Statistical Inferences for Functional Data

Jin-Ting Zhang
Department of Statistics
National University of Singapore

July 20, 2004

Joint work with Dr. Jianwei Chen, Univ. of Rochester
OUTLINE

- **Introduction**
  - The Motivation Data
  - Nonparametric Mixed-Effects Model
  - Existing Approaches

- **Estimation**
  - Reconstructing the Individual Functions
  - Estimation of the Mean and Covariance Functions
  - Estimation of the Noise Variance Function
  - Bandwidth Selection

- **Hypothesis Testing**
  - Test about a Specific Mean Function
  - Polynomial Goodness of Fit Test
  - Implementation of the Test Procedures

- **Simulation Study**

- **A Real Data Example**
1 Introduction

1.1 The Motivation Data

![Graphs showing Raw Data and Pointwise Means](image)

Figure 1: Raw curves of the nonconceptive group of the progesterone data (Brumback & Rice 1998)


- Consist of a sample of \( n = 69 \) urinary metabolite progesterone curves (Munro et al. 1991) of women’s menstrual cycles.
- Scheduled design time points: 24 and Missing rate: 10.02%
1.2 Nonparametric Mixed-Effects Model

A NPME Model shall be used to model the above curve data:

\[ y_{ij} = \eta(t_{ij}) + v_i(t_{ij}) + \epsilon_i(t_{ij}), \ j = 1, \ldots, n_i, \ i = 1, \ldots, n \]

- **Overall mean function** (Fixed Effects): \( \eta(t) \)
- **Subject effects** (Random Effects): \( v_i(t) \sim SP(0, \gamma) \)

- **\( i \)-th individual curve:** \( f_i(t) = E[y_i(t)|v_i(t)] = \eta(t) + v_i(t) \)
- **Uncorrelated measurements errors:** \( \epsilon_i(t) \sim SP(0, \sigma^2(t)) \)
- \( n_i \) are **quite large**, often much larger than \( n \)
- \( t \in \mathcal{T} = [a, b], \ -\infty < a < b < \infty. \)
Targets of Interest:

- Estimate the Overall Mean $\eta(t)$ of the curves
- Estimate each of the individual curves $f_i(t)$
- Estimate the Covariance Structure $\gamma(s, t)$ of the individual curves
- Hypothesis tests about the Overall Mean Curve
  - Test $H_0$: $\eta(t) = 0$ (No effect test)
  - Test $H_0$: $\eta(t)$ is a polynomial
    (Polynomial goodness of fit test)
1.3 Existing Approaches

Most of the existing approaches focus on estimation of $\eta(t)$ and PCA of $\gamma(s,t)$, involving one smoother or another.


- **Local Polynomial Smoothing** (Fan & Zhang 2000)

- **Basis-based Approaches** (Ramsay and Silverman 1997)

- **Reproducing kernel Hilbert spline basis** (Ramsay 1986 among others)
Work on Hypothesis Tests about $\eta(t)$.

- **Pointwise t-test/F-test** (Ramsay and Silverman 1997, 2000)
  - **Main Idea**: at each time point, Conduct a t-test/F-test
  - **Advantage**: Simple to understand, easy to implement
  - **Drawback**: Overall Test Result NOT available

- **Adaptive Neyman Test** (Fan and Lin 1998)
  - **Main Idea**: Apply the Adaptive Neyman Test of Fan (1996) to the Fourier Coefficients of the curve data
  - **Advantage**: Powerful
  - **Drawback**: Only for curve data yielded from Gaussian white noise process or stationary Gaussian process
• **Methodologies of the current talk**

  – Reconstructing $f_i(t), i = 1, 2, \cdots, n$ using Local Polynomial Smoothing
  – Estimation of $\eta(t), \gamma(s, t), \sigma^2(t)$ based on the Reconstructions
  – Hypothesis Tests about $\eta(t)$ based on the Reconstructions
  – Theoretical Analysis of the **Estimators** and **Test Statistics**
2 Estimation

