Two Sample Problem for Functional Data

Radek Hendrych

Stochastic Modelling in Economics and Finance 1
November 25, 2013
Outline

1. Introduction

2. Testing Equality of Mean Functions

3. Testing Equality of Covariance Operators

4. Bibliography
Two Sample Problem for Functional Data

- testing the equality of the means in two independent samples
- testing the equality of the covariance operators in two independent samples

Asymptotic procedures will be introduced.
Outline

1. Introduction

2. Testing Equality of Mean Functions

3. Testing Equality of Covariance Operators

4. Bibliography
Testing Equality of Mean Functions

Model

Consider two samples $X_1, \ldots, X_N$ and $X_1^*, \ldots, X_M^*$ satisfying the model

\begin{align*}
X_i(t) &= \mu(t) + \varepsilon_i(t), \quad i = 1 \ldots, N, \\
X_j^*(t) &= \mu^*(t) + \varepsilon_j^*(t), \quad j = 1 \ldots, M.
\end{align*}

Assumptions

(A1) the two samples are independent

(A2) $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. with $E\varepsilon_1(t) = 0$ and $E||\varepsilon_1||^4 < \infty$

(A3) $\varepsilon_1^*, \ldots, \varepsilon_M^*$ are i.i.d. with $E\varepsilon_1^*(t) = 0$ and $E||\varepsilon_1^*||^4 < \infty$

The $\varepsilon_1$ and $\varepsilon_1^*$ do not have to follow the same distribution.
Main Goal

Testing Hypothesis

\[ H_0 : \mu = \mu^* \text{ in } L^2 \text{ against } H_1 : \neg H_0 \]
Method I

\[ \bar{X}_N(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t) \quad \text{and} \quad \bar{X}_M^*(t) = \frac{1}{M} \sum_{j=1}^{M} X_j^*(t) \]

are unbiased estimators for \( \mu(t) \) and \( \mu^*(t) \), respectively.

It is natural to reject the null hypothesis if

\[ U_{N,M} = \frac{NM}{N + M} \int_0^1 (\bar{X}_N(t) - \bar{X}_M^*(t))^2 \, dt \]  \hspace{1cm} (3)

is large.
Convergence of $U_{N,M}$ under $H_0$

**Theorem**

If $H_0$ and the assumptions (A1), (A2) and (A3) hold, and

$$
\frac{N}{N + M} \to \theta, \quad \text{for some } \theta \in [0, 1], \quad \text{as } N \to \infty,
$$

then

$$
U_{N,M} \xrightarrow{d} \int_0^1 \Gamma^2(t) dt, \quad N, M \to \infty,
$$

(4)

where $\{\Gamma(t), t \in [0, 1]\}$ is a Gaussian process satisfying $\mathbb{E}\Gamma(t) = 0$ and

$$
\mathbb{E} [\Gamma(t)\Gamma(s)] = (1 - \theta)c(t, s) + \theta c^*(t, s),
$$

with $c(t, s) = \text{cov}(X_1(t), X_1(s))$ and $c^*(t, s) = \text{cov}(X_1^*(t), X_1^*(s))$.

*Proof.* See [1].
The limit distribution of $U_{N,M}$ in (4) depends on the unknown covariance functions $c$ and $c^*$.

According to the Karhunen-Loève expansion, one can suppose that

$$
\Gamma(t) = \sum_{k=1}^{\infty} \tau_k^{1/2} N_k \varphi_k(t),
$$

where $N_k$, $k \in \mathbb{N}$, are independent $N(0, 1)$ random variables, $\tau_1 \geq \tau_2 \geq \ldots$ and $\varphi_1, \varphi_2, \ldots$ are the eigenvalues and eigenfunctions of the operator determined by $(1 - \theta)c + \theta c^*$. 
Since
\[ \int_0^1 \Gamma^2(t) dt = \sum_{k=1}^{\infty} \tau_k N_k^2, \]
to provide a reasonable approximation for \( \int_0^1 \Gamma^2(t) dt \), one only need to estimate \( \tau_k \).

This can be done using \( \hat{\tau}_k \), the eigenvalues of the empirical covariance function

\[
\hat{z}_{N,M}(t, s) = \frac{M}{N+M} \frac{1}{N} \sum_{i=1}^{N} (X_i(t) - \bar{X}_N(t)) (X_i(s) - \bar{X}_N(s)) \\
+ \frac{N}{N+M} \frac{1}{M} \sum_{j=1}^{M} (X^*_j(t) - \bar{X}_M^*(t)) (X^*_j(s) - \bar{X}_M^*(s)).
\]

Thus, the sum \( \sum_{k=1}^{d} \hat{\tau}_k N_k^2 \) offers an approximation to the limit distribution in (4) if \( d \) is large enough.
Asymptotic Consistency of Method I

Theorem

If the assumptions (A1), (A2) and (A3) hold,

\[
\frac{N}{N + M} \xrightarrow{\text{as } N \to \infty} \theta, \quad \text{for some } \theta \in [0, 1], \quad \text{as } N \to \infty,
\]

and

\[
\int_0^1 (\mu(t) - \mu^*(t))^2 \, dt > 0,
\]

then \( U_{N,M} \xrightarrow{P} \infty, \text{ as } N, M \to \infty. \)

Proof. See [1].
Method II

The method is a *projection version* of the first procedure based on $U_{N,M}$.

