

**Politechnika Gdańska**  
Wydział Fizyki Technicznej i Matematyki Stosowanej

**Anna Wieżel**

Nr albumu: 132540

# **Funkcjonalne Modele Liniowe**

**Praca magisterska**  
**na kierunku MATEMATYKA**  
**w zakresie MATEMATYKA FINANSOWA**

Praca wykonana pod kierunkiem  
**dra hab. Karola Dziedziula**  
Katedra Analizy Matematycznej i Numerycznej

Wrzesień 2015

## **Oświadczenie kierującego prac**

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data

Podpis kierującego prac

## **Oświadczenie autora (autorów) pracy**

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data

Podpis autora (autorów) pracy

## **Streszczenie**

coś

## **Słowa kluczowe**

funkcjonalna analiza danych, dane funkcyjne, funkcyjne modele liniowe, test istotności

## **Dziedzina pracy (kody wg programu Socrates-Erasmus)**

11.1 Matematyka

11.2 Statystyka

## **Klasyfikacja tematyczna**

62 Statistics

62-07 Data analysis

62J12 Generalized linear models

## **Tytuł pracy w języku angielskim**

Functional Linear Models



# Spis treści

<b>Wstęp</b>	5
<b>1. Preliminaria</b>	7
1.1. Klasyfikacja operatorów liniowych	7
1.2. Przestrzeń $L^2$	9
1.3. Zmienne funkcyjne w $L^2$ . Pojęcie operatora kowariancji	9
1.4. Funkcyjny model liniowy	10
<b>2. Test istotności w funkcyjnym modelu liniowym</b>	13
2.1. Procedura testowa	13
2.2. Formalne podstawy	15
<b>3. Przykład zastosowania</b>	19
<b>A. Kod w R</b>	21
<b>Bibliografia</b>	23



# Wstęp

Odpowiednik testu istotności dla prostego modelu regresji = F-test (+ t-test) [patrz: artykuł]





# Rozdział 1

## Preliminaria

Przestrzenią funkcyjną  $E$  nazywać będziemy przestrzeń liniową funkcji z dowolnego zbioru  $A$  do zbioru  $B$ .

**Definicja 1.0.1** [Ferraty, Vieu]

Zmienną losową  $X$  nazywamy **zmienną funkcjonalną** wtedy i tylko wtedy, gdy przyjmuje wartości w nieskończenie wymiarowej przestrzeni (przestrzeni funkcyjnej). Obserwację  $\chi$  zmiennej  $X$  nazywamy **daną funkcjonalną** (ang. *functional data*).

Jeśli zmienna funkcjonalna  $X$  (odpowiednio obserwacja  $\chi$ ) jest krzywą, to zachodzi  $X = \{X(t), t \in T\}$  (odp.  $\chi = \{\chi(t), t \in T\}$ ), gdzie zbiór indeksów  $T \subset \mathbb{R}$ . Taką zmienną funkcjonalną możemy zatem utożsamiać z procesem stochastycznym z nieskończenie wymiarową przestrzenią stanów. W szczególności, zmienna funkcjonalna może być powierzchnią, czyli dwuwymiarowym wektorem krzywych - wtedy, analogicznie,  $T$  będzie dwuwymiarowym zbiorem indeksów tj.  $T \subset \mathbb{R}^2$  - lub dowolnie wymiarowym wektorem krzywych.

W niniejszej pracy skupimy się na zmiennych funkcjonalnych przyjmujących postać krzywych.

Aby zbudować pojęcie operatora kowariancji dla zmiennych funkcjonalnych wprowadzimy niezbędne pojęcia z dziedziny operatorów liniowych.

### 1.1. Klasyfikacja operatorów liniowych

Niech  $(\Omega, \mathcal{F}, P)$  będzie przestrzenią probabilistyczną,  $\Omega$  jest zatem zbiorem scenariuszy  $\omega$ ,  $\mathcal{F}$  jest  $\sigma$ -algebrą podzbiorów  $\Omega$ , a  $P$  miarą prawdopodobieństwa nad  $\mathcal{F}$ . Dla uproszczenia zakładamy zupełność zadanej przestrzeni probabilistycznej. Rozważmy proces stochastyczny z czasem ciągłym  $X = \{X_t, t \in T\}$ , gdzie  $T$  jest przedziałem w  $\mathbb{R}$ , zdefiniowany na przestrzeni probabilistycznej  $(\Omega, \mathcal{F}, P)$ , taki, że  $X_t(\omega)$  należy do przestrzeni funkcyjnej  $E$  dla wszystkich  $\omega \in \Omega$ .

W pracy rozważać będziemy zmienne funkcjonalne przyjmujące wartości w przestrzeni Hilberta.

Rozważmy ośrodkową nieskończenie wymiarową przestrzeń Hilberta  $H$  z iloczynem skalarnym  $\langle \cdot, \cdot \rangle$  zadającym normę  $\|\cdot\|$  i oznaczmy przez  $\mathcal{L}$  przestrzeń ciągłych (ograniczonych) operatorów liniowych w  $H$  z normą

$$\|\Psi\|_{\mathcal{L}} := \sup\{\|\Psi(x)\| : \|x\| \leq 1\}.$$

**Definicja 1.1.1** [Horváth, Kokoszka]

Operator  $\Psi \in \mathcal{L}$  nazywamy **operatorem zwartym**, jeśli istnieją dwie ortonormalne bazy  $\{\nu_j\}_{j=1}^\infty$  i  $\{f_j\}_{j=1}^\infty$ , oraz ciąg liczb rzeczywistych  $\{\lambda_j\}_{j=1}^\infty$  zbieżny do zera, takie że

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, \nu_j \rangle f_j, \quad x \in H. \quad (1.1)$$

Bez straty ogólności możemy założyć, że w przedstawionej reprezentacji  $\lambda_j$  są wartościami dodatnimi, w razie konieczności wystarczy  $f_j$  zamienić na  $-f_j$ .

