Streszczenie

Słowa kluczowe: dane funkcjonalne, analiza danych funkcjonalnych, funkcjonalny model liniowy, test istotności

Dziedzina nauki i techniki, zgodnie z wymogami OECD: 1.1 Matematyka.

Abstract

The paper's motivation is to contribute to popularization of mathematical statistics on infinite dimensional function Hilbert spaces. The author presents the fully functional linear model in form $Y = \Psi X + \varepsilon$ and its significance test proposed by Kokoszka et al. The test detects nullity of Hilbert-Schmidt operator Ψ, which implies the lack of linear dependence between X and Y. Using the principal component decomposition it is concluded with test statistic convergent by distribution to chi-squared.
The test is further used for magnetic field data collected in some stations in different latitudes. The results show linear dependence between horizontal intensities of the magnetic field in mid- and low-latitude stations with high-latitude station data with a day or two delay but they contradict the linear dependence between data with more than a two-day lag.

Keywords: functional data, functional data analysis, functional linear model, significance test.

Spis treści

Wstęp **7**

1 Preliminaria **9**
1.1 Klasyfikacja operatorów liniowych 9
1.2 L^2-elementy losowe. Pojęcie funkcji średniej i operatora kowariancji 13
1.3 Estymacja średniej, funkcji kowariancji i operatora kowariancji. FPC 18

2 Test istotności w funkcjonalnym modelu liniowym **21**
2.1 Funkcjonalny model liniowy 21
2.2 Procedura testowa 24
2.3 Rozkład statystyki testowej 26

3 Przykład zastosowania **33**
3.1 Opis danych magnetometrycznych 33
3.2 Ameryka Północna (Kanada) 34
3.3 Europa (Polska) 37

A Kod w R **39**

Bibliografia **43**

Wstęp

Analiza danych funkcjonalnych (ang. *Functional Data Analysis*)
W pracy przedstawiona zostanie teoria estymacji elementów losowych przyjmujących wartości w ośrodkowej (rzeczywistej) przestrzeni Hilberta $L^2(T)$, gdzie $T \subset \mathbb{R}$ jest przedziałem. Ponieważ elementy przestrzeni $L^2(T)$ są formalnie klasami abstrakcji funkcji równych prawie wszędzie, to powyższe podejście uniemożliwia rozważanie wartości obserwacji elementu losowego w ustalonym punkcie $t \in T$. Z kolei w praktyce mamy do dyspozycji dane historyczne będące wartościami wylosowanej funkcji w pewnej ilości punktów z przedziału T. Dlatego naturalne jest rozważanie jako wyjściowego obiektu procesu stochastycznego $\{X_t\}_{t\in T}$, a następnie związanie z nim $L^2(T)$-elementu losowego.
Praca opiera się głównie na artykule [Kokoszka et al.], który został rozwinięty w książce [Horváth, Kokoszka].

[pakiet w R: fda] W załączniku na końcu pracy załączony został kod napisany w języku R na potrzeby przykładu zaprezentowanego w pracy.

Ze względu na to, że analiza danych funkcjonalnych (*ang.* Functional Data Analysis, FDA) jest stosunkowo nowym działem statystyki i jest wciąż mało popularna w polskiej literaturze, wiele pojęć czy określeń zawartych w pracy nie posiada jeszcze ogólnie przyjętych polskich odpowiedników. Dlatego zostały one przetłumaczone przez autora według własnego uznania, przytaczając oryginalne (angielskie) nazwy.

W pracy wykorzystano dane o polu magnetycznym Ziemi publikowane na stronie programu INTERMAGNET oraz organizacji SuperMAG. Załączam zatem specjalne podziękowania:

ACKNOWLEDGEMENTS

The results presented in this paper rely on data collected at magnetic observatories. We thank the national institutes that support them and INTERMAGNET for promoting high standards of magnetic observatory practice (www.intermagnet.org).

For the ground magnetometer data we gratefully acknowledge: Intermagnet; USGS, Jeffrey J. Love; CARISMA, PI Ian Mann; CANMOS; The S-RAMP Database, PI K. Yumoto and Dr. K. Shiokawa; The SPIDR database; AARI, PI Oleg Troshichev; The MACCS program, PI M. Engebretson, Geomagnetism Unit of the Geological Survey of Canada; GIMA; MEASURE, UCLA IGPP and Florida Institute of Technology; SAMBA, PI Eftyhia Zesta; 210 Chain, PI K. Yumoto; SAMNET, PI Farideh Honary; The institutes who maintain the IMAGE magnetometer array, PI Eija Tanskanen; PENGUIN; AUTUMN, PI Martin Connors; DTU Space, PI Dr. Juergen Matzka; South Pole and McMurdo Magnetometer, PI's Louis J. Lanzarotti and Alan T. Weatherwax; ICESTAR; RAPIDMAG; PENGUIn; British Artarctic Survey; McMac, PI Dr. Peter Chi; BGS, PI Dr. Susan Macmillan; Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation (IZMIRAN); GFZ, PI Dr. Juergen Matzka; MFGI, PI B. Heilig; IGFPAS, PI J. Reda; University of L'Aquila, PI M. Vellante; SuperMAG, PI Jesper W. Gjerloev.

Rozdział 1

Preliminaria

Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną. Ω jest zatem zbiorem scenariuszy ω, $\mathcal{F}$ jest σ-algebrą podzbiorów Ω, a P miarą probabilistyczną na $\mathcal{F}$.

Definicja 1.1 *Niech B będzie przestrzenią Banacha. **σ-ciałem zbiorów borelowskich na $\mathbf{B}$** nazywamy σ-ciało $\mathcal{B}(B)$ generowane przez rodzinę zbiorów otwartych w normie przestrzeni B.*

Definicja 1.2 *Niech B będzie przestrzenią Banacha, zaś $(\Omega, \mathcal{F}, P)$ przestrzenią probabilistyczną. Odwzorowanie $X : \Omega \to B$ nazywamy **$\mathbf{B}$-elementem losowym**, gdy X jest mierzalne, tzn. dla każdego zbioru borelowskiego $A \in \mathcal{B}(B)$ zachodzi $X^{-1}(A) \in \mathcal{F}$.*

W pracy skupimy się na elementach losowych przyjmujących wartości w nieskończenie wymiarowej ośrodkowej (rzeczywistej) przestrzeni Hilberta. W środowisku statystyków przyjęło się aby, w przypadku gdy $X(\omega)$ jest funkcją (lub ogólnie krzywą), taką zmienną nazywać **zmienną funkcjonalną** (ang. *functional variable*), zaś obserwację χ zmiennej X nazywać **daną funkcjonalną** (ang. *functional data*). Statystyki takich obiektów są bardziej skomplikowane niż dla zmiennych losowych przyjmujących wartości w $\mathbb{R}$ lub $\mathbb{R}^n$ ($n \in \mathbb{N}$), dlatego, aby zbudować pojęcia funkcji średniej oraz operatora kowariancji dla zmiennych tego typu, wprowadzimy najpierw niezbędne pojęcia z dziedziny operatorów liniowych.

