F-type Tests for Linear Models with Functional Responses

Jin-Ting Zhang
National University of Singapore

stazjt@nus.edu.sg
OUTLINE

• Introduction
• F-type Test for the GLHT Problem
• Special Cases of the GLHT Problem
• Variable Selection in FLM
• Outlier Detection for FLM
• Summary and Discussion
Introduction

Outline

• Functional Data
• An motivating example
• Functional linear models
• General linear hypothesis testing problem
• Some related works
Introduction

Functional data

• In many situations, data are often observed over time for a number of subjects. Examples include Temperature data (Ramsay and Silverman 2003), Progesterone data (Brunback and Rice 1998), Ergonomics data (Faraway 1997), and Children growth curves (Zhang, Liang and Xiao 2010) among others. An observation unit is now a function (a curve or a surface) instead of a vector or a scalar.

• Smoothing techniques are often needed to reconstruct the functions (Ramsay and Silverman 2003). Traditional inferences often fail to apply properly due to the fact that the covariance operator is no longer invertible. New techniques are needed to develop
Introduction

Possible approaches

• **Type 1**: Techniques with dimension-reduction, e.g., basis-based or PCA-based (earlier works for FDA, e.g., Zhang 1999). **Advantage**: traditional inference tools can be applied to dimension-reduced data and can be powerful. **Drawback**: need to determine the number of dimensionality and may lose some information

• **Type 2**: Techniques without involving dimension-reduction, e.g., L2-norm based techniques (Faraway 1997, Shen and Faraway 2004, Zhang and Chen 2007, Zhang, Liang and Xiao 2010, Zhang 2011 among others). **Advantage**: straightforward and no dimension-reduction involved. **Drawback**: may lose of some power
Introduction

Motivating Example: the Ergonomics Data (Shen and Faraway 2004): responses

- Data collected for study of motion of automobile drivers
- Angles formed at the right elbow between the upper and lower arms observed on an equally spaced grid of points over a period of time, rescaled to $[0, 1]$ for convenience
- Number of the time points varies from observation to observation
- 20 locations within a test car for a single subject observed
- At each location, 3 replicates taken
- Totally, 60 angle response curves observed
Introduction

The Ergonomics Data

Figure 1: Six selected curves from the ergonomics data.
Introduction

The Ergonomics Data

Figure 2: The smoothed ergonomics curves.
Introduction

The Ergonomics Data: covariates

- The coordinates \((c_x, c_y, c_z)\) of the target
- \(x\) represents the left to right direction
- \(y\) represents the close to far direction
- \(z\) represents the down to up direction
Introduction

The Ergonomics Data: the model
Shen and Faraway (2004) suggested the following quadratic FLM:

\[ y(t) = \beta_0(t) + c_x\beta_1(t) + c_y\beta_2(t) + c_z\beta_3(t) + c_x^2\beta_4(t) + c_y^2\beta_5(t) + c_z^2\beta_6(t) + c_x c_y\beta_7(t) + c_x c_z\beta_8(t) + c_y c_z\beta_9(t) + \epsilon(t), \]

- Overall test significant?
- Each coefficient function significant?
- How to select the best model?
- How to detect the unusual curves?
Introduction

**Functional Linear Model:**

\[ y(t) = X\beta(t) + \epsilon(t), \; \epsilon(t) \sim \text{GP}(0, \gamma I_n), \; t \in T = [a, b], \]

where

- \( y(t) \): vector of response functions
- \( X \): \( n \times (p + 1) \) full rank design matrix
- \( \beta(t) = [\beta_0(t), \beta_1(t), \cdots, \beta_p(t)]^T \): vector of coefficient functions
- \( \epsilon(t) \): vector of subject-effects
**Introduction**