Równoważną definicją operatora zwartego jest spełnienie następującego warunku: zbieżność  $\langle y, x_n \rangle \rightarrow \langle y, x \rangle$  dla każdego  $y \in H$  implikuje  $\|\Psi(x_n) - \Psi(x)\| \rightarrow 0$ .

Inną klasę operatorów są operatory Hilberta-Schmidta, którą oznaczać będziemy przez  $\mathcal{S}$ .

**Definicja 1.1.2** [Bosq]

**Operatorem Hilberta-Schmidta** nazywamy taki operator zwarty  $\Psi \in \mathcal{L}$ , dla którego ciąg  $\{\lambda_j\}_{j=1}^\infty$  w reprezentacji (1.1) spełnia  $\sum_{j=1}^\infty \lambda_j^2 < \infty$ .

**Uwaga 1.1.1** [Bosq], [Horváth, Kokoszka]

Klasa  $\mathcal{S}$  jest przestrzenią Hilberta z iloczynem skalarnym

$$\langle \Psi_1, \Psi_2 \rangle_{\mathcal{S}} := \sum_{j=1}^{\infty} \langle \Psi_1(e_j), \Psi_2(e_j) \rangle,$$

gdzie  $\{e_j\}_{j=1}^\infty$  jest dowolną bazą ortonormalną w  $H$ .

Powyższy iloczyn skalarny zadaje normę  $\|\Psi\|_{\mathcal{S}} := \left( \sum_{j=1}^\infty \lambda_j^2 \right)^{1/2}$ .

**Definicja 1.1.3** [Bosq]

Operator liniowy nazywamy **operatorem śladowym** (ang. nuclear operator), jeśli równość (1.1) spełniona jest dla ciągu takiego, że  $\sum_{j=1}^\infty |\lambda_j| < \infty$ .

**Uwaga 1.1.2** [Bosq]

Klasa operatorów śladowych  $\mathcal{N}$  z normą  $\|\Psi\|_{\mathcal{N}} := \sum_{j=1}^\infty |\lambda_j|$  jest przestrzenią Banacha.

**Definicja 1.1.4** [Horváth, Kokoszka]

Operator  $\Psi \in \mathcal{L}$  nazywamy **symetrycznym**, jeśli

$$\langle \Psi(x), y \rangle = \langle x, \Psi(y) \rangle, \quad x, y \in H,$$

oraz **nieujemnie określonym** (połowicznie pozytywnie określonym, ang. positive semidefinite), jeśli

$$\langle \Psi(x), x \rangle \geq 0, \quad x \in H.$$

**Uwaga 1.1.3** [Horváth, Kokoszka]

Symetryczny nieujemnie określony operator Hilberta-Schmidta  $\Psi$  możemy przedstawić w reprezentacji

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, \nu_j \rangle \nu_j, \quad x \in H, \quad (1.2)$$

gdzie ortonormalne  $\nu_j$  są **funkcjami własnymi**  $\Psi$ , tj.  $\Psi(\nu_j) = \lambda_j \nu_j$ . Funkcje  $\nu_j$  mogą być rozszerzone do bazy, przez dopełnienie ortogonalne podprzestrzeni rozpiętej przez oryginalne  $\nu_j$ . Możemy zatem założyć, że funkcje  $\nu_j$  w (1.2) tworzą bazę, a pewne wartości  $\lambda_j$  mogą być równe zero.

## 1.2. Przestrzeń $L^2$

Przestrzeń  $L^2 = L^2(K, \mathcal{A}, \mu)$  nad pewną przestrzenią liniową  $K$  jest zbiorem mierzalnych funkcji rzeczywistych określonych na  $K$  spełniających  $\int_K x^2(t)dt < \infty$ . Przestrzeń  $L^2$  jest ośrodkową przestrzenią Hilberta z iloczynem skalarnym

$$\langle x, y \rangle := \int_K x(t)y(t)dt.$$

Tak jak zwyczajowo zapisujemy  $L^2$  zamiast  $L^2(K)$ , tak w przypadku symbolu całki bez wskazania obszaru całkowania będziemy mieć na myśli całkowanie po całej przestrzeni  $K$ . Jeśli  $x, y \in L^2$ , równość  $x = y$  zawsze oznaczać będzie  $\int [x(t) - y(t)]^2 dt = 0$ .

Ważną klasę operatorów liniowych na przestrzeni  $L^2$  stanowią operatory całkowe.

**Definicja 1.2.1** *Operatorem całkowym* nazywamy operator liniowy  $\Psi$  dający się przedstawić w formie

$$\Psi(x)(t) = \int \psi(t, s)x(s)ds, \quad x \in L^2,$$

gdzie  $\psi$  stanowi **jądro całkowe** operatora  $\Psi$ .