1.1 Klasyfikacja operatorów liniowych

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

$$\Psi \in \mathcal{L} \iff \exists_{C>0}\, \forall_{x \in H} \ \|\Psi x\| \leqslant C \|x\|.$$

Każdy operator $\Psi \in \mathcal{L}$ posiada **operator sprzężony** Ψ^*, zdefiniowany następująco

$$\langle \Psi^* x, y \rangle = \langle x, \Psi y \rangle.$$

Przestrzeń $\mathcal{L}$ z normą

$$\|\Psi\|_{\mathcal{L}} := \sup\{\|\Psi(x)\| : \ \|x\| \leqslant 1\} = \min\{C > 0 : \|\Psi x\| \leqslant C \|x\|, x \in H\}, \quad \Psi \in \mathcal{L}$$

jest przestrzenią Banacha.

Definicja 1.3 *[Horváth, Kokoszka]*
Operator $\Psi \in \mathcal{L}$ nazywamy ***operatorem zwartym****, jeśli istnieją dwie ortonormalne bazy w H $\{\upsilon_j\}_{j=1}^{\infty}$ i $\{f_j\}_{j=1}^{\infty}$, oraz ciąg liczb rzeczywistych $\{\lambda_j\}_{j=1}^{\infty}$ zbieżny do zera, takie że*

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

Klasę operatorów zwartych oznacza się przez $\mathcal{C}$.

Bez straty ogólności możemy założyć, że w przedstawionej reprezentacji λ_j są wartościami dodatnimi, w razie konieczności wystarczy f_j zamienić na $-f_j$.
Równoważną definicją operatora zwartego jest spełnienie przez Ψ następującego warunku: zbieżność $\langle y, x_n \rangle \to \langle y, x \rangle$ dla każdego $y \in H$ implikuje $\| \Psi(x_n) - \Psi(x) \| \to 0$.

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

Definicja 1.4 *[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 $\sum_{j=1}^{\infty} \lambda_j^2 < \infty$.*

Przytoczymy teraz *tożsamość Parsevala*, z której wielokrotnie będziemy korzystać w pracy.

Lemat 1.1 *(Tożsamość Parsevala)*
Niech $\{e_j\}_{j=1}^{\infty}$ będzie bazą ortonormalną w przestrzeni Hilberta H. Wtedy dla każdego $x \in H$ mamy $\|x\|^2 = \sum_{j=1}^{\infty} |\langle x, e_j \rangle|^2$.

Uwaga 1.1 *[Bosq], [Horváth, Kokoszka]*
Klasa $\mathcal{S}$ jest przestrzenią Hilberta z iloczynem skalarnym

$$\langle \Psi_1, \Psi_2 \rangle_{\mathcal{S}} := \sum_{j=1}^{\infty} \langle \Psi_1(e_j), \Psi_2(e_j) \rangle, \tag{1.2}$$

gdzie $\{e_j\}_{j=1}^{\infty}$ jest dowolną bazą ortonormalną w H.
Powyższy iloczyn skalarny zadaje normę

$$\| \Psi \|_{\mathcal{S}} := \left(\sum_{j=1}^{\infty} \lambda_j^2 \right)^{1/2},$$

co wynika z szeregu równości

$$\begin{aligned} \| \Psi \|_{\mathcal{S}}^2 = \langle \Psi, \Psi \rangle_{\mathcal{S}} &= \sum_{n=1}^{\infty} \left\langle \sum_{j=1}^{\infty} \lambda_j \langle e_n, \upsilon_j \rangle f_j, \sum_{k=1}^{\infty} \lambda_k \langle e_n, \upsilon_k \rangle f_k \right\rangle \\ &= \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \lambda_j \lambda_k \langle e_n, \upsilon_j \rangle \langle e_n, \upsilon_k \rangle \langle f_j, f_k \rangle = \sum_{n=1}^{\infty} \sum_{j=1}^{\infty} \lambda_j^2 \langle e_n, \upsilon_j \rangle^2 \\ &= \sum_{j=1}^{\infty} \lambda_j^2 \sum_{n=1}^{\infty} \langle e_n, \upsilon_j \rangle^2 \overset{\text{tożsamość Parsevala}}{=} \sum_{j=1}^{\infty} \lambda_j^2 \|\upsilon_j\|^2 = \sum_{j=1}^{\infty} \lambda_j^2. \end{aligned}$$

□

Definicja 1.5 *[Bosq]*
Zwarty operator liniowy nazywamy ***operatorem śladowym*** *(ang. nuclear operator), jeśli równość (1.1) spełniona jest dla ciągu $\{\lambda_j\}_{j=1}^{\infty}$ takiego, że $\sum_{j=1}^{\infty} |\lambda_j| < \infty$.*

Uwaga 1.2 *[Bosq]*
Klasa operatorów śladowych $\mathcal{N}$ z normą $\|\Psi\|_{\mathcal{N}} := \sum_{j=1}^{\infty} |\lambda_j|$ jest przestrzenią Banacha.

Uwaga 1.3 *[Bosq]*
Prawdziwe są inkluzje: $\mathcal{N} \subset \mathcal{S} \subset \mathcal{C} \subset \mathcal{L}$.

Definicja 1.6 *[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*** *(lub połowicznie pozytywnie określonym, ang. positive semidefinite), jeśli*

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

Uwaga 1.4 *[Horváth, Kokoszka]*
Symetryczny nieujemnie określony operator Hilberta-Schmidta Ψ możemy przedstawić w postaci

$$\Psi(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j\rangle v_j, \quad x \in H, \tag{1.3}$$

gdzie ortonormalne v_j są ***funkcjami (wektorami) własnymi*** *Ψ, a λ_j odpowiadającymi im* ***wartościami własnymi****, tj. $\Psi(v_j) = \lambda_j v_j$. Funkcje v_j mogą być rozszerzone do bazy, przez dodanie bazy ortonormalnej dopełnienia ortogonalnego podprzestrzeni rozpiętej przez oryginalne v_j. Możemy zatem założyć, że funkcje v_j w* (1.3) *tworzą bazę, a pewne wartości λ_j mogą być równe zero.*

W dalszej części pracy ograniczymy się do przypadku $H = L^2(T, \mathcal{B}(T), \lambda)$.

