>
<td>99.84</td>
</tr>
<tr>
<td></td>
<td>AT 9.47</td>
<td>99.98</td>
</tr>
<tr>
<td></td>
<td>PLRT 23.95</td>
<td>99.98</td>
</tr>
</tbody>
</table>
Power Comparison with Hilgert, Mas and Verzelen (2013)

<table>
<thead>
<tr>
<th>Test</th>
<th>$B = 0.1$</th>
<th>$B = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\xi = 0.1$</td>
<td>HMV 8.48</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>AT 9.57</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>PLRT 21.27</td>
<td>100</td>
</tr>
<tr>
<td>$\xi = 1$</td>
<td>HMV 16.13</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>AT 26.51</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>PLRT 34.08</td>
<td>100</td>
</tr>
</tbody>
</table>
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!