**Uwaga 1.2.1** [Horváth, Kokoszka]

Operatory całkowe są operatorami Hilberta-Schmidta wtedy i tylko wtedy, gdy

$$\iint \psi^2(t, s)dtds < \infty.$$

Ponadto zachodzi

$$\|\Psi\|_{\mathcal{S}}^2 = \iint \psi^2(t, s)dtds.$$

**Uwaga 1.2.2** (Twierdzenie Mercera) [Horváth, Kokoszka]

Jeśli operator spełnia również  $\psi(s, t) = \psi(t, s)$  oraz  $\iint \psi(t, s)x(t)x(s)dtds \geq 0$ , to operator całkowy  $\Psi$  jest symetryczny i nieujemnie określony, zatem z uwagi 1.1.3 mamy

$$\psi(t, s) = \sum_{j=1}^{\infty} \lambda_j \nu_j(t) \nu_j(s) \quad \text{w } L^2(K) \times L^2(K).$$

Jeżeli funkcja  $\psi$  jest ciągła, powyższe rozwinięcie jest prawdziwe dla wszystkich  $s, t \in K$  i szereg jest zbieżny jednostajnie.

## 1.3. Zmienne funkcyjne w $L^2$ . Pojęcie operatora kowariancji

Rozważmy zmienną funkcyjną  $X = \{X(t), t \in T\}$  będącą krzywą ( $T \subset \mathbb{R}$ ) jako element losowy z przestrzeni  $L^2(T)$  zaopatrzonej w  $\sigma$ -algebrę borelowskich podzbiorów  $T$ .

Mówimy, że zmienna  $X$  jest **całkowalna**, jeśli  $\mathbb{E} \|X\| = \mathbb{E} [\int X^2(t)dt]^{1/2} < \infty$ .

**Definicja 1.3.1** [Bosq]

**Operator kowariancji** scentrowanej zmiennej funkcyjnej  $X$  (tj.  $\mathbb{E}X = 0$ ) przyjmującej wartości w przestrzeni funkcyjnej  $L^2$  spełniającej  $\mathbb{E} \|X\|^2 < \infty$  definiujemy następująco

$$C_X(x) := \mathbb{E}[\langle X, x \rangle X], \quad x \in L^2.$$

Jeśli  $Y$  jest zmienną funkcjonalną spełniającą powyższe warunki, wtedy operator kowariancji między zmiennymi  $X$  i  $Y$  przedstawiamy jako

$$C_{X,Y}(x) := \mathbb{E}[\langle X, x \rangle Y], \quad x \in L^2$$

oraz

$$C_{Y,X}(x) := \mathbb{E}[\langle Y, x \rangle X], \quad x \in L^2.$$

Operator kowariancji jest operatorem całkowym, czyli

$$C_X(x)(t) = \int c(t, s)x(s)ds, \quad \text{gdzie } c(t, s) = \mathbb{E}[X(t)X(s)].$$

Oczywistym jest, że  $c(t, s) = c(s, t)$  i mamy

$$\iint c(t, s)x(t)x(s)dtds = \iint \mathbb{E}[X(t)X(s)]x(t)x(s)dtds = \mathbb{E}\left[\left(\int X(t)x(t)dt\right)^2\right] \geq 0.$$

Zatem operator kowariancji  $C_X$  jest symetryczny oraz nieujemnie określony. Wartości własne  $\lambda_j$  operatora  $C_X$  są dodatnie i spełniony jest warunek  $\sum_{j=1}^{\infty} \lambda_j = \mathbb{E}\|X\|^2 < \infty$ .  $C_X$  jest operatorem Hilberta-Schmidta (a nawet operatorem śladowym) i posiada on następującą reprezentację

$$C_X(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, \nu_j \rangle \nu_j, \quad x \in L^2.$$

[już tu: estymatory operatorów kowariancji?]

## 1.4. Funkcjonalny model liniowy

Standardowy model liniowy dla par zmiennych skalarnych  $Y_n$  i wektorów  $\mathbf{X}_n$  ( $n = 1, \dots, N$ ), przy założeniu  $\mathbb{E}Y_n = 0$ ,  $\mathbb{E}\mathbf{X}_n = \mathbf{0}^1$ , ma postać

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.3}$$

gdzie

- $\mathbf{Y}$  jest wektorem zmiennych objaśnianych długości  $N$ ,
- $\mathbf{X}$  jest macierzą zmiennych objaśniających wymiaru  $N \times p$ ,
- $\boldsymbol{\beta}$  jest wektorem parametrów długości  $p$ ,
- $\boldsymbol{\varepsilon}$  jest wektorem błędów losowych długości  $N$ .

[ Mając dane realizacje zmiennych  $\mathbf{Y}$  oraz  $\mathbf{X}$  poszukiwany wektor współczynników modelu  $\boldsymbol{\beta}$  znajdujemy metodą najmniejszych kwadratów. ]

Poza narzuconym już założeniem o scentrowanych zmiennych losowych  $\mathbf{Y}$  i  $\mathbf{X}$  (tu: jedynie aby uniknąć uwzględniania wyrazu wolnego<sup>2</sup>) najważniejszymi założeniami powyższego modelu liniowego są wymagania, aby zmienna losowa  $\boldsymbol{\varepsilon}$  opisująca błąd modelu również spełniała  $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$  oraz aby nie była skorelowana ze zmiennymi  $X_n$ ).

Rozważać będziemy odpowiednik modelu liniowego dla zmiennych funkcjonalnych. Dla uproszczenia (podobnie jak wyżej) zakładamy, że zmienne objaśniane i objaśniające mają

<sup>1</sup>przenieść tę uwagę/wytłumaczenie do przypisu?

<sup>2</sup>przenieść tę uwagę/wytłumaczenie do przypisu?