**General Linear Hypothesis Testing Problem:**

\[ H_0 : \mathbf{C}\beta(t) = \mathbf{c}(t), \ vs \ H_1 : \mathbf{C}\beta(t) \neq \mathbf{c}(t), \]

where

- **\( \mathbf{C} \):** \( q \times (p + 1) \) full rank matrix with \( q < p + 1 \)
- **\( \mathbf{c}(t) = [c_1(t), \cdots, c_q(t)]^T \):** given vector of known functions
- **Example 1:** \( \beta_r(t) = 0 \)? One takes \( \mathbf{C} = \mathbf{e}_{r,p+1}^T \) and \( \mathbf{c}(t) = 0 \)
- **Example 2:** \( \beta_0(t) = \beta_1(t) \)? One takes \( \mathbf{C} = (\mathbf{e}_{1,p+1} - \mathbf{e}_{2,p+1})^T \) and \( \mathbf{c}(t) = 0 \).
- **Here \( \mathbf{e}_{r,k} \) denotes a \( k \)-dim unit vector whose \( r \)-th entry being 1. For example, \( \mathbf{e}_{2,4} = [0, 1, 0, 0]' \).**
Introduction

Some related works:

• An $L^2$-norm based Test for two nested models with bootstrapping (Faraway 1997)
• An F-type Test for two nested models (Shen and Faraway 2004)
• An $L^2$-norm based Test for the GLHT problem (Zhang and Chen 2007)
• An F-type Test for the GLHT problem (Zhang 2011)
• One-Sample Test (Mas 2007)
• Two-sample Test (Hall and Van Keilegom 2007, Zhang, Peng and Zhang 2010)
• One-way ANOVA (Cuevas, Febrero and Fraiman 2004)
• Outlier detection (Bande and Manuel 2006, Shen and Xu 2006)
F-type Test for the GLHT Problem

Outline

• $F$-type test statistic
• Asymptotic expressions derivation
• Asymptotic power derivation
• $F$-type test is root-$n$ consistent
• Null distribution Approximation
F-type Test for the GLHT Problem

F-type test statistic:

\[ F = \frac{(SSE_R - SSE_F)/q}{SSE_F/(n - p - 1)} = \frac{SSH/q}{SSE_F/(n - p - 1)}, \]

where

- \( SSE_R = \int_T \| y(t) - X \hat{\beta}_R(t) \|^2 dt \): SSE under the RM
- \( SSE_F = \int_T \| y(t) - X \hat{\beta}_F(t) \|^2 dt = \int_T y(t)^T(I_n - P_X)y(t)dt \): SSE under the FM, where \( P_X = X(X^TX)^{-1}X^T \).
- \( SSH = \int_T (C\hat{\beta}_F(t) - c)^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta}_F - c)dt \).
- SSH and SSE \(_F\) are independent.
- \( q \) and \( n - p - 1 \) no longer dfs of SSH and SSE \(_F\).
**F-type Test for the GLHT Problem**

**Asymptotic Expressions:**

**Theorem 1** Assume that \( \text{tr}(\gamma) < \infty \) and \( X \) has full rank. Given \( X \), we have

\[
SSH = \sum_{r=1}^{m} \lambda_r A_r + \sum_{r=m+1}^{\infty} \pi_r^2, \quad SSE_F = \sum_{r=1}^{m} \lambda_r B_r,
\]

where

- \( A_r \sim \chi^2_q(\lambda_r^{-1} \pi_r^2) \) and \( B_r \sim \chi^2_{n-p-1} \) are independent
- \( \pi_r^2 = \| \int_T \eta_w(t) \phi_r(t) dt \|^2, \quad r = 1, \cdots, m \)
- \( \eta_w(t) = (C(X^T X)^{-1}C^T)^{-1/2}[C\beta(t) - c(t)] \)
- \( \lambda_1, \lambda_2, \cdots, \lambda_m \): positive eigenvalues of \( \gamma(s,t) \)
- \( \phi_1(t), \phi_2(t), \cdots, \phi_m(t) \): orthonormal eigenfunctions of \( \gamma(s,t) \)
- \( m \): number of the positive eigenvalues of \( \gamma(s,t) \)
F-type Test for the GLHT Problem