Na przedziale $T \subset \mathbb{R}$ rozważmy σ-algebrę zbiorów borelowskich $\mathcal{B}(T)$ wraz z miarą Lebesgue'a λ. Przestrzeń $L^2 = L^2(T) = L^2(T, \mathcal{B}, \lambda)$ nad przedziałem T jest zbiorem mierzalnych funkcji rzeczywistych całkowalnych z kwadratem określonych na T, tj.

$$x \in L^2(T) \iff x : T \to \mathbb{R} \ \wedge \int_T x^2(t)dt < \infty,$$

z utożsamieniem funkcji równych prawie wszędzie. Przestrzeń L^2 jest ośrodkową przestrzenią Hilberta z iloczynem skalarnym

$$\langle x, y\rangle := \int_T x(t)y(t)dt, \quad x, y \in L^2,$$

wyznaczającym normę

$$\|x\|^2 = \langle x, x\rangle = \int x^2(t)dt, \quad x \in L^2.$$

Tak jak zwyczajowo zapisujemy L^2 zamiast $L^2(T)$, tak w przypadku symbolu całki bez wskazania obszaru całkowania będziemy mieć na myśli całkowanie po całym przedziale T. 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. Przedstawimy pomocnicze definicje i twierdzenia, a następnie twierdzenie opisujące warunki, które powinny być spełnione, aby taki operator był dobrze określony.

Definicja 1.7 *[Beśka] Niech $(T, \mathcal{A})$ i $(S, \mathcal{C})$ będą przestrzeniami mierzalnymi.* ***σ-algebrę produktową*** *na $T \times S$ określa wzór*

$$\mathcal{A} \otimes \mathcal{C} = \sigma(\{A \times C : A \in \mathcal{A},\ C \in \mathcal{C}\}).$$

Twierdzenie 1.1 *[Beśka]*
Niech $(T, \mathcal{A}, \mu)$ i $(S, \mathcal{C}, \nu)$ będą przestrzeniami z miarami σ-skończonymi. Wówczas istnieje jedyna miara na $\mathcal{A} \otimes \mathcal{C}$ oznaczana symbolem $\mu \times \nu$ taka, że

$$(\mu \times \nu)(A \times C) = \mu(A)\nu(C), \quad A \in \mathcal{A},\ C \in \mathcal{C}.$$

Taką miarę nazywamy ***miarą produktową****.*

Twierdzenie 1.2 *[Billingsley] (Twierdzenie Fubiniego)*
Niech $(T, \mathcal{A}, \mu)$ i $(S, \mathcal{C}, \nu)$ będą przestrzeniami z miarami σ-skończonymi. Niech $f : T \times U \to \mathbb{R}$ będzie funkcją mierzalną względem σ-algebry produktowej $\mathcal{A} \otimes \mathcal{C}$.

(a) Załóżmy, że całka $\int_S f(t,s)d\nu(s)$ istnieje dla μ-prawie każdego $t \in T$ oraz całka $\int_T f(t,s)d\mu(t)$ istnieje dla ν-prawie każdego $s \in S$. Wówczas funkcja $T \ni t \mapsto \int_S f(t,s)d\nu(s) \in \mathbb{R}$ jest $\mathcal{A}$-mierzalna i funkcja $S \ni s \mapsto \int_T f(t,s)d\mu(t) \in \mathbb{R}$ jest $\mathcal{C}$-mierzalna.

(b) Załóżmy, że przynajmniej jedna z całek jest skończona:

$$\int_{T\times S} |f| d\mu \otimes \nu, \ \int_T \left(\int_S |f(t,s)| d\nu(s) \right) d\mu(t), \ \int_S \left(\int_T |f(t,s)| d\mu(t) \right) d\nu(s).$$

Wtedy dla μ-prawie wszystkich $t \in T$ funkcja $f(t, \cdot) : S \to \mathbb{R}$ jest ν-skończenie całkowalna i dla ν-prawie wszystkich $s \in S$ funkcja $f(\cdot, s) : T \to \mathbb{R}$ jest μ-skończenie całkowalna. Ponadto, funkcja $T \ni t \mapsto \int_S f(t,s)d\nu(s) \in \mathbb{R}$ jest μ-skończenie całkowalna i funkcja $S \ni s \mapsto \int_T f(t,s)d\mu(t) \in \mathbb{R}$ jest ν-skończenie całkowalna. Prawdziwe są poniższe równości

$$\int_{T\times S} f d\mu \otimes \nu = \int_T \left(\int_S f(t,s) d\nu(s) \right) d\mu(t) = \int_S \left(\int_T f(t,s) d\mu(t) \right) d\nu(s).$$

Uwaga 1.5 *[Beśka]*
Zauważmy, że funkcje $T \ni t \mapsto \int_S f(t,s)d\nu(s) \in \mathbb{R}$, $S \ni s \mapsto \int_T f(t,s)d\mu(t) \in \mathbb{R}$ mogą nie być poprawnie określone dla wszystkich $t \in T$ oraz $s \in S$. Są one zdefiniowane μ- i ν-prawie wszędzie, co wystarcza, aby poprawnie zdefiniować ich całki.

Lemat 1.2 *[Hsing, Eubank]*
Niech (T, μ) będzie przestrzenią z miarą. Dla funkcji $\psi \in L^2(T \times T)$ zdefiniujmy

$$\Psi x(t) = \int_T \psi(t,s)x(s)d\mu(s), \quad x \in L^2,\ t \in T. \tag{1.4}$$

Wówczas $\Psi : L^2(T) \to L^2(T)$ jest ograniczonym operatorem liniowym spełniającym

$$\|\Psi\| \leqslant \left(\int_T \int_T |\psi(t,s)|^2 d\mu(t)d\mu(s) \right)^{1/2} = \|\psi\|_{L^2(T\times T)}.$$

Dowód. Pokażemy, że dla $x \in L^2(T)$ zachodzi $\Psi x \in L^2(T)$.
Oznaczmy $M := \Big(\iint |\psi(t,s)|^2 d\mu(t) d\mu(s) \Big)^{1/2}$. Mierzalność funkcji Ψx wynika z twierdzenia Fubiniego. Z nierówności Cauchy'ego-Schwarza mamy

$$\begin{aligned} \| \Psi x\|^2 &= \int | \Psi x(t)|^2 d\mu(t) = \int \Big| \int \psi(t,s) x(s) d\mu(s) \Big|^2 d\mu(t) \\ &\leqslant \int \Big(\int |\psi(t,s)|^2 d\mu(s) \cdot \int |x(s)|^2 d\mu(s) \Big) d\mu(t) \\ &= \iint |\psi(t,s)|^2 d\mu(s) d\mu(t) \cdot \int |x(s)|^2 d\mu(s) = M^2 \|x\|^2, \end{aligned}$$

więc $\Psi x \in L^2(T)$ oraz Ψ jest operatorem ograniczonym z normą $\| \Psi\| \leqslant M$. Liniowość operatora wynika z liniowości całki. □

Tak określony operator Ψ (wzór (1.4)) nazywamy **operatorem całkowym**, zaś funkcję ψ nazywamy **jądrem całkowym** operatora Ψ.