średnie równe zero. **Pełen model funkcjonalny** (ang. *fully functional model*) przyjmuje postać

$$Y_n = \Psi X_n + \varepsilon_n, \quad n = 1, \dots, N, \quad (1.4)$$

gdzie krzywe  $Y_n$ ,  $X_n$  oraz nieobserwowalny błąd  $\varepsilon_n$  należą do przestrzeni Hilberta  $L^2(T)$ . Operator  $\Psi : L^2 \rightarrow L^2$  jest ograniczonym operatorem liniowym, który w szczególności jest również operatorem całkowym, którego jądro całkowe  $\psi(t, s)$  jest funkcją całkowalną z kwadratem na  $T \times T$ . Równość (1.4) rozumiemy zatem następująco

$$Y_n(t) = \int \psi(s, t) X_n(s) ds + \varepsilon_n(t), \quad n = 1, \dots, N.$$

[ Nazwa powyższego modelu wynika z faktu, że zarówno zmienne objaśniane  $Y_n$  jak i zmienne objaśniające  $X_n$  są zmiennymi funkcjonalnymi. Niewielkim uproszczeniem są pozostałe typy funkcjonalnych modeli liniowych, tj.

- model z odpowiedzią skalarną (ang. *scalar response model*)

$$Y_n = \int \psi(s) X_n(s) ds + \varepsilon_n, \quad n = 1, \dots, N,$$

w którym tylko zmienne objaśniające  $X$  są zmiennymi funkcjonalnymi,

- model z odpowiedzią funkcyjną (ang. *functional response model*)

$$Y_n(t) = \psi(t) X_n + \varepsilon_n(t), \quad n = 1, \dots, N,$$

w którym zmienne objaśniające  $X_n$  są skalarami. ]

Naturalnym problemem pojawiającym się przy funkcjonalnym modelu liniowym jest estymacja operatora  $\Psi$ ...



## Rozdział 2

# Test istotności w funkcjonalnym modelu liniowym

### 2.1. Procedura testowa

Jednym z podstawowych testów na efektywność modelu jest test istotności zmiennych objaśniających. Jak w przypadku modelu liniowego dla zmiennych skalarnych (postaci (1.3)) testuje się hipotezę o zerowaniu się wektora  $\beta$ , tak w przypadku funkcjonalnego modelu liniowego badamy zerowanie się operatora  $\Psi$ , tj. hipotezy

$$H_0 : \quad \Psi = 0 \quad \text{przeciw} \quad H_A : \quad \Psi \neq 0.$$

Zauważmy, że przyjęcie  $H_0$  nie oznacza braku związku między zmienną objaśnianą a objaśniającą. Prowadzi jedynie do stwierdzenia braku zależności liniowej.

Zakładamy, że zmienna objaśniana  $Y_n$ , zmienne objaśniające  $X_n$  i błędy  $\varepsilon_n$  są scentrowanymi zmiennymi losowymi przyjmującymi wartości w przestrzeni Hilberta  $L^2$ . Oznaczając przez  $X$  (analogicznie  $Y$ ) losową funkcję o tym samym rozkładzie co  $X_n$  ( $Y_n$ ) wprowadzamy operatory

$$C(x) = \mathbb{E}[\langle X, x \rangle X], \quad \Gamma(x) = \mathbb{E}[\langle Y, x \rangle Y], \quad \Delta(x) = \mathbb{E}[\langle X, x \rangle Y].$$

Przez  $\hat{C}$ ,  $\hat{\Gamma}$ ,  $\hat{\Delta}$  oznaczamy ich estymatory, np.

$$\hat{C}(x) = \frac{1}{N} \sum_{n=1}^N \langle X_n, x \rangle X_n.$$

Definiujemy również wartości i wektory własne  $C$  i  $\Gamma$

$$C(v_k) = \lambda_k v_k, \quad \Gamma(u_j) = \gamma_j u_j,$$

których estymatory będziemy oznaczać  $(\hat{\lambda}_k, \hat{v}_k)$ ,  $(\hat{\gamma}_j, \hat{u}_j)$ .

Test obejmuje obcięcie powyższych operatorów na podprzestrzeń skończenie wymiarowe. Podprzestrzeń  $\mathcal{V}_p = \text{span}\{v_1, \dots, v_p\}$  zawiera najlepsze przybliżenia  $X_n$ , które są liniowymi kombinacjami pierwszych  $p$  głównych składowych (ang. *Functional Principal Components, FPC*). Metodą głównych składowych wyznaczamy  $p$  największych wartości własnych operatora  $\hat{C}$  tak, że  $\hat{\mathcal{V}}_p = \text{span}\{\hat{v}_1, \dots, \hat{v}_p\}$  zawiera najlepsze przybliżenie  $X_n$ . Analogicznie  $\mathcal{U}_q = \text{span}\{u_1, \dots, u_q\}$  zawiera przybliżenia  $\text{span}\{Y_1, \dots, Y_N\}$ .