Asymptotic Expressions: By the above theorem,

\[ F \overset{d}{=} \frac{\left[ \sum_{r=1}^{m} \lambda_r A_r + \sum_{r=m+1}^{\infty} \pi_r^2 \right]}{\sum_{r=1}^{m} \lambda_r B_r/(n - p - 1)} \]

As \( n \to \infty \), the asymptotic null random expression of \( F \) is

\[ F^* \overset{d}{=} \frac{\sum_{r=1}^{m} \lambda_r A_r/q}{\sum_{r=1}^{m} \lambda_r B_r/(n - p - 1)} \overset{d}{=} \frac{T^*}{q \text{tr}(\gamma)} + o_p(1), \]

where

- \( T^* = \sum_{r=1}^{m} \lambda_r A_r \),
- \( \sum_{r=1}^{m} \lambda_r B_r/(n - p - 1) \to \text{tr}(\gamma) = \sum_{r=1}^{m} \lambda_r \) almost surely,
- \( A_r \overset{i.i.d}{\sim} \chi_q^2, \ B_r \overset{i.i.d}{\sim} \chi_{n-p-1}^2. \)
Asymptotic Power: a sequence of local alternatives:

\[ H_{1n} : C\beta(t) - c(t) = n^{-\tau/2}d(t), \]

where

- \(0 < \tau < 1\) some constant
- \(d(t)\) any fixed vector of functions such that
  \[ 0 < \int_T \|d(t)\|^2dt < \infty \]
- As \(n\) gets large, the local alternatives turn to the null hypothesis.
F-type Test for the GLHT Problem

Asymptotic Power: Assume $n^{-1}X^TX \to \Omega$, invertible. Then under $H_{1n}$, we have $\pi_r^2 = n^{1-\tau} \delta_r^2$ where

$$\delta_r^2 = \| \int_T \left( C\Omega^{-1}C^T \right)^{-1/2} d(t) \phi_r(t) dt \|^2 [1 + o(1)], r = 1, 2, \ldots .$$

Three possible cases for $m$ and $\delta_r^2$:

- $m < \infty$ and $\delta_r^2 = 0$ for all $r \in \{1, 2, \cdots , m\}$
- $m < \infty$ and $\delta_r^2 \neq 0$ for at least one $r \in \{1, 2, \cdots , m\}$,
- $m = \infty$.

Can show that the F-type test is root $n$-consistent under the 3 cases.
F-type Test for the GLHT Problem

**Asymptotic Power:** Assumption A

(A1) $0 < \text{tr}(\gamma) < \infty$.

(A2) As $n \to \infty$, $n^{-1}X^TX \to \Omega$, invertible.

(A3) $0 < \tau < 1$ and $0 < \delta^2 < \infty$ where $\delta^2 = \sum_{r=1}^{\infty} \delta_r^2$.

**Theorem 2** For Case 1, under Assumption A, the asymptotic power of $F$ is

$$P(F \geq F^*_{\alpha}|H_{1n}) = P(T^* > T^*_\alpha - n^{1-\tau}\delta^2) + o(1),$$

which tends to 1 as $n \to \infty$.

- $F^*$: asymptotic null distr of $F_n$
- $T^* = \sum_{r=1}^{m} \lambda_r A_r$, $A_r \sim \chi^2_q$,
- $F^*_{\alpha}$ and $T^*_\alpha$: upper 100$\alpha$ percentiles of $F^*$ and $T^*$
F-type Test for the GLHT Problem

Asymptotic Power under Cases 2 and 3:

**Theorem 3** For Cases 2 and 3, under Assumption A, as \( n \to \infty \), we have

\[
\frac{F - n^{1-\tau} \delta^2/[qtr(\gamma)]}{2n^{(1-\tau)/2} \delta \lambda /[qtr(\gamma)]} \xrightarrow{L} N(0, 1).
\]

In addition, the asymptotic power of \( F \) is

\[
P (F \geq F_\alpha^*|H_{1n}) = \Phi \left( \frac{n^{(1-\tau)/2} \delta^2}{2\delta \lambda} \right) + o(1),
\]

which tends to 1 as \( n \to \infty \), where

- \( \Phi(\cdot) \) denotes the cdf of \( N(0, 1) \)
- \( \delta^2 = \sum_{r=1}^{\infty} \lambda_r \delta_r^2 \leq \lambda_{\text{max}} \delta^2 \).
**F-type Test for the GLHT Problem**