Uwaga 1.6 *[Horváth, Kokoszka]*
Operatory całkowe są operatorami Hilberta-Schmidta wtedy i tylko wtedy, gdy

$$\iint \psi^2(t,s) dt ds < \infty. \tag{1.5}$$

Ponadto zachodzi

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

Twierdzenie 1.3 *(Twierdzenie Mercera) [Horváth, Kokoszka]*
Niech operator Ψ będzie operatorem całkowym spełniającym (1.5)*. Jeśli ponadto jego jądro całkowe ψ spełnia $\psi(s,t) = \psi(t,s)$ oraz $\iint \psi(t,s)x(t)x(s)dtds \geqslant 0$, to operator całkowy Ψjest symetryczny i nieujemnie określony, zatem z Uwagi 1.4 mamy*

$$\psi(t,s) = \sum_{j=1}^{\infty} \lambda_j \upsilon_j(t) \upsilon_j(s) \quad w\ L^2(T \times T),$$

gdzie λ_j, υ_j są odpowiednio wartościami własnymi i funkcjami własnymi operatora Ψ. Jeżeli funkcja ψ jest ciągła, powyższe rozwinięcie jest prawdziwe dla wszystkich $t, s \in T$ i szereg jest zbieżny jednostajnie.

1.2 L^2-elementy losowe. Pojęcie funkcji średniej i operatora kowariancji

W pracy przedstawiona zostanie teoria estymacji elementów losowych przyjmujących wartości w ośrodkowej (rzeczywistej) przestrzeni Hilberta $L^2(T)$, gdzie $T \subset \mathbb{R}$ jest przedziałem. Ponieważ elementy przestrzeni $L^2(T)$ są formalnie klasami abstrakcji funkcji równych prawie wszędzie, to powyższe podejście uniemożliwia rozważanie wartości obserwacji elementu losowego w ustalonym punkcie $t \in T$. Z kolei w praktyce mamy do dyspozycji dane historyczne będące wartościami wylosowanej funkcji w pewnej ilości punktów z przedziału T. Dlatego naturalne jest rozważanie jako wyjściowego obiektu procesu stochastycznego $\{X_t\}_{t \in T}$, a następnie związanie z nim $L^2(T)$-elementu losowego. W tym celu potrzebujemy warunku który zagwarantuje, że odwzorowanie $\Omega \ni \omega \mapsto X(\omega, \cdot)$ będzie $L^2(T)$-elementem losowym.

Definicja 1.8 *Niech $T \subset \mathbb{R}$ będzie przedziałem, a $(\Omega, \mathcal{F}, P)$ przestrzenią probabilistyczną. Proces stochastyczny $\{X_t\}_{t \in T}$ nazywamy* ***mierzalnym****, gdy jest mierzalny jako odwzorowanie z przestrzeni mierzalnej $(\Omega \times T, \mathcal{F} \otimes \mathcal{B}(T))$ w przestrzeń $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

Lemat 1.3 *[Hsing, Eubank]*
Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną, a H ośrodkową przestrzenią Hilberta. Odwzorowanie $X : \Omega \to H$ jest H-elementem losowym wtedy i tylko wtedy, gdy dla każdego $y \in H$ odwzorowanie $\Omega \ni \omega \mapsto \langle y, X(\omega)\rangle \in \mathbb{R}$ jest mierzalne (rozważając na $\mathbb{R}$ σ-ciało zbiorów borelowskich).

Dowód. Załóżmy najpierw, że X jest H-elementem losowym. Ponieważ dla dowolnego $y \in H$ odwzorowanie $H \ni x \mapsto \langle y, x\rangle \in \mathbb{R}$ jest ciągłe, to jest także mierzalne (rozpatrując σ-ciała borelowskie zarówno na H oraz $\mathbb{R}$). W takim razie odwzorowanie $\omega \mapsto \langle y, X(\omega)\rangle$ jest mierzalne jako złożenie odwzorowań mierzalnych.
Aby przeprowadzić dowód w drugą stronę wystarczy pokazać, że przeciwobrazy kul domkniętych w H są mierzalne (ponieważ generują one $\mathcal{B}(H)$). Ustalmy w tym celu $h \in H$ oraz $\varepsilon > 0$ oraz bazę ortonormalną $\{e_j\}_{j=1}^{\infty}$ przestrzeni H. Z tożsamości Parsevala mamy

$$\{x \in H : \|x - h\| \leqslant \varepsilon\} = \{x \in H : \|x - h\|^2 \leqslant \varepsilon^2\} = \{x \in H : \sum_{j=1}^{\infty} \langle e_j, x - h\rangle^2 \leqslant \varepsilon^2\}$$

$$= \{x \in H : \sum_{j=1}^{\infty} \left(\langle e_j, x\rangle - \langle e_j, h\rangle\right)^2 \leqslant \varepsilon^2\}.$$

Podobnie

$$X^{-1}(\{x \in H : \|x - h\| \leqslant \varepsilon\}) = \{\omega \in \Omega : \|X(\omega) - h\| \leqslant \varepsilon\}$$

$$= \{\omega \in \Omega : \sum_{j=1}^{\infty} \left(\langle e_j, X(\omega)\rangle - \langle e_j, h\rangle\right)^2 \leqslant \varepsilon^2\}$$

Ponieważ wszystkie odwzorowania $\omega \mapsto \langle e_j, X(\omega)\rangle$ są mierzalne z założenia, a skończone sumy oraz granice funkcji mierzalnych są mierzalne, to mierzalne jest także odwzorowanie $\omega \mapsto \sum\limits_{j=1}^{\infty} \left(\langle e_j, X(\omega)\rangle - \langle e_j, h\rangle\right)^2$, więc (na mocy powyższej równości) zbiór $X^{-1}(\{x \in H : \|x - h\| \leqslant \varepsilon\})$ jest mierzalny. □

Podamy teraz kryterium gwarantujące, że proces stochastyczny zada L^2-element losowy.