2.1 Reconstruction of the Individual Curves

**The Approach**

- Based on *i*-th Individual measurements \((t_{ij}, y_{ij})\)
- Use the usual **Nonpar. Regr. Model**: \(y_i(t_{ij}) = f_i(t_{ij}) + \epsilon_{ij}\),
- Target at the **Individual function** \(f_i(t) = \eta(t) + v_i(t)\)
- Use **Local Polynomial Smoothing** (Fan and Gijbels 1996)
  - \(p\)-th order polynomial Taylor expansion at \(t_0\):
    \[ f_i(t) \approx f_i(t_0) + (t - t_0)f_i^{(1)}(t_0) + \cdots + (t - t_0)^p f_i^{(p)}(t_0)/p! , \]
  - Weighted Least Squares Estimation: \(K\)-kernel, \(h_i\)-bandwidth
    \[ \sum_{j=1}^{n_i} \left\{ y_{ij} - \sum_{r=0}^{p} (t_{ij} - t_0)^r \beta_r \right\}^2 K_{h_i}(t_{ij} - t_0), \]
  - Reconstruction at \(t_0\): \(\hat{f}_i(t_0) = \hat{\beta}_0\).
Theorem 1  Under some regularization conditions, and in particular when there are two positive constants $C_0$ and $\delta_0$ such that
\[ n_i \geq C_0 n^{\delta_0}, \; i = 1, 2, \cdots, n. \]

Then for the $\hat{f}_i(t)$ using the individual optimal bandwidths, we have
\[ \hat{f}_i(t) = f_i(t) + n^{-\delta} O_P(1), \; i = 1, 2, \cdots, n, \]
where $O_P(1)$ is uniform for any $t$ within the interior of $T$ and all
\[ i = 1, 2, \cdots, n, \text{ and } \delta = \begin{cases} \frac{p+1}{2p+3} \delta_0, & \text{as } p \text{ odd,} \\ \frac{p+2}{2p+5} \delta_0, & \text{as } p \text{ even.} \end{cases} \]

Remarks

- Reconstruct $\hat{f}_i(t) \approx f_i(t), i = 1, 2, \cdots, n$ provided that $n_i \to \infty$ slightly faster than $n$.

- **Often require** $\delta > 1/2$. To make this satisfied, need
\[ \delta_0 > \begin{cases} 1 + 1/(2p + 2), & \text{as } p \text{ odd,} \\ 1 + 1/(2p + 4), & \text{as } p \text{ even.} \end{cases} \]
when $p = 0$ or 1, need $\delta_0 > 5/4$.

- **For Functional Data**, the above condition is easy to satisfy.
2.2 Estimation of the Mean and Covariance Functions

- **Proposed Estimators** using the reconstructions \( \hat{f}_i(t) \):
  \[
  \hat{\eta}(t) = n^{-1} \sum_{i=1}^{n} \hat{f}_i(t) \quad \hat{\gamma}(s, t) = n^{-1} \sum_{i=1}^{n} \left\{ \hat{f}_i(s) - \hat{\eta}(s) \right\} \left\{ \hat{f}_i(t) - \hat{\eta}(t) \right\} .
  \]
  - **Advantage**: Simple, Natural and Computatable

- **Ideal Estimators** using the true individuals \( f_i(t) \):
  \[
  \tilde{\eta}(t) = n^{-1} \sum_{i=1}^{n} f_i(t) , \quad \tilde{\gamma}(s, t) = n^{-1} \sum_{i=1}^{n} \left\{ f_i(s) - \tilde{\eta}(s) \right\} \left\{ f_i(t) - \tilde{\eta}(t) \right\} .
  \]
  - **Drawback**: Not Computatable