**Null Distribution Approximation:** Under $H_0$,

$$
F = \frac{\text{SSH}/q}{\text{SSE}_F/(n-p-1)} \approx F_{q\kappa,(n-p-1)\kappa}.
$$

Under $H_0$, by Welch-Satterthwaite approximation, one has

- $\text{SSH} = \sum_{r=1}^{m} \lambda_r A_r$, $A_r \overset{i.i.d}{\sim} \chi^2_q$, approximated by $R_1 = \alpha_1 \chi^2_{d_1}$ via matching two cumulants of SSH and $R_1$ such that

$$
\alpha_1 = \frac{\sum_{r=1}^{m} \lambda_r^2}{\sum_{r=1}^{m} \lambda_r}, \quad d_1 = q\kappa, \quad \kappa = \frac{(\sum_{r=1}^{m} \lambda_r)^2}{\sum_{r=1}^{m} \lambda_r^2},
$$

- $\text{SSE}_F = \sum_{r=1}^{m} \lambda_r B_r$, $B_r \overset{i.i.d}{\sim} \chi^2_{n-p-1}$, approximated by $R_2 = \alpha_2 \chi^2_{d_2}$ via matching two cumulants of SSH and $R_2$ such that

$$
\alpha_2 = \frac{\sum_{r=1}^{m} \lambda_r^2}{\sum_{r=1}^{m} \lambda_r} = \alpha_1, \quad d_2 = (n-p-1)\kappa,
$$
F-type Test for the GLHT Problem

Null Distribution Approximation:

\[ \kappa = \frac{\left(\sum_{r=1}^{m} \lambda_r \right)^2}{\sum_{r=1}^{m} \lambda_r^2} = \frac{\left(\sum_{r=1}^{\infty} \lambda_r \right)^2}{\sum_{r=1}^{\infty} \lambda_r^2} = \frac{\text{tr}^2(\gamma)}{\text{tr}(\gamma \otimes 2)} \]

where

- \( \gamma \otimes 2(s, t) = \int_T \gamma(s, u)\gamma(u, t)du \), cross-square of \( \gamma(s, t) \),
  generalization of the product of two matrices

- \( \text{tr}(\gamma) = \sum_{r=1}^{\infty} \lambda_r \)

- \( \text{tr}(\gamma \otimes 2) = \sum_{r=1}^{\infty} \lambda_r^2 \)

- No need to estimate \( m \)

- No need to estimate the eigenvalues of \( \gamma(s, t) \)
F-type Test for the GLHT Problem

Null Distribution Approximation: estimation of $\kappa$

The naive method

- Estimate $\gamma(s, t)$ by its unbiased estimator

$$\hat{\gamma}(s, t) = (n - p - 1)^{-1} \sum_{i=1}^{n} (y_i(s) - \hat{y}_i(s))(y_i(t) - \hat{y}_i(t))$$

- Substitute $\text{tr}^2(\gamma), \text{tr}(\gamma^{\otimes 2})$ by $\text{tr}^2(\hat{\gamma}), \text{tr}(\hat{\gamma}^{\otimes 2})$

- Drawback: $\text{tr}^2(\hat{\gamma}), \text{tr}(\hat{\gamma}^{\otimes 2})$ are biased.
Special Cases of the GLHT Problem

Outline

- Overall test
- Test of an individual coefficient function
- Test of a subset of coefficient functions
- One-sample test
- Two-sample test with a common covariance function
- Multi-sample test with a common covariance function

By properly choosing $y(t)$, $X$, $\beta(t)$ and $C$, $c(t)$, the GLHT problem reduces to any of the above special cases.
Special Cases of the GLHT Problem