Lemat 1.4 *[Hsing, Eubank]*
Niech $\{X_t\}_{t\in T}$ będzie mierzalnym procesem stochastycznym. Jeśli dla każdego $\omega \in \Omega$ funkcja $X(\omega, \cdot)$ jest całkowalna z kwadratem (względem miary Lebesgue'a na T), to odwzorowanie $\Omega \ni \omega \mapsto X(\omega, \cdot) \in L^2(T)$ jest L^2-elementem losowym (gdzie $X(\omega, \cdot)$ rozumiemy już jako klasę abstrakcji funkcji równych prawie wszędzie).

Dowód. Założenie o całkowalności z kwadratem gwarantuje, że funkcja $X(\omega, \cdot)$ rzeczywiście należy do $L^2(T)$. Wystarczy zatem wykazać mierzalność odwzorowania. Skorzystamy w tym celu z Lematu 1.3. Weźmy $y \in L^2(T)$. Ponieważ odwzorowanie $\omega \mapsto \langle X(\omega, \cdot), y\rangle = \int X(\omega, t)y(t)dt$ nie zmieni się gdy funkcje podcałkowe zmienimy na zbiorze miary zero, to y możemy traktować jako reprezentanta klasy abstrakcji. Skoro $X : \Omega \times T \to \mathbb{R}$ jest mierzalne względem σ-ciała produktowego $\mathcal{F} \otimes \mathcal{B}(T)$, a $y : T \to \mathbb{R}$ jest mierzalne względem $\mathcal{B}(T)$, to odwzorowanie $\Omega \times T \ni (\omega, t) \mapsto X(\omega, t)y(t)$ także jest mierzalne względem $\mathcal{F} \otimes \mathcal{B}(T)$. Z twierdzenia Fubiniego (Twierdzenie 1.2) wynika teraz, że odwzorowanie $\omega \mapsto \int X(\omega, t)y(t)dt = \langle X(\omega, \cdot), y\rangle$ jest mierzalne. Z Lematu 1.3 otrzymujemy, że odwzorowanie $\omega \mapsto X(\omega, \cdot) \in L^2(T)$ jest L^2-elementem losowym. □

Definicja 1.9 *Niech X będzie $L^2(T)$-elementem losowym. Mówimy, że X jest* ***całkowalny****, jeśli $\mathbb{E}\,\|X\| = \mathbb{E}\left[\int X^2(t)dt\right]^{1/2} < \infty$.*

Zauważmy, że jeśli X jest L^2-elementem losowym, to odwzorowanie $\omega \mapsto \|X(\omega)\|$ jest mierzalne (gdyż odwzorowanie $L^2 \ni h \mapsto \|h\|$ jest ciągłe), więc można rozważać powyższą wartość oczekiwaną. Wprowadzimy teraz pojęcie wartością oczekiwnej dla L^2-elementu losowego.

Definicja 1.10 *Niech X będzie całkowalnym $L^2(T)$-elementem losowym. Jedyny element $\mu \in L^2$ taki, że $\langle y, \mu\rangle = \mathbb{E}\langle y, X\rangle$ dla dowolnego $y \in L^2$ nazywamy* ***wartością oczekiwaną (funkcją średniej)*** *elementu losowego X. Ozn. $\mathbb{E}X := \mu$.*

Aby uzasadnić powyższa definicję, zauważmy najpierw, że odwzorowanie $\omega \mapsto \langle y, X(\omega)\rangle$ jest zmienną losową na mocy Lematu 1.3. Odwzorwanie $f : L^2 \to \mathbb{R}$ zadane jako $f(y) := \mathbb{E}\langle y, X\rangle$ jest ograniczonym funkcjonałem liniowym na L^2. Liniowość wynika z liniowości wartości oczekiwanej oraz iloczynu skalarnego, zaś ograniczoność z nierówności Cauchy'ego-Schwarza:

$$|f(y)| = |\mathbb{E}\langle y, X\rangle| \leqslant \mathbb{E}|\langle y, X\rangle| \leqslant \mathbb{E}\|y\|\|X\| = \|y\| \cdot \mathbb{E}\left[\int X^2(t)dt\right]^{1/2} = \mathbb{E}\,\|X\| \cdot \|y\|.$$

Istnienie i jednoznaczność funkcji $\mu \in L^2$ takiej, że $f(y) = \langle y, \mu\rangle$ wynika teraz z twierdzenia Riesza o reprezentacji funkcjonału liniowego ciągłego na przestrzeni Hilberta.
Wartość oczekiwana jest przemienna z operatorami ograniczonymi, tj. jeśli X jest całkowalna oraz $\Psi \in \mathcal{L}$, to $\Psi(X)$ także jest całkowalna (gdyż $\mathbb{E}\|\Psi(X)\| \leqslant \mathbb{E}\|\Psi\|\|X\| = \|\Psi\| \cdot \mathbb{E}\|X\| < \infty$) oraz mamy $\mathbb{E}\,\Psi(X) = \Psi(\mathbb{E}X)$. Istotnie, niech $\mu = \mathbb{E}X$ oraz $\nu = \mathbb{E}\,\Psi(X)$. Wówczas dla dowolnego $y \in L^2$ mamy

$$\langle y, \nu\rangle = \mathbb{E}\langle y, \Psi(X)\rangle = \mathbb{E}\langle \Psi^* y, X\rangle = \langle \Psi^* y, \mu\rangle = \langle y, \Psi\mu\rangle,$$

gdzie Ψ^* oznacza operator sprzężony do operatora Ψ. W takim razie $\Psi(\mathbb{E}X) = \Psi\mu = \nu = \mathbb{E}\,\Psi(X)$.
Jeśli dodatkowo założymy, że $\{X_t\}_{t\in T}$ jest mierzalnym procesem stochastycznym o całkowalnych z kwadratem trajektoriach (więc, z Lematu 1.4 zadaje L^2-element losowy) oraz takim, że $\int |\mathbb{E}X(t)|^2 dt < \infty$, to funkcję średniej zadanego przez niego L^2-elementu losowego możemy znaleźć wprost jako $\mu(t) = \mathbb{E}X(t)$ dla prawie wszystkich $t \in T$. Po pierwsze zauważmy, że dodatkowe założenie gwarantuje, że funkcja $t \mapsto \mathbb{E}X(t)$ należy do $L^2(T)$ (w szczególności wartość oczekiwana $\mathbb{E}X(t)$ istnieje dla prawie każdego $t \in T$). Dla dowolnego $y \in L^2$ zachodzi (na mocy tw. Fubiniego, 1.2)