Shall Show that under some regular conditions,

**Proposed Estimators \( \approx \) Ideal Estimators.**
Theorem 2 Under the conditions of Theorem 1, we have
\[ \hat{\eta}(t) = \tilde{\eta}(t) + n^{-\delta} O_P(1), \quad \hat{\gamma}(s,t) = \tilde{\gamma}(s,t) + n^{-\delta} O_P(1), \]
where \( O_P(1) \) is uniform for all the interior points of \( T \). In addition, assume \( \delta > 1/2 \). Then as \( n \to \infty \), we have
\[ n^{1/2} \{ \hat{\eta}(t) - \eta(t) \} \sim AGP(0, \gamma), \quad n^{1/2} \{ \hat{\gamma}(s,t) - \gamma(s,t) \} \sim AGP(0, \gamma_*), \]
where
- \( AGP(\eta, \gamma) \)—asymptotical Gaussian process with mean function \( \eta(t) \) and covariance function \( \gamma(s,t) \),
- \( \gamma_* \{(s_1, t_1), (s_2, t_2)\} = E \{ v_1(s_1)v_1(t_1)v_1(s_2)v_1(t_2) \} - \gamma(s_1, t_1)\gamma(s_2, t_2), \)
- \( v_1(t) \)—the subject effect of \( f_1(t) \); when \( v_1(t) \) is Gaussian,
  \[ \gamma_* \{(s_1, t_1), (s_2, t_2)\} = \gamma(s_1, t_2)\gamma(s_2, t_1) + \gamma(s_1, s_2)\gamma(t_1, t_2). \]
2.3 Estimation of the Noise Variance Function

- Use the individual residuals \( \hat{e}_{ij} = y_{ij} - \hat{f}_i(t_{ij}) \).

- Use kernel estimator

\[
\hat{\sigma}^2(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} K_h(t_{ij} - t) \hat{e}_{ij}^2}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} K_h(t_{ij} - t)}.
\]

- Kernel \( K \) and bandwidth \( h \) MAY different from those for \( \hat{f}_i(t) \).

- \( N = \sum_{i=1}^{n} n_i \): total number of the measurements.

**Theorem 3** Assume the conditions for Theorem 1 are satisfied. In addition, assume a bandwidth \( h = O(N^{-1/5}) \) is used for \( \hat{\sigma}^2(t) \). Then as \( n \to \infty \)

\[
\hat{\sigma}^2(t) = \sigma^2(t) + n^{-\delta}O_P(1),
\]

where \( O_P(1) \) is uniform for all the interior points of \( \mathcal{T} \).
2.4 Bandwidth Selection

- **Pooled Bandwidth Method:** $h_{\text{pool}}$
  Select the bandwidth from the pooled functional data using usual bandwidth selector such as GCV.

- **Individual Bandwidth Method:** $h_i$
  Select an individual bandwidth per subject using usual bandwidth selector such as GCV.

- **Subject-out Bandwidth Method:** $h_{\text{SCV}}$
  Select the bandwidth using the one-subject-out cross-validation (Rice and Silverman 1991).

- **Generally** $h_{\text{SCV}} < h_{\text{pool}} < h_i$. 
3 Hypothesis Tests about the Mean Function

3.1 Test about a Specific Mean Function

- **The Testing Problem:**
  \[ H_0 : \eta(t) = \eta_0(t) \text{ for any } t, \]

  - **"No Effect" Test** when \( \eta_0(t) \equiv 0, \)
  - **Pairwise tests** easily transformed into "no effect" tests.

- **Proposed Test Statistic:** \( L_2 \)-norm based
  \[ T_n = n \| \hat{\eta} - \eta_0 \|^2 = n \int \{ \hat{\eta}(t) - \eta_0(t) \}^2 \, dt. \]

- **Ideal Test Statistic:** assume \( f_i(t), i = 1, 2, \ldots, n \) known
  \[ T_{n,0} = n \| \hat{\eta} - \eta_0 \|^2 = n \int \{ \hat{\eta}(t) - \eta_0(t) \}^2 \, dt, \]
Theorem 4 Assume the conditions of Theorem 1 are satisfied. In addition, assume $\gamma$ has a finite trace. Then as $n \to \infty$, we have