**Overall test:** When $C = [0, I_p]$ and $c(t) = 0$, the GLHT problem reduces to

$$H_0 : y(t) = 1_n \beta_0(t) + v(t), \text{ vs } H_1 : y(t) = X \beta(t) + v(t),$$

**The associated test statistic**

$$F = \frac{\sum_{i=1}^{n} \int (\hat{y}_i(t) - \bar{y}(t))^2 dt / p}{\sum_{i=1}^{n} \int_T (y_i(t) - \hat{y}_i(t))^2 dt / (n - p - 1)}$$

**Under $H_0$**

$$F \overset{approx.}{\sim} F_{p \kappa, (n - p - 1) \kappa}.$$
Special Cases of the GLHT Problem

**Overall test:** Application to the ergonomics data.

**Reconstruction of the response functions**

- The angle curves fitted using a quadratic regression spline with 7 equally spaced inner knots
- The number of knots was selected by GCV (Zhang and Chen 2007).
- The angle curves evaluated at a grid of $m = 1000$ equally spaced time points over $[0, 1]$. 
Special Cases of the GLHT Problem

**Overall test:** Application to the ergonomics data.

Table 1: Overall test for the ergonomics data with $m = 1000$. The estimated $\kappa = 1.69$ and the 95% critical value for the $F$-type test is $F(.95, 15, 83) = 1.79$.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>81885</td>
<td>15</td>
<td>5459</td>
<td>31.84</td>
<td>0</td>
</tr>
<tr>
<td>Residual</td>
<td>14228</td>
<td>83</td>
<td>171.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>96113</td>
<td>98</td>
<td>980.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Special Cases of the GLHT Problem

**Test of an individual coefficient function** When \( C = e_{j,p+1}^T \) and \( c(t) = 0 \), the GLHT problem reduces to

\[
H_0 : \beta_j(t) = 0, \quad \text{vs} \quad H_1 : \beta_j(t) \neq 0.
\]

The associated test statistic

\[
F_j = \frac{\int_T \hat{\beta}_j^2(t) dt / e_{j,p+1}^T (X^T X)^{-1} e_{j,p+1}}{\sum_{i=1}^n \int_T [y_i(t) - \hat{y}_i(t)]^2 dt / (n - p - 1)}
\]

**Under** \( H_0 \)

\[
F_j \overset{\text{approx.}}{\sim} F_{\kappa, (n-p-1)\kappa}.
\]

- Strongly related to Variable Selection for FLM.
Special Cases of the GLHT Problem

**Test of an individual coefficient function** Application to the ergonomics data.

Table 2: Coefficient table for the ergonomics data with $m = 1000$. The 95% critical value is $F(.95, 1.69, 83) = 3.29$.

<table>
<thead>
<tr>
<th>Estimated Coef.</th>
<th>F</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_0(t)$</td>
<td>1813</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{\beta}_1(t)$</td>
<td>10.19</td>
<td>$1.10 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\hat{\beta}_2(t)$</td>
<td>20.79</td>
<td>$4.78 \times 10^{-8}$</td>
</tr>
<tr>
<td>$\hat{\beta}_3(t)$</td>
<td>6.17</td>
<td>$3.15 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\hat{\beta}_4(t)$</td>
<td>5.98</td>
<td>$3.74 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\hat{\beta}_5(t)$</td>
<td>8.74</td>
<td>$3.56 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\hat{\beta}_6(t)$</td>
<td>2.73</td>
<td>$7.08 \times 10^{-2}$</td>
</tr>
<tr>
<td>$\hat{\beta}_7(t)$</td>
<td>6.63</td>
<td>$2.12 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\hat{\beta}_8(t)$</td>
<td>1.12</td>
<td>$3.30 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\hat{\beta}_9(t)$</td>
<td>4.13</td>
<td>$1.95 \times 10^{-2}$</td>
</tr>
</tbody>
</table>
Special Cases of the GLHT Problem

Test of a subset of coefficient functions When $C = [0_{q \times (p-q+1)}, I_q]$ and $c(t) = 0$, the GLHT problem reduces to

$$H_0 : y(t) = X_1 \beta_1(t) + \nu(t), \text{ vs } H_1 : y(t) = X \beta(t) + \nu(t),$$