$$\int y(t)\mathbb{E}X(t)dt = \mathbb{E}\int y(t)X(t)dt = \mathbb{E}\langle y, X\rangle,$$

więc funkcja $t \mapsto EX(t)$ spełnia definicję bycia wartością oczekiwaną elementu losowego $\omega \mapsto X(\omega, \cdot)$. Z uwagi na jednoznaczność wartości oczekiwanej zachodzi $\mu(t) = \mathbb{E}X(t)$. Na koniec zauważmy, że jeśli proces $\{X_t\}_{t\in T}$ spełnia $\mathbb{E}\|X\|^2 < \infty$ (założenie to będzie obowiązywało w dalszej części pracy), to spełnia także $\int |\mathbb{E}X(t)|^2 dt < \infty$, gdyż (z nierówności Jensena oraz ponownie twierdzenia Fubiniego)

$$\int |\mathbb{E}X(t)|^2 dt \leqslant \int \left(\mathbb{E}|X(t)|\right)^2 dt \leqslant \int \mathbb{E}(|X(t)|^2) dt = \mathbb{E}\int |X(t)|^2 dt = \mathbb{E}\|X\|^2 < \infty.$$

Definicja 1.11 *[Bosq]*
Operator kowariancji *całkowalnego $L^2(T)$-elementu losowego X o funkcji średniej μ_X spełniającego $\mathbb{E}\,\|X\|^2 < \infty$ definiujemy jako ograniczony operator liniowy według wzoru*

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

Jeśli Y jest $L^2(T)$-elementem losowym o funkcji średniej μ_Y spełniającym powyższe warunki, wtedy operator kowariancji między zmiennymi X i Y (ang. cross-covariance operator) przedstawiamy jako

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

oraz

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

Uzasadnimy, że powyższe operatory są dobrze określonymi operatorem ograniczonymi na L^2. Wystarczy to zrobić dla $C_{X,Y}$, gdyż $C_X = C_{X,X}$. W pierwszej kolejności należy sprawdzić, że $Z := \langle X - \mu_X, x\rangle (Y - \mu_Y)$ jest L^2-elementem losowym. Z Lematu 1.3 wystarczy sprawdzić, że dla każdego $z \in L^2$ funkcja $\omega \mapsto \langle Z(\omega), z\rangle$ jest zmienną losową. Mamy

$$\langle Z, z\rangle = \langle \langle X - \mu_X, x\rangle (Y - \mu_Y), z\rangle = \langle X - \mu_X, x\rangle \langle Y - \mu_Y, z\rangle$$

$$= \langle X, x\rangle \langle Y, z\rangle - \langle X, x\rangle \langle \mu_Y, z\rangle - \langle \mu_X, x\rangle \langle Y, z\rangle + \langle \mu_X, x\rangle \langle \mu_Y, z\rangle.$$

Skoro X oraz Y są L^2-elementami losowymi, to $\langle X, x\rangle$ oraz $\langle Y, z\rangle$ są zmiennymi losowymi. Pozostałe wyrażenia są stałe, więc ostatecznie $\langle Z, z\rangle$ jest zmienną losową. Z jest całkowalny, gdyż (z nierówności Cauchy'ego-Schwarza)

$$\mathbb{E}\|Z\| = \mathbb{E}\|\langle X - \mu_X, x\rangle (Y - \mu_Y)\| \leqslant \mathbb{E}\Big[|\langle X - \mu_X, x\rangle| \cdot \|(Y - \mu_Y)\|\Big]$$

$$\leqslant \|x\| \cdot \mathbb{E}\Big[\|X - \mu_X\| \cdot \|(Y - \mu_Y)\|\Big] \leqslant \|x\| \cdot \mathbb{E}\Big[\big(\|X\| + \|\mu_X\|\big)\big(\|Y\| + \|\mu_Y\|\big)\Big]$$

$$\leqslant \|x\| \cdot \Big[\mathbb{E}\|X\|\|Y\| + \|\mu_X\|\mathbb{E}\|Y\| + \|\mu_Y\|\mathbb{E}\|X\| + \|\mu_X\|\|\mu_Y\|\Big].$$

Skoro $\mathbb{E}\|X\|^2$, $\mathbb{E}\|Y\|^2 < \infty$, to także $\mathbb{E}\|X\|\|Y\|$, $\mathbb{E}\|X\|$, $\mathbb{E}\|Y\| < \infty$, więc zachodzi też $\mathbb{E}\|Z\| < \infty$. W takim razie, $C_{X,Y}(x) = \mathbb{E}Z$ istnieje i należy do $L^2(T)$. Powyższy rachunek pokazuje także, że operator $C_{X,Y}$ jest ograniczony, zaś jego liniowość wynika z liniowości iloczynu skalarnego oraz wartości oczekiwanej (liniowość wartości oczekiwanej dla L^2-elementów losowych wynika wprost z definicji).

[twierdzenie spinające właściwości operatora kowariancji?]

Pokażemy, że operatorem sprzężonym do $C_{X,Y}$ jest $C_{Y,X}$. Wystarczy sprawdzić, ze dla dowolnych $x, y \in L^2$ zachodzi $\langle C_{X,Y}(x), y\rangle = \langle x, C_{Y,X}(y)\rangle$. Dla uproszczenia załóżmy, że $\mathbb{E}X = \mathbb{E}Y = 0$. Korzystając z definicji wartości oczekiwanej dla elementu losowego mamy

$$\langle C_{X,Y}(x), y\rangle = \langle \mathbb{E}\langle X, x\rangle Y, y\rangle = \mathbb{E}\langle \langle X, x\rangle Y, y\rangle = \mathbb{E}\langle X, x\rangle \langle Y, y\rangle$$

$$= \mathbb{E}\langle \langle Y, y\rangle X, x\rangle = \mathbb{E}\langle x, \langle Y, y\rangle X\rangle = \langle x, \mathbb{E}\langle Y, y\rangle X\rangle = \langle x, C_{Y,X}(y)\rangle.$$