Z równości

$$Y(t) = \int \psi(s, t)X(s)ds + \epsilon(t)$$

wynika  $\Delta = \psi C$  i dla  $k \leq p$  mamy

$$\psi(v_k) = \lambda_k^{-1} \Delta(v_k).$$

Stąd,  $\psi$  zeruje się na  $\text{span}\{v_1, \dots, v_p\}$  wtedy i tylko wtedy, gdy  $\Delta(v_k) = 0$  dla każdego  $k = 1, \dots, p$ . Zauważmy, że

$$\Delta(v_k) \approx \hat{\Delta}(v_k) = \frac{1}{N} \sum_{n=1}^N \langle X_n, v_k \rangle Y_n.$$

Skoro zatem  $\text{span}\{Y_1, \dots, Y_N\}$  są dobrze aproksymowane przez  $\mathcal{U}_q$ , to możemy ograniczyć się do sprawdzania czy

$$\langle \hat{\Delta}(v_k), u_j \rangle = 0, \quad k = 1, \dots, p, \quad j = 1, \dots, q. \quad (2.1)$$

Jeśli  $H_0$  jest prawdziwa, to dla każdego  $x \in \mathcal{V}_p$ ,  $\psi(x)$  nie należy do  $\mathcal{U}_q$ . Co znaczy, że żadna funkcja  $Y_n$  nie może być opisana jako liniowa kombinacja  $X_n$ ,  $n = 1, \dots, N$ . Statystyka testowa powinna zatem sumować kwadraty iloczynów skalarnych (2.1). Poniższe twierdzenia prowadzą do wyznaczenia statystyki

$$\hat{T}_N(p, q) = N \sum_{k=1}^p \sum_{j=1}^q \hat{\lambda}_k^{-1} \hat{\gamma}_j^{-1} \langle \hat{\Delta}(\hat{v}_k), \hat{u}_j \rangle^2, \quad (2.2)$$

która zbiega według rozkładu do rozkładu  $\chi^2$  z  $pq$  stopniami swobody. Przy czym

$$\langle \hat{\Delta}(\hat{v}_k), \hat{u}_j \rangle = \left\langle \frac{1}{N} \sum_{n=1}^N \langle X_n, \hat{v}_k \rangle Y_n, \hat{u}_j \right\rangle = \frac{1}{N} \sum_{n=1}^N \langle X_n, \hat{v}_k \rangle \langle Y_n, \hat{u}_j \rangle$$

oraz  $\lambda_k = \mathbb{E} \langle X, v_k \rangle^2$  i  $\gamma_j = \mathbb{E} \langle Y, u_j \rangle^2$ .

**Uwaga 2.1.1** Oczywiście jest, że jeśli odrzucamy  $H_0$ , to  $\psi(v_k) \neq 0$  dla pewnego  $k \geq 1$ . Jednak ograniczając się do  $p$  największych wartości własnych, test jest skuteczny tylko jeśli  $\psi$  nie zanika na którymś wektorze  $v_k$ ,  $k = 1, \dots, p$ . Aczkolwiek takie ograniczenie jest intuicyjnie niegroźne, ponieważ test ma za zadanie sprawdzić czy główne źródła zmienności  $Y$  mogą być opisane przez główne źródła zmienności zmiennych  $X$ .

### Schemat przebiegu testu

1. Sprawdzamy założenie o liniowości metodą *FPC score predictor-response plots*.
2. Wybieramy liczbę głównych składowych  $p$  i  $q$  metodami *scree test* oraz *CPV*.
3. Wyliczamy wartość statystyki  $\hat{T}_N(p, q)$  (2.2).
4. Jeśli  $\hat{T}_N(p, q) > \chi_{pq}^2(1 - \alpha)$ , to odrzucamy hipotezę zerową o braku liniowej zależności. W przeciwnym razie nie mamy podstaw do odrzucenia  $H_0$ .

...



## 2.2. Formalne podstawy

**Założenie 2.2.1** [Kokoszka et al. (2008)], [Horváth, Kokoszka]

Trójka  $(Y_n, X_n, \varepsilon_n)$  tworzy ciąg niezależnych elementów losowych o jednakowym rozkładzie, takich że  $\varepsilon_n$  jest niezależne od  $X_n$  oraz

$$\mathbb{E}X_n = 0, \quad \mathbb{E}\varepsilon_n = 0,$$

$$\mathbb{E}\|X_n\|^4 < \infty \quad i \quad \mathbb{E}\|\varepsilon_n\|^4 < \infty.$$

**Założenie 2.2.2** [Kokoszka et al. (2008)], [Horváth, Kokoszka]

Wartości własne operatorów  $C$  oraz  $\Gamma$  spełniają, dla pewnych  $p > 0$  i  $q > 0$

$$\lambda_1 > \lambda_2 > \dots > \lambda_p > \lambda_{p+1}, \quad \gamma_1 > \gamma_2 > \dots > \gamma_q > \gamma_{q+1}.$$

**Twierdzenie 2.2.1** [Kokoszka et al. (2008)], [Horváth, Kokoszka]

Jeśli spełnione są  $H_0$  i powyższe Założenia 2.2.1, 2.2.2, to  $\hat{T}_N(p, q) \xrightarrow{d} \chi_{pq}^2$  przy  $N \rightarrow \infty$ .

**Twierdzenie 2.2.2** [Kokoszka et al. (2008)], [Horváth, Kokoszka]

Przy Założeniach 2.2.1, 2.2.2 oraz jeśli  $\langle \psi(v_k), u_j \rangle \neq 0$  dla pewnych  $k \leq p$  oraz  $j \leq q$ , to  $\hat{T}_N(p, q) \xrightarrow{P} \chi_{pq}^2$  przy  $N \rightarrow \infty$ .

Dowody...

□

**Lemat 2.2.1** [Kokoszka et al. (2008)], [Bosq]

Przy powyższych Założeniach spełnione są nierówności

$$\limsup_{N \rightarrow \infty} N \mathbb{E} \|\nu_k - \hat{\nu}_k\|^2 < \infty, \quad \limsup_{N \rightarrow \infty} N \mathbb{E} \|u_j - \hat{u}_j\|^2 < \infty,$$

$$\limsup_{N \rightarrow \infty} N \mathbb{E} \left[ |\gamma_k - \hat{\gamma}_k|^2 \right] < \infty, \quad \limsup_{N \rightarrow \infty} N \mathbb{E} \left[ |\lambda_j - \hat{\lambda}_j|^2 \right] < \infty,$$

dla  $k \leq p$  oraz  $j \leq q$ .