- $T_n = T_{n,0} + n^{-(\delta-1/2)}O_P(1)$.
- $T_{n,0} \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r A_r + o_P(1), \quad A_r \sim \chi_{1}^2(nu_r^2/\lambda_r)$.
- In addition, when $\delta > 1/2$,

$$T_n \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r A_r + o_P(1).$$

where

- $u_r = \int (\eta(t) - \eta_0(t))\phi_r(t)dt, r = 1, 2, \ldots$,
- $\lambda_r, r = 1, 2, \ldots$ — eigenvalues of $\gamma$,
- $\phi_r(t), r = 1, 2, \ldots$ — eigen functions of $\gamma$. 

16
3.2 Polynomial Goodness of Fit Test

- **Polynomial Goodness of Fit Test:**
  \[ H_0 : \eta(t) = \Psi_m(t)^T \beta \text{ for some } \beta, \quad \Psi_m(t) = (1, t, t^2, \ldots, t^m)^T. \]

- **Examples:** No-effect Test \((m = 0)\)/ Linearity Test \((m = 1)\)/ Quadratic Test \((m = 2)\)

- **Least Squares Estimation** of \(\beta\)
  - Minimize \(\sum_{i=1}^{n} \int \left( \hat{y}_i(t) - \Psi_m(t)^T \beta \right)^2 dt\)
  - The Solution
  \[
  \hat{\beta} = \left\{ \int \Psi_m(t)\Psi_m(t)^T dt \right\}^{-1} \int \Psi_m(t)\hat{\eta}(t)dt = \int w(t)\hat{\eta}(t)dt,
  \]
  where \(w(t) = \left( \int \Psi_m(u)\Psi_m(u)^T du \right)^{-1} \Psi_m(t)\).
**Proposed Test Statistic:**

\[ D_n = n \int \left\{ \hat{\eta}(t) - \Psi_m(t)^T \hat{\beta} \right\}^2 dt = n \int d_n^2(t) dt, \]

where \( d_n(t) = \hat{\eta}(t) - \Psi_m(t)^T \hat{\beta} \).

**Asymptotic Property of** \( d_n(t) \):

**Theorem 5** Assume the conditions of Theorem 4 are satisfied. Then as \( n \to \infty \),

\[ n^{1/2} [d_n(t) - d(t)] \sim AGP(0, \gamma_d), \]

where

- \( \gamma_d(s, t) = \sum_{r=1}^{\infty} \lambda_r d_r(s)d_r(t) \),
- \( d(t) = \eta(t) - \Psi_m(t)^T \int w(u)\eta(u)du \),
- \( d_r(t) = \phi_r(t) - \Psi_m(t)^T \int w(u)\phi_r(u)du \),
- \( [\lambda_r, \phi_r(t)], r = 1, 2, \ldots — eigenvalues/eigenfunctions of \gamma \),
Asymptotic Random Expression of $D_n$:

**Theorem 6** Assume the conditions in Theorem 1 are satisfied. In addition, assume $\gamma_d(s,t)$ has a finite trace. Then

$$D_n \xrightarrow{d} \sum_{r=1}^{\infty} \tau_r A_r + n^{-(\delta-1/2)}O_P(1),$$

where

- $A_r \sim \chi^2_{1}(nu_r^2/\tau_r)$, independent of each other,
- $u_r = \int [d_n(t) - d(t)]\psi_r(t)dt$,
- $[\tau_r, \psi_r(t)], r = 1, 2, \cdots$—eigenvalues/eigenfunctions of $\gamma_d$. 