It does not require the numerical evaluation of the integral in the definition of $U_{N,M}$ (⇒ an easier implementation).

Consider projections onto the space determined by the leading eigenfunctions of the operator $Z = (1 - \theta)C + \theta C^*$.  

In particular, assume that the eigenvalues of $Z$ satisfy

$$\tau_1 > \tau_2 > \cdots > \tau_d > \tau_{d+1}. \quad (5)$$

One wants to project observations onto the space spanned by $\varphi_1, \ldots, \varphi_d$, i.e. the corresponding eigenfunctions.
In fact, the functions $\varphi_1, \ldots, \varphi_d$ are unknown.

The corresponding eigenfunctions of $\hat{Z}_{N,M}$, denoted by $\hat{\varphi}_i$, are used.

This delivers a projection of $\bar{X}_N - \bar{X}_M^*$ into the linear space spanned by $\hat{\varphi}_1, \ldots, \hat{\varphi}_d$. Let

$$\hat{a} = (\hat{a}_1, \ldots, \hat{a}_d)^\top,$$

where $\hat{a}_i = \langle \bar{X}_N - \bar{X}_M^*, \hat{\varphi}_i \rangle$.

Under the conditions of the first theorem, it can be shown that

$$\sqrt{NM/(N+M)}\hat{a}$$

has approximately $d$-variate normal distribution (up to some random signs) with the asymptotic variance $Q = \{Q(i,j)\}_{i,j=1}^d$,

$$Q(i,j) = (1-\theta)\mathbb{E}\langle X_1 - \mu, \varphi_i \rangle\langle X_1 - \mu, \varphi_j \rangle + \theta\mathbb{E}\langle X_1^* - \mu^*, \varphi_i \rangle\langle X_1^* - \mu^*, \varphi_j \rangle.$$

Thus, $Q(i,i) = \tau_i$ and $Q(i,j) \neq i \neq j = 0$. 

Radek Hendrych

Two Sample Problem for Functional Data
According to the mentioned facts, testing procedures can be based on

\[ T_{N,M}^{(1)} = \frac{NM}{N + M} \sum_{k=1}^{d} \frac{\hat{a}_k^2}{\hat{\tau}_k}, \]  

(6)

\[ T_{N,M}^{(2)} = \frac{NM}{N + M} \sum_{k=1}^{d} \hat{a}_k^2. \]  

(7)
Convergence of $T^{(1)}_{N,M}$ and $T^{(2)}_{N,M}$ under $H_0$

**Theorem**

If $H_0$, the assumptions (A1), (A2), (A3) and (5) hold, and

$$\frac{N}{N+M} \rightarrow \theta, \quad \text{for some } \theta \in [0,1], \text{ as } N \rightarrow \infty,$$

then

$$T^{(1)}_{N,M} \xrightarrow{d} \chi^2_d \quad \text{and} \quad T^{(2)}_{N,M} \xrightarrow{d} \sum_{k=1}^{d} \tau_k N_k^2, \quad N, M \rightarrow \infty,$$

where $N_1, \ldots, N_k$ are independent standard normal random variables.

**Proof.** See [1].

$T^{(2)}_{N,M}$ is a projection version of $U_{N,M}$, where only first $d$ terms in $L^2$ expansion of $\tilde{X}_N - \tilde{X}_M^*$ are used. The statistic $T^{(1)}_{N,M}$ is an asymptotically distribution free modification of $T^{(2)}_{N,M}$ (and hence of $U_{N,M}$).
Asymptotic Consistency of Method II

**Theorem**

If the assumptions (A1), (A2), (A3) and (5) hold,

\[
\frac{N}{N + M} \to \theta, \quad \text{for some } \theta \in [0, 1], \text{ as } N \to \infty,
\]

and \( \mu - \mu^* \) is not orthogonal to the linear span of \( \varphi_1, \ldots, \varphi_d \), then

\[
T_{N,M}^{(1)} \overset{P}{\to} \infty \quad \text{and} \quad T_{N,M}^{(2)} \overset{P}{\to} \infty, \quad \text{as } N, M \to \infty.
\]

**Proof.** See [1].
Empirical Example

Data

- The data set consists of egg-laying trajectories of Mediterranean fruit flies (medflies).
- Consider 534 egg-laying curves of flies who lived at least 30 days.
- Each function is defined over an interval \([0, 30]\).
- Its value on day \(t \in [0, 30]\) is the number of eggs laid by the fly \(i\) on that day.
- The 534 medflies are classified into:
  - 256 short-lived (died before the end of the 44th day after birth),
  - 278 long-lived (lived longer than 44 days).