**Lemat 2.2.2** [Kokoszka et al. (2008)], [Horváth, Kokoszka]

Jeśli spełnione są  $H_0$  i powyższe Założenia, to dla  $j \leq q$ ,  $k \leq p$

$$\sqrt{N} \langle \hat{\Delta} \nu_k, u_j \rangle \xrightarrow{d} \eta_{kj} \sqrt{\gamma_k \lambda_j},$$

gdzie  $\eta_{kj} \sim N(0, 1)$ . Przy czym  $\eta_{k,j}$  oraz  $\eta_{k',j'}$  są niezależne dla  $(k, j) \neq (k', j')$ .

Dowód. Przy  $H_0$

$$\sqrt{N} \langle \hat{\Delta} \nu_k, u_j \rangle = N^{-1/2} \sum_{n=1}^N \langle X_n, \nu_k \rangle \langle \varepsilon_n, u_j \rangle.$$

... Aby udowodnić niezależność między  $\eta_{kj}$  i  $\eta_{k'j'}$  dla  $(k, j) \neq (k', j')$ , wystarczy pokazać, że  $\sqrt{N}(\hat{\Delta}(\nu_k), u_j)$  i  $\sqrt{N}(\hat{\Delta}(\nu_{k'}), u_{j'})$  są nieskorelowane

$$\begin{aligned} & \mathbb{E} \left[ \sqrt{N} \langle \hat{\Delta}(\nu_k), u_j \rangle, \sqrt{N} \langle \hat{\Delta}(\nu_{k'}), u_{j'} \rangle \right] \\ &= \frac{1}{N} \mathbb{E} \left[ \sum_{n=1}^N \langle X_n, \nu_k \rangle \langle \varepsilon_n, u_j \rangle \sum_{n'=1}^N \langle X_{n'}, \nu_{k'} \rangle \langle \varepsilon_{n'}, u_{j'} \rangle \right] \\ &= \frac{1}{N} \sum_{n, n'=1}^N \mathbb{E} [\langle X_n, \nu_k \rangle \langle X_{n'}, \nu_{k'} \rangle] \mathbb{E} [\langle \varepsilon_n, u_j \rangle \langle \varepsilon_{n'}, u_{j'} \rangle] \\ &= \frac{1}{N} \sum_{n=1}^N \mathbb{E} [\langle X_n, \nu_k \rangle \langle X_n, \nu_{k'} \rangle] \mathbb{E} [\langle \varepsilon_n, u_j \rangle \langle \varepsilon_n, u_{j'} \rangle] \\ &= \langle C(\nu_k), \nu_{k'} \rangle \langle \Gamma u_j, u_{j'} \rangle = \gamma_k \delta_{kk'} \gamma_j \delta_{jj'}. \end{aligned}$$

□

Przypomnijmy, że norma Hilberta-Schmidta operatora Hilberta-Schmidta  $S$  zdefiniowana jest wzorem  $\|S\|_S^2 = \sum_{j=1}^{\infty} \|S(e_j)\|^2$ , gdzie ciąg  $\{e_1, e_2, \dots\}$  stanowi bazę ortonormalną oraz, że norma ta jest nie mniejsza od normy operatorowej, tj.  $\|S\|_{\mathcal{L}}^2 \leq \|S\|_S^2$ .

**Lemat 2.2.3** [Kokoszka et al. (2008)], [Horváth, Kokoszka]

Przy założeniach Twierdzenia 2.2.1 mamy

$$\mathbb{E} \left\| \hat{\Delta} \right\|_S^2 = N^{-1} \mathbb{E} \|X\|^2 \mathbb{E} \|\varepsilon_1\|^2.$$

*Dowód.* Zauważmy, że

$$\left\| \hat{\Delta}(e_j) \right\|^2 = N^{-2} \sum_{n, n'=1}^N \langle X_n, e_j \rangle \langle X_{n'}, e_j \rangle \langle Y_n, Y_{n'} \rangle.$$

Stąd mamy

$$\begin{aligned} \mathbb{E} \left\| \hat{\Delta} \right\|_S^2 &= N^{-2} \sum_{j=1}^{\infty} \sum_{n, n'=1}^N \mathbb{E} [\langle X_n, e_j \rangle \langle X_{n'}, e_j \rangle \langle \varepsilon_n, \varepsilon_{n'} \rangle] \\ &= N^{-2} \sum_{j=1}^{\infty} \sum_{n, n'=1}^N \mathbb{E} \langle X_n, e_j \rangle^2 \mathbb{E} \|\varepsilon_n\|^2 \\ &= N^{-1} \mathbb{E} \|\varepsilon_1\|^2 \sum_{j=1}^{\infty} \langle X, e_j \rangle^2 = N^{-1} \mathbb{E} \|\varepsilon_1\|^2 \|X\|^2. \end{aligned}$$

□

**Lemat 2.2.4** [Kokoszka et al. (2008)], [Horváth, Kokoszka]

Załóżmy, że  $\{U_n\}_{n=1}^{\infty}$  oraz  $\{V_n\}_{n=1}^{\infty}$  są ciągami elementów losowych z przestrzeni Hilberta takich, że  $\|U_n\| \xrightarrow{P} 0$  i  $\|V_n\| = O_P(1)$ , tj.