3.3 Implementation of the Testing Procedures

- **Remarks and Challenge**
  - Test statistics $T_n$ and $D_n$ are \( \chi^2 \)-type mixtures
  - Under $H_0$, the noncentral parameters $u_r$ are 0
  - Eigenvalues and eigenfunctions of $\gamma$ or $\gamma_d$ **Unknown**

- **Implementation**
  - **Replace the Unknowns by their Estimates**
  - **Use \( \chi^2 \)-approximation** for the resulting test statistics; see Buckley and Eagleson (1988)
    Ramil-Novos and Gonzalez-Manteiga (1998)
    Zhang (2004)
4 Simulation Study

- **Aim:** Compare the Bandwidth Selection Methods via comparing
  - the resulting bandwidths
  - the associated subject-out CV scores
  - the MSE for estimating $\eta(t)$:
    \[
    \text{MSE}_\eta = M^{-1} \sum_{j=1}^{M} [\eta(\tau_j) - \hat{\eta}(\tau_j)]^2
    \]
  - the MSE for estimating $f_i(t), i = 1, 2, \ldots, n$:
    \[
    \text{MSE}_f = (nM)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M} \left\{ \hat{f}_i(\tau_j) - f_i(\tau_j) \right\}^2
    \]
  where $\tau_1, \ldots, \tau_M$ equally-spaced in $[0, 1]$, for some $M$ large.
• **Simulation Model:**

\[ y_i(t) = \eta(t) + v_i(t) + \epsilon_i(t), \]
\[ \eta(t) = a_0 + a_1 \phi_1(t) + a_2 \phi_2(t), \]
\[ v_i(t) = b_{i0} + b_{i1} \psi_1(t) + b_{i2} \psi_2(t), \]
\[ b_i = [b_{i0}, b_{i1}, b_{i2}]^T \sim N[0, \text{diag}(\sigma_0^2, \sigma_1^2, \sigma_2^2)], \]
\[ \epsilon_i(t) \sim N[0, \sigma^2(1 + t)], \quad i = 1, 2, \cdots, n, \]

- \( b_i \) and \( \epsilon_i(t) \) are independent
- Scheduled time points \( t_j = j/(m + 1), j = 1, 2, \cdots, m \)
- Randomly remove some responses at a rate \( r_{\text{min}} \) to yield unbalance design

• **Simulation Parameters Used:**

- \([a_0, a_1, a_2] = [1.2, 2.3, 4.2], \quad [\sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma^2] = [1, 2, 3, .5]\)
- \( \phi_1(t) = \psi_1(t) = \cos(2\pi t), \phi_2(t) = \psi_2(t) = \sin(2\pi t) \)
- \( r_{\text{min}} = 10\%, [n, m] = [20, 40], M = 400.\)
Figure 2: *Simulation results.*

**Simulation Results**

- Panel (a) $h_i$ (C2) > $h_{pool}$ (C1) > $h_{SCV}$ (C3) for 200 simulations;
- Panel (b) One-subject-out CV scores associated with 3 methods;
- Panel (c) MSE$_{\eta}$ values for 3 methods, and the method using $\hat{\eta}(t)$;
- Panel (d) MSE$_{f}$ values for 3 methods.
- Overall, **Ind. Bandwidths Preferred** for $\hat{\eta}(t)$ and $\hat{f}_i(t)$.
A Real Data Example

![Graphs showing reconstructed curves, mean ±2 stddev bands, covariance function, and noise variance function.](image)

Figure 3: *FDA of the Nonconceptive Progesterone data.*

- **Estimation**
  - Panel (a), \( \hat{f}_i(t) \) by a loc. lin. smoother using ind. bandwidths;
  - Panel (b), \( \hat{\eta}(t) \) based on the reconstructions;
  - Panel (c), \( \hat{\gamma}(s, t) \) based on the reconstructions;
  - Panel (d), \( \hat{\sigma}^2(t) \) by loc. const. kernel smoother using \( h_{GCV} \).
Hypothesis Tests

- $\eta(t)$ definitely not constant or linear
- Quadratic, cubic or quartic tests Rejected with Pvalue 0
- No-effect Test for the residuals after $\hat{\eta}(t)$ removed Accepted with Pvalue 1.

Thank You