Two Samples

- \(X_i(t), \ t \in [0, 30], \ i = 1, \ldots, 256\) (for short-lived medflies)
- \(X_j^*(t), \ t \in [0, 30], \ j = 1, \ldots, 278\) (for long-lived medflies)
**Figure:** Randomly selected egg-laying curves for short- and long-lived medflies.
Figure: The estimated mean functions for the medfly data: short-lived by the solid line, long-lived by the dashed line.
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{256,278}^{(1)}$</td>
<td>1.0</td>
<td>2.2</td>
<td>3.0</td>
<td>5.7</td>
<td>10.3</td>
<td>15.3</td>
<td>3.2</td>
<td>2.7</td>
<td>5.0</td>
</tr>
<tr>
<td>$T_{256,278}^{(2)}$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.1</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**Table:** The $p$-values (in %) of the tests based on the statistics $T_{256,278}^{(1)}$ and $T_{256,278}^{(2)}$ applied to the medfly data.

The $p$-values for $T_{256,278}^{(2)}$ are much more stable.

The test based on $T_{N,M}^{(1)}$ is easier to apply (due to the standard chi-squared critical values), the test based on $T_{N,M}^{(2)}$ may be more reliable.
Outline

1. Introduction

2. Testing Equality of Mean Functions

3. Testing Equality of Covariance Operators

4. Bibliography
Model Assumptions

Consider two samples $X_1, \ldots, X_N$ and $X_1^*, \ldots, X_M^*$ satisfying:

(A1) the two samples are independent,

(A2) $X_1, \ldots, X_N$ are i.i.d. elements of $L^2$ with $\mathbb{E}X_1(t) = 0$,

(A3) $X_1^*, \ldots, X_M^*$ are i.i.d. elements of $L^2$ with $\mathbb{E}X_1^*(t) = 0$.

The $X_1$ and $X_1^*$ do not have to follow the same distribution.
Main Goal

Testing Hypothesis

Suppose the variance operators

\[ C(x) = \mathbb{E}[\langle X, x \rangle X], \quad C^*(x) = \mathbb{E}[\langle X^*, x \rangle X^*], \quad x \in L^2, \]

where \( X \) has the same distribution as \( X_1 \), and \( X^* \) has the same distribution as \( X_1^* \).

One wants to test

\[ H_0 : C = C^* \quad \text{against} \quad H_1 : \neg H_0. \]
Let \( \hat{R} \) be the empirical covariance operator of the pooled data, i.e.

\[
\hat{R}(x) = \frac{1}{N+M} \left[ \sum_{i=1}^{N} \langle X_i, x \rangle X_i + \sum_{j=1}^{M} \langle X_j^*, x \rangle X_j^* \right] = \hat{\theta} \hat{C}(x) + (1 - \hat{\theta}) \hat{C}^*(x),
\]

where \( \hat{C} \) and \( \hat{C}^* \) are the empirical counterparts of \( C \) and \( C^* \), \( x \in L^2 \), and \( \hat{\theta} = \frac{N}{N+M} \).

The operator \( \hat{R} \) has \( N + M \) eigenfunctions (denoted by \( \hat{\phi}_k \)).

Set

\[
\hat{\lambda}_k = \frac{1}{N} \sum_{i=1}^{N} \langle X_i, \hat{\phi}_k \rangle^2 \quad \text{and} \quad \hat{\lambda}_k^* = \frac{1}{M} \sum_{j=1}^{M} \langle X_j^*, \hat{\phi}_k \rangle^2.
\]

These are the sample variances of the Fourier coefficients of \( X \) and \( X^* \) with respect to the orthonormal system \( \{\hat{\phi}_k, k = 1, \ldots, N + M\} \).
Convergence of $\hat{T}$ under $H_0$ and Gaussianity

The test statistic $\hat{T}$ is defined by

$$\hat{T} = \frac{N + M}{2} \hat{\theta} (1 - \hat{\theta}) \sum_{i, j=1}^{N+M} \frac{\langle (\hat{C} - \hat{C}^*) \hat{\phi}_i, \hat{\phi}_j \rangle^2}{(\hat{\theta} \hat{\lambda}_i + (1 - \hat{\theta}) \hat{\lambda}_i^*) (\hat{\theta} \hat{\lambda}_j + (1 - \hat{\theta}) \hat{\lambda}_j^*)}.$$

**Theorem**

Suppose $X$ and $X^*$ are Gaussian elements of $L^2$ such that $\mathbb{E}||X||^4 < \infty$ and $\mathbb{E}||X^*||^4 < \infty$. Suppose also that $\hat{\theta} \to \theta \in (0, 1)$, as $N \to \infty$. If $H_0$ holds, then

$$\hat{T} \overset{d}{\to} \chi^2_{(N+M)(N+M+1)/2}, \quad N, M \to \infty. \quad (9)$$

**Proof.** See [1].
Outline

1. Introduction
2. Testing Equality of Mean Functions
3. Testing Equality of Covariance Operators
4. Bibliography
Thank you for your attention.