$$\lim_{C \rightarrow \infty} \limsup_{n \rightarrow \infty} P(\|V_n\| > C) = 0.$$

Wtedy zachodzi

$$\langle U_n, V_n \rangle \xrightarrow{P} 0.$$

*Dowód.* Prawdziwość lematu wynika z analogicznej własności dla losowych ciągów liczb rzeczywistych i nierówności  $|\langle U_n, V_n \rangle| \leq \|U_n\| \|V_n\|$ .  $\square$

**Lemat 2.2.5** [Kokoszka et al. (2008)], [Horváth, Kokoszka]  
Przy założeniach Twierdzenia 2.2.1, dla  $j \leq q$ ,  $k \leq p$  zachodzi

$$\sqrt{N} \langle \hat{\Delta}(\hat{\nu}_k), \hat{u}_j \rangle \xrightarrow{d} \eta_{kj} \sqrt{\gamma_k \lambda_j},$$

gdzie  $\eta_{kj}$  definiowane są jak w Lemacie 2.2.2.

*Dowód.* Na mocy Lematu 2.2.2, wystarczy pokazać

$$\sqrt{N} \langle \hat{\Delta}(\hat{\nu}_k), \hat{u}_j \rangle - \sqrt{N} \langle \hat{\Delta}(\nu_k), u_j \rangle \xrightarrow{P} 0. \quad (2.3)$$

Równość (2.3) wynika z

$$\sqrt{N} \langle \hat{\Delta}(\hat{\nu}_k), \hat{u}_j - u_j \rangle \xrightarrow{P} 0 \quad (2.4)$$

i

$$\sqrt{N} \langle \hat{\Delta}(\hat{\nu}_k - \nu_k), \hat{u}_j \rangle \xrightarrow{P} 0. \quad (2.5)$$

Aby udowodnić równość (2.4), zauważmy, że  $\sqrt{N}(\hat{u}_j - u_j) = O_P(1)$  oraz, na mocy Lematu 2.2.3,  $\mathbb{E} \|\hat{\Delta}(\nu_k)\| \leq \mathbb{E} \|\hat{\Delta}\|_{\mathcal{S}} = O(N^{-1/2})$ . Stąd równość (2.4) wynika z Lematu 2.2.4...  $\square$

**Wniosek 2.2.1** [Kokoszka et al. (2008)], [Horváth, Kokoszka]  
Przy założeniach Twierdzenia 2.2.1, dla  $j \leq q$ ,  $k \leq p$  zachodzi

$$\sqrt{N} \langle \hat{\lambda}_k^{-1/2} \hat{\gamma}_j^{-1/2} \hat{\Delta}(\hat{\nu}_k), \hat{u}_j \rangle \xrightarrow{d} \eta_{kj},$$

gdzie  $\eta_{kj}$  definiowane są jak w Lemacie 2.2.2.

**Lemat 2.2.6** [Kokoszka et al. (2008)], [Horváth, Kokoszka] Jeśli  $\{Y_n\}_{n \geq 1}$  są elementami losowymi o jednakowych rozkładach, to zachodzi  $\mathbb{E} \|\hat{\Delta}\| \leq \mathbb{E} \|Y\|^2$ .

*Dowód.* Dla dowolnego  $u \in L^2$  takiego, że  $\|u\| \leq 1$ , mamy

$$\|\hat{\Delta}u\| \leq N^{-1} \sum_{n=1}^N |\langle Y_n, u \rangle| \|Y_n\| \leq N^{-1} \sum_{n=1}^N \|Y_n\|^2.$$

Co ze względu na założenie, że  $Y_n$  mają jednakowy rozkład, jest równoważne tezie lematu.  $\square$

**Lemat 2.2.7** [Kokoszka et al. (2008)], [Horváth, Kokoszka]  
Jeżeli spełnione jest Założenie 2.2.1, wtedy dla dowolnych funkcji  $\nu, u \in L^2$

$$\langle \hat{\Delta}(\nu), u \rangle \xrightarrow{P} \langle \Delta(\nu), u \rangle.$$

*Dowód.* Tezę otrzymujemy korzystając z Prawa Wielkich Liczb zauważając

$$\langle \hat{\Delta}(\nu), u \rangle = \frac{1}{N} \sum_{n=1}^N \langle X_n, \nu \rangle \langle Y_n, u \rangle$$

oraz

$$\mathbb{E}[\langle X_n, \nu \rangle \langle Y_n, u \rangle] = \mathbb{E}[\langle \langle X_n, \nu \rangle Y_n, u \rangle] = \langle \Delta(\nu), u \rangle.$$

$\square$

**Lemat 2.2.8** [Kokoszka et al. (2008)], [Horváth, Kokoszka]  
*Jeżeli spełnione są Założenia 2.2.1 i 2.2.2, to*

$$\langle \widehat{\Delta}(\hat{\nu}_k), \hat{u}_j \rangle \xrightarrow{P} \langle \Delta(\nu_k), u_j \rangle, \quad \text{dla } k \leq p, \ j \leq q.$$

*Dowód.* Na mocy Lematu 2.2.7 wystarczy pokazać

$$\langle \widehat{\Delta}(\nu_k), \hat{u}_j - u_j \rangle \xrightarrow{P} 0$$

i

$$\langle \widehat{\Delta}(\hat{\nu}_k) - \widehat{\Delta}(\nu_k), \hat{u}_j \rangle \xrightarrow{P} 0.$$