The associated test statistic

$$F = \frac{\int_T y(t)^T (P_X - P_{X_1}) y(t) dt / q}{\int_T y(t)^T (I_n - P_X) y(t) dt / (n - p - 1)}$$

Under $H_0$

$$F \overset{approx}{\sim} F_{q \kappa, (n-p-1) \kappa}.$$  

• Considered by Faraway (1997) by a bootstrap method.
• Considered by Shen and Faraway (2004) by an F-type test.
Special Cases of the GLHT Problem

**One-Sample test** Given a sample $y_1(t), \cdots, y_n(t) \sim GP(\eta, \gamma)$, test $H_0 : \eta(t) = \eta_0(t)$, vs $H_1 : \eta(t) \neq \eta_0(t)$. The associated GLHT problem

$$H_0 : y(t) = v(t), \text{ vs } H_1 : y(t) = 1_n \eta(t) + v(t).$$

The associated test statistic

$$F = \frac{n \int_T (\bar{y}(t) - \eta_0(t))^2 dt}{\sum_{i=1}^n \int_T (y_i(t) - \bar{y}(t))^2 dt / (n - 1)}.$$

**Under $H_0$**

$$F \approx F_{\kappa, (n-1)\kappa}.$$

- $\hat{\gamma}(s, t) = (n - 1)^{-1} \sum_{i=1}^n (y_i(s) - \bar{y}(s))(y_i(t) - \bar{y}(t))$. 
Special Cases of the GLHT Problem

Two-sample test Given $y_{i1}(t), \cdots, y_{in_i}(t) \sim \text{GP}(\eta_i, \gamma), i = 1, 2$, test: $H_0 : \eta_1(t) = \eta_2(t)$, vs $H_1 : \eta_1(t) \neq \eta_2(t)$. The GLHT problem

$\mathbf{y}(t) = \mathbf{1}_n \eta(t) + \mathbf{v}(t)$, vs $H_1 : \mathbf{y}(t) = \text{diag} (\mathbf{1}_{n_1} \mathbf{1}_{n_2}) [\eta_1(t), \eta_2(t)]^T + \mathbf{v}(t)$

The associated test statistic

$$F = \frac{\frac{n_1 n_2}{n} \int_T (\bar{y}_1(t) - \bar{y}_2(t))^2 dt}{\sum_{i=1}^2 \sum_{j=1}^{n_i} \int_T (y_{ij}(t) - \bar{y}_i(t))^2 dt / (n - 2)}$$

Under $H_0$

$$F \overset{\text{approx.}}{\sim} F_{\kappa, (n-2)\kappa}.$$ 

• Considered by Hall and Van Keilegolom (2007), Zhang, Peng and Zhang (2010)

• $\hat{\gamma}(s, t) = (n - 2)^{-1} \sum_{i=1}^2 \sum_{j=1}^{n_i} (y_{ij}(s) - \bar{y}_i(s))(y_{ij}(t) - \bar{y}_i(t))$. 
Special Cases of the GLHT Problem

**Multi-sample test** Given $q$ samples

$y_{i1}(t), \ldots, y_{in_i}(t) \sim \text{GP}(\eta_i, \gamma), i = 1, \ldots, q$, test

$H_0 : \eta_1(t) = \eta_2(t) = \ldots = \eta_q(t)$. The GLHT problem

$y(t) = 1_n \eta(t) + \nu(t)$, vs $H_1 : y(t) = \text{diag}(1_{n_1}, \ldots, 1_{n_q})[\eta_1(t), \ldots, \eta_q(t)]^T + \nu(t)$

**The associated test statistic**

$$F = \frac{\sum_{i=1}^q n_i \int_T (\bar{y}_i(t) - \bar{y}(t))^2 dt / (q - 1)}{\sum_{i=1}^q \sum_{j=1}^{n_i} \int_T (y_{ij}(t) - \bar{y}_i(t))^2 dt / (n - q)}$$

**Under $H_0$**

$$F \overset{\text{approx.}}{\sim} F_{(q-1)\kappa, (n-q)\kappa}.$$ 

- One-way ANOVA considered by Cuevas, Febrero and Fraiman (2004) by a bootstrap method.