Załóżmy teraz ponownie, że $\{X_t\}_{t\in T}$ jest mierzalnym procesem stochastycznym takim, że $\mathbb{E}\|X\|^2 < \infty$ z funkcją średniej $\mu \in L^2(T)$. Wówczas operator kowariancji jest operatorem całkowym, czyli

$$C_X(x)(t) = \int c(t,s)x(s)ds,$$

z jądrem całkowym $c(t,s)$ zdefiniowanym następująco:

$$c(t,s) = \mathbb{E}\Big[\big(X(t) - \mu(t)\big)\big(X(s) - \mu(s)\big)\Big].$$

Zauważmy, że mierzalność procesu implikuje, że funkcja $c(t,s)$ jest mierzalna na produkcie $T \times T$ (twierdzenie Fubiniego). Najpierw pokażemy, że jądro $c(t,s)$ podanej postaci

rzeczywiście zadaje ograniczony operator liniowy na L^2. W tym celu, na mocy Lematu 1.2, wystarczy sprawdzić, że $c \in L^2(T \times T)$. Mamy

$$\iint |c(t,s)|^2 dtds = \iint \Big|\mathbb{E}\big(X(t)-\mu(t)\big)\big(X(s)-\mu(s)\big)\Big|^2 dtds$$

$$\leqslant \iint \Big(\mathbb{E}|X(t)-\mu(t)||X(s)-\mu(s)|\Big)^2 dtds = (*).$$

Zauważmy, że dla prawie każdego $t \in T$ zmienna losowa $X(t)$ jest całkowalna z kwadratem, tzn. $X(t) \in L^2(\Omega, \mathcal{F}, P)$. Wynika to z faktu, że

$$\int \mathbb{E}X^2(t)dt = \mathbb{E}\int X^2(t)dt = \mathbb{E}\|X\|^2 < \infty,$$

więc funkcja $t \mapsto \mathbb{E}X^2(t)$ musi być skończona prawie wszędzie. W takim razie także $\big(X(t)-\mu(t)\big) \in L^2(\Omega, \mathcal{F}, P)$ i możemy skorzystać z nierówności Cauchy'ego-Schwarza:

$$(*) \leqslant \iint \mathbb{E}|X(t)-\mu(t)|^2 \mathbb{E}|X(s)-\mu(s)|^2 dtds = \Big(\int \mathbb{E}|X(t)-\mu(t)|^2 dt\Big)^2$$

$$= \Big(\mathbb{E}\int |X(t)-\mu(t)|^2 dt\Big)^2 = \Big(\mathbb{E}\|X-\mu\|^2\Big)^2 \leqslant \Big(\mathbb{E}\|X\|^2 + \|\mu\|\mathbb{E}\|X\| + \|\mu\|^2\Big)^2 < \infty.$$

Pokażemy teraz, że przy powyższych założeniach operator kowariancji istotnie jest operatorem całkowym z jądrem c. Ponieważ mamy do czynienia z L^2-elementem losowym pochodzącym od mierzalnego procesu stochastycznego, to wartość oczekiwaną elementu losowego $Z := \langle X-\mu, x\rangle(X-\mu)$ możemy liczyć punktowo (por. uwaga po Definicji 1.10), czyli dla $x \in L^2(T)$ zachodzi dla prawie wszystkich $t \in T$:

$$\begin{aligned} C_X(x)(t) &= \mathbb{E}\langle X-\mu, x\rangle(X(t)-\mu(t)) \\ &= \mathbb{E}\Big[\int (X(s)-\mu(s))x(s)ds\Big]\big(X(t)-\mu(t)\big) \\ &= \mathbb{E}\Big[\int \big(X(s)-\mu(s)\big)\big(X(t)-\mu(t)\big)x(s)ds\Big] \\ &= \int \Big[\mathbb{E}\big(X(s)-\mu(s)\big)\big(X(t)-\mu(t)\big)\Big]x(s)ds \\ &= \int c(t,s)x(s)ds, \end{aligned}$$

zatem funkcja c jest jądrem operatora C_X, który, jako operator całkowy, jest operatorem Hilberta-Schmidta (Uwaga 1.6). Oczywistym jest, że $c(t,s) = c(s,t)$ i mamy

$$\begin{aligned} \iint c(t,s)x(t)x(s)dtds &= \iint \mathbb{E}\big[\big(X(t)-\mu(t)\big)\big(X(s)-\mu(s)\big)\big]x(t)x(s)dtds \\ &= \mathbb{E}\left[\left(\int \big(X(t)-\mu(t)\big)x(t)dt\right)^2\right] \geqslant 0. \end{aligned}$$

Zatem operator kowariancji C_X jest symetryczny oraz nieujemnie określony (Twierdzenie 1.3). Z Uwagi 1.4 wynika, że posiada on reprezentację

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

gdzie λ_j są wartościami własnymi operatora C_X (lub zerami), a v_j odpowiadającymi im wektorami własnymi tworzącymi bazę ortonormalną przestrzeni $L^2(T)$. Co więcej, spełnione jest

$$\begin{aligned}\lambda_j &= \lambda_j \|v_j\|^2 = \langle \lambda_j v_j, v_j\rangle = \langle C_X v_j, v_j\rangle = \langle \mathbb{E}\langle X-\mu, v_j\rangle (X-\mu), v_j\rangle \\ &= \int \mathbb{E}\Big[\Big(\int (X(s)-\mu(s))v_j(s)ds\Big)\Big(X(t)-\mu(t)\Big)\Big]v_j(t)dt \\ &= \mathbb{E}\iint \Big((X(s)-\mu(s))v_j(s)\Big)\Big((X(t)-\mu(t))v_j(t)\Big)dsdt \\ &= \mathbb{E}\left[\left(\int (X(t)-\mu(t))v_j(t)dt\right)^2\right] = \mathbb{E}\langle X-\mu, v_j\rangle^2.\end{aligned}$$

W takim razie wartości własne operatora kowariancji są nieujemne oraz tożsamość Parsevala pokazuje, że

$$\sum_{j=1}^{\infty}\lambda_j = \sum_{j=1}^{\infty}\mathbb{E}\langle X-\mu, v_j\rangle^2 = \mathbb{E}\sum_{j=1}^{\infty}\langle X-\mu, v_j\rangle^2 = \mathbb{E}\|X-\mu\|^2 < \infty.$$

Widzimy zatem, że operator C_X jest operatorem nuklearnym.

W dalszej części pracy będziemy skupiać się na **scentrowanych** elementach losowych X, tj. takich że $\mathbb{E}X = 0$, dlatego operator kowariancji przyjmować będzie postać

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

Niezależność elementów losowych przyjmujących wartości w przestrzeni Hilberta oznacza dokładnie to samo co w przypadku niezależności zmiennych losowych. W pracy wielokrotnie korzystać będziemy z założenia o niezależności oraz z następującej konsekwencji:

Lemat 1.5 *[Horváth, Kokoszka]*
Jeśli X_1 i X_2 są niezależnymi L^2-elementami losowymi takimi, że $\mathbb{E}X_1 = 0$ i $\mathbb{E}\|X_1\|^2 < \infty$ oraz $\mathbb{E}\|X_2\|^2 < \infty$, to $\mathbb{E}[\langle X_1, X_2\rangle] = 0$.