Relacje te wynikają z Lematu 2.2.4 oraz Lematu 2.2.6. □

*Dowód Twierdzenia 2.2.2.* Wprowadźmy oznaczenie

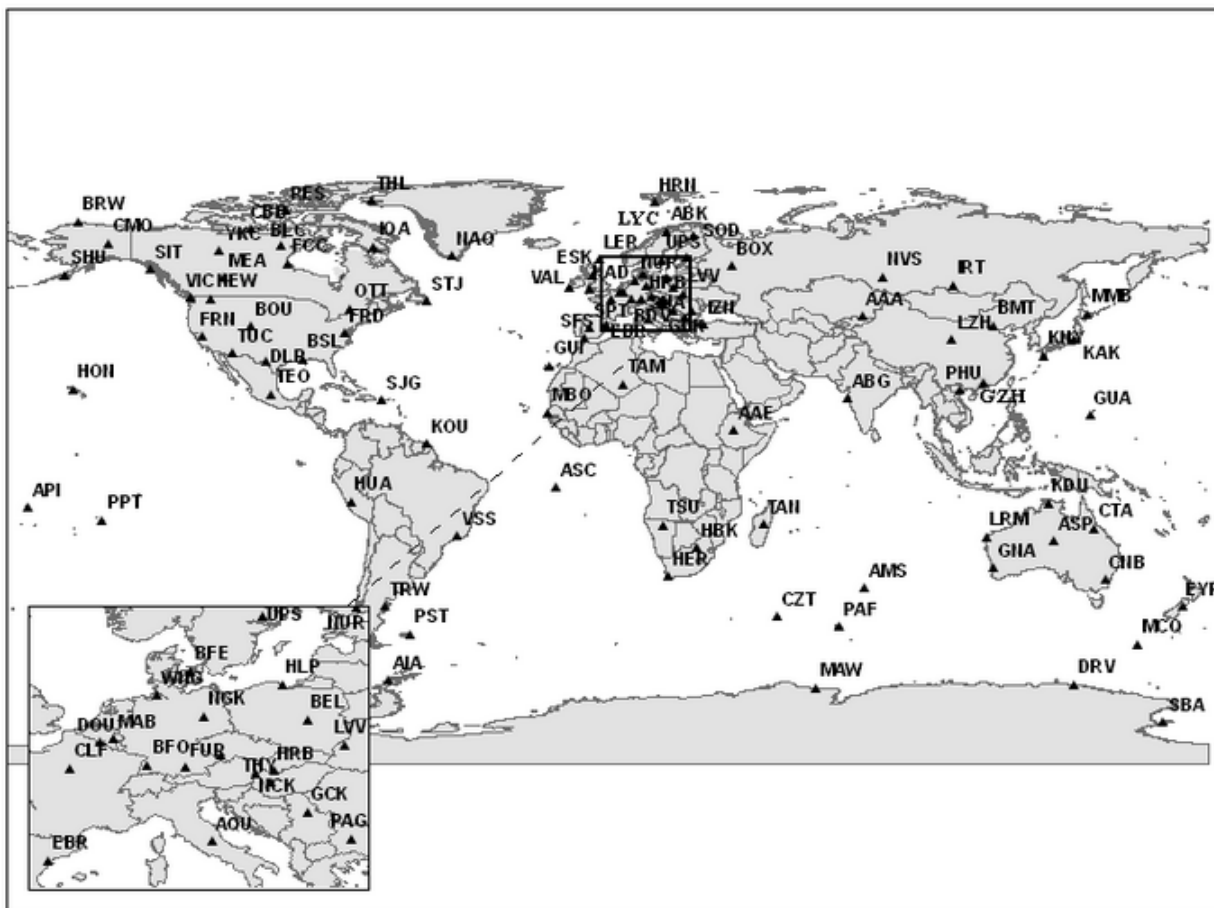
$$\widehat{S}_N(p, q) = \sum_{k=1}^p \sum_{j=1}^q \hat{\lambda}_k^{-1} \hat{\gamma}_j^{-1} \langle \widehat{\Delta}(\hat{\nu}_k), \hat{u}_j \rangle^2.$$

Na mocy Lematu 2.2.8 ... □

## Rozdział 3

# Przykład zastosowania

Magnetometer data... dostępne na stronie INTERMAGNET [1]



Rysunek 3.1: Mapa stacji meteorologicznych należących do programu INTERMAGNET

Korzystając z dostępnego pakietu *fda* ([R: fda])



**Dodatek A**

**Kod w R**

...





# Bibliografia

- [Bosq] D. Bosq, *Linear Processes in Function Spaces*. Springer, 2000.
- [Ferraty, Vieu] F. Ferraty, P. Vieu, *Nonparametric Functional Data Analysis. Theory and practice*. Springer, 2006.
- [Horváth, Kokoszka] L. Horváth, P. Kokoszka, *Interference for Functional Data with Applications*. Springer, 2012.
- [I] INTERMAGNET <http://www.intermagnet.org/index-eng.php>
- [Kokoszka et al. (2008)] P. Kokoszka, I. Maslova, J. Sojka, L. Zhu, *Testing for lack of dependence in the functional linear model*. Canadian Journal of Statistics, 2008, 36, 207-222.
- [R: fda] J. O. Ramsay, H. Wickham, S. Graves, G. Hooker, *Package 'fda'*. On-line: <https://cran.r-project.org/web/packages/fda/fda.pdf>
- [Ramsay, Silverman] J. O. Ramsay, B. W. Silverman, *Functional Data Analysis*. Springer, 2005.