- $\hat{\gamma}(s, t) = (n - q)^{-1} \sum_{i=1}^q \sum_{j=1}^{n_i} (y_{ij}(s) - \bar{y}_i(s))(y_{ij}(t) - \bar{y}_i(t))$. 
Variable Selection for FLM

Aims:

- To get parsimonious regression models
- To get good prediction regression models

Related Questions:

- Which variable should be included?
- In what form, $X, X^2$ or other forms?
- Which variable should be excluded from the model?
- Which criterion can be used?
Variable Selection for FLM

**Methods:**

- **Forward selection:** introduce the variables one after another
- **Backward elimination:** delete the variables one after another
- **Stepwise:** a mixture of FS and BE
Variable Selection for FLM

The related F-type test

\[ H_0 : \beta_j(t) = 0, \text{ vs } H_1 : \beta_j(t) \neq 0. \]

\[ F_j = \frac{\int_T \hat{\beta}_j^2(t) dt / e_{j,p+1}^T (X^T X)^{-1} e_{j,p+1}}{\sum_{i=1}^n \int_T [y_i(t) - \hat{y}_i(t)]^2 dt / (n - p - 1)} \approx F_{\kappa,(n-p-1)\kappa}. \]

- The cutoff values can be specified as \( F_{in} = F_{\kappa,(n-p-1)\kappa}(\alpha_{in}) \) and \( F_{out} = F_{\kappa,(n-p-1)\kappa}(\alpha_{out}) \). For example, \( \alpha_{in} = .05 \) and \( \alpha_{out} = .05 \).
Variable Selection for FLM

Application to the ergonomics data

Table 3: P-values for variable selection for the ergonomics data with $m = 1000$.

<table>
<thead>
<tr>
<th>Estimated Coef. function</th>
<th>P-values for Step</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$\hat{\beta}_0(t)$</td>
<td>0</td>
</tr>
<tr>
<td>$\hat{\beta}_1(t)$</td>
<td>$1.10 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\hat{\beta}_2(t)$</td>
<td>$4.78 \times 10^{-8}$</td>
</tr>
<tr>
<td>$\hat{\beta}_3(t)$</td>
<td>$3.15 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\hat{\beta}_4(t)$</td>
<td>$3.74 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\hat{\beta}_5(t)$</td>
<td>$3.56 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\hat{\beta}_6(t)$</td>
<td>$7.08 \times 10^{-2}$</td>
</tr>
<tr>
<td>$\hat{\beta}_7(t)$</td>
<td>$2.12 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\hat{\beta}_8(t)$</td>
<td>$3.30 \times 10^{-1}$</td>
</tr>
<tr>
<td>$\hat{\beta}_9(t)$</td>
<td>$1.95 \times 10^{-2}$</td>
</tr>
</tbody>
</table>
Outlier Detection for FLM

Motivation and Definition:

• Outliers can be frequently found not only in observational studies but also in designed experiments.

• According to Bande and Manuel (2006), a curve considered as an outlier if it has a different distribution from the rest of curves, which are assumed to be identically distributed.

• Outliers may be curves with gross errors such as measurement, recording and typing mistakes. These errors should be identified and corrected whenever possible.

• Outliers may be real curves in the sense that they are not due to gross errors but are somehow suspicious or surprising since they do not follow the same pattern as the rest of curves.
Outlier Detection for FLM

Motivation and Definition:

- Outliers may bias our functional estimates, which we would like to prevent.
- Analysis with outliers often leads to misleading conclusions.
- Some methods should be found to detect and examine these surprising, unusual curves, known as outliers.
Outlier Detection for FLM

Existing approaches:

• Bande and Manuel (2006) proposed a bootstrap-based approach which is time-consuming.

• Shen and Xu (2006) proposed two approaches, based on standardized residual curves and Jackknife residual curves respectively.

• Here we state the Jackknife method which is based on the Jackknife residual curves.
Outlier Detection for FLM

Jackknife method for FLM The Jackknife score is defined as

\[ J_i^2 = \frac{\|y_i - \hat{y}_i^{(-i)}\|^2}{(1 - h_{ii})^{-1} \text{tr}(\hat{\gamma}^{(-i)})} \]

\[ , i = 1, 2, \cdots, n, \]

- \( \hat{y}_i^{(-i)}(t) \): i-th fitted response curve when curve i excluded.
- \( \hat{\gamma}^{(-i)}(s, t) \): estimated covar. funct. when curve i excluded.
- When curve i is NOT an outlier, \( y_i(t) - \hat{y}_i^{(-i)}(t) \sim \text{GP}(0, \gamma_i) \),
  \( \gamma_i(s, t) = (1 + x_i^T (X^{(-i)T} X^{(-i)})^{-1} x_i) \gamma(s, t) = (1 - h_{ii})^{-1} \gamma(s, t) \)
  where \( X^{(-i)} \) is the design matrix when curve i excluded.
- \( \text{E} \|y_i - y_i^{(-i)}\|^2 = (1 - h_{ii})^{-1} \text{tr}(\gamma) \) where \( h_{ii} \) the i-th diagonal entry of \( P_X = X(X^T X)^{-1} X^T \).
Outlier Detection for FLM

Jackknife method for FLM

\[ J_i^2 \overset{d}{=} \frac{\sum_{r=1}^{\infty} \lambda_r A_r}{\sum_{r=1}^{\infty} \lambda_r B_r} \sim F_{\kappa,(n-p-2)\kappa} \] approximately,

where \( \lambda_r \): \( r \)-th eigenvalue of \( \gamma(s,t) \) and \( A_r \sim \chi^2_1, B_r \sim \chi^2_{n-p-2} \) and \( \kappa = \frac{\text{tr}^2(\gamma)}{\text{tr}(\gamma \otimes 2)} \).

- When curve \( i \) is NOT an outlier,
  \[ \|y_i - y_i^{(-i)}\|^2 (1 - h_{ii}) \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r A_r. \]

- since \( (n - p - 2)\hat{\gamma}^{(-i)}(s,t) \sim WP(n - p - 2, \gamma) \),
  \( (n - p - 2)\text{tr}(\hat{\gamma}^{(-i)}) \overset{d}{=} \sum_{r=1}^{\infty} \lambda_r B_r. \)

- \( y_i(t) - \hat{y}_i^{(-i)}(t) \) and \( \hat{\gamma}^{(-i)}(s,t) \) are independent.
Outlier Detection for FLM

Calculation for the Jackknife method By some simple calculation, Shen and Xu (2006) showed that

\[ J_i^2 = \frac{(n - p - 1)S_i^2}{n - p - S_i^2} \] where \( S_i^2 = \frac{\|\hat{\epsilon}_i\|^2}{\text{tr}(\hat{\gamma})(1 - h_{ii})} \),

- \( \hat{\epsilon}_i(t) \): \( i \)-th residual curve.
- \( \hat{\gamma}(s, t) \): estimated covariance function using all the data.
- Cut off value for \( J_i^2 \) is \( F_{\hat{\kappa},(n-p-2)\hat{\kappa}}(1 - \alpha) \) approximately for any given \( \alpha \) where \( \hat{\kappa} = \frac{\text{tr}^2(\hat{\gamma})}{\text{tr}(\hat{\gamma} \otimes 2)} \).
Outlier Detection for FLM

Application to the ergonomics data

Figure 3: Outlier detection for the ergonomics data.
Outlier Detection for FLM

Application to the ergonomics data

Figure 4: Unusual curves in the smoothed ergonomics data.
Summary and Discussion

• An F-type test and its applications (ANOVA, Variable Selection and Outlier detection) discussed.

• The methodologies considered are not dimension reduction based.

• Dimension reduction based methods can also be considered.

• Further studies are warranted.
THANK YOU