1.3 Estymacja średniej, funkcji kowariancji i operatora kowariancji. FPC

Naturalnym problemem pojawiającym się przy L^2-elementach losowych jest wnioskowanie o obiektach nieskończenie wymiarowych na podstawie skończonej próbki danych. Ze względu na fakt, że w statystyce ograniczamy się głównie do przypadku, w którym elementy losowe to funkcje gładkie lub krzywe, to nazywa się je także zmiennymi funkcjonalnymi, zaś obserwacje zmiennej funkcjonalnej - danymi funkcjonalnymi.
Przedmiotem funkcjonalnej analizy danych jest zatem ciąg $X_1, ..., X_N$ niezależnych danych funkcjonalnych w $L^2(T)$ o jednakowym rozkładzie jak zmienna funkcjonalna $X \in L^2$ spełniająca założenie $\mathbb{E}\|X\|^2 < \infty$. W dowolnie wybranej bazie $\{v_j\}_{j\geqslant 1}$ w L^2, $X_1, ..., X_n$ można przedstawić jako kombinacje liniowe funkcji bazowych, tj.

$$X_n(t) = \sum_{k=1}^{\infty}\xi_{kn}v_k(t), \quad n = 1, ..., N.$$

W praktyce obserwujemy tylko punkty z danych funkcjonalnych, tj. wartości z N nieznanych funkcji dla wybranych argumentów $t_1 < ... < t_n$ (niekoniecznie równomiernie rozłożonych) należących do przedziału T. Aby zatem otrzymać N danych funkcjonalnych

$X_1, ..., X_N$ należy znaleźć ich przybliżenie przez dopasowanie kombinacji liniowej tylko K funkcji bazowych $\{\upsilon_j\}_{j\geqslant 1}$ w L^2, czyli

$$X_n(t) \approx \sum_{k=1}^{K} \widehat{\xi}_{kn}\upsilon_k(t), \quad n = 1, ..., N,$$

gdzie $\widehat{\xi}_{kn}$ można znaleźć metodą najmniejszych kwadratów. Dla uproszczenia obliczeń najlepiej, kiedy baza $\{\upsilon_j\}_{j\geqslant 1}$ jest ortonormalna, jak np. baza Fouriera, ale możemy wykorzystać nawet bazę nieortogonalną, np. bazę tworzoną przez B-splajny.

Z punktu widzenia analizy danych funkcjonalnych parametrami koniecznymi do estymacji są funkcja średniej, funkcja kowariancji oraz operator kowariancji, określone następująco

funkcja średniej:	$\mu(t) = \mathbb{E}[X(t)]$;
funkcja kowariancji:	$c(t,s) = \mathbb{E}[(X(t) - \mu(t)(X(s) - \mu(s))]$;
operator kowariancji:	$C = \mathbb{E}[\langle (X - \mu), \cdot\rangle (X - \mu)]$.

Funkcję średniej μ estymujemy średnią z funkcji z próby

$$\widehat{\mu}(t) = \frac{1}{N}\sum_{n=1}^{N} X_n(t), \quad t \in T,$$

funkcję kowariancji ze wzoru

$$\hat{c}(t,s) = \frac{1}{N}\sum_{n=1}^{N} \big(X_n(t) - \hat{\mu}(t)\big)\big(X_n(s) - \hat{\mu}(s)\big), \quad t, s \in T,$$

zaś operator kowariancji estymujemy

$$\widehat{C}(x)(t) = \frac{1}{N}\sum_{n=1}^{N} \langle X_n - \hat{\mu}, x\rangle (X_n(t) - \hat{\mu}(t)), \quad x \in L^2,\ t \in T. \tag{1.6}$$

Zauważmy, że powyższa równość ilustruje wspomniany problem wnioskowania statystycznego o zmiennych funkcjonalnych. Estymator $\widehat{C}$ rzutuje L^2 na skończenie wymiarową podprzestrzeń generowaną przez $X_1, ..., X_N$, co ogranicza dokładność znalezienia obiektu nieskończenie wymiarowego posiadając skończoną próbę.
Niemniej jednak powyższe estymatory są dobrze określone, a estymator funkcji średniej jest estymatorem nieobciążonym. Dowody poprawności tych estymatorów można znaleźć w Rozdziale 2 w książce [Horváth, Kokoszka].

W dalszej części pracy istotne będzie dla nas oszacowanie również wartości i funkcji własnych operatora kowariancji C. W szczególności interesować nas będzie p największych wartości własnych λ_j spełniających

$$\lambda_1 > \lambda_2 > ... > \lambda_p > \lambda_{p+1} \geqslant 0. \tag{1.7}$$

Funkcje własne υ_j zdefiniowane są przez równanie $C\upsilon_j = \lambda_j\upsilon_j$. Zauważmy, że (z definicji operatora liniowego), jeśli υ_j jest funkcją własną, to również $a\upsilon_j$ jest funkcją własną, gdzie $a \neq 0$ jest skalarem. Funkcje własne υ_j są zazwyczaj normalizowane, tak aby $\|\upsilon_j\| = 1$. Przy estymacji może pojawić się problem ze znakiem $\widehat{\upsilon}_j$, dlatego wprowadzamy dodatkowe parametry $\widehat{c}_j = \text{sign}(\langle\hat{\upsilon}_j, \upsilon_j\rangle)$ tak aby $\widehat{c}_j\widehat{\upsilon}_j$ były możliwie blisko υ_j. Niemniej jednak $\widehat{c}_j$ nie są możliwe do uzyskania z danych, dlatego sposób estymacji funkcji własnych nie powinien zależeć od $\widehat{c}_j$. Wartości własne λ_j i funkcje własne υ_j estymujemy zatem według wzoru

$$\int \hat{c}(t,s)\hat{\upsilon}_j(s)ds = \hat{\lambda}_j\hat{\upsilon}_j(t), \quad j = 1, 2, ..., N.$$

Podamy teraz (bez dowodu) podstawową własność przyjętego sposobu estymacji wartości oraz wektorów własnych operatorów kowariancji - średniokwadratowy błąd estymacji maleje do zera szybciej niż liniowo:

Twierdzenie 1.4 *[Kokoszka et al.], [Bosq]*
Według powyższych oznaczeń, jeżeli dla pewnego $p > 0$ *prawdziwe jest* (1.7)*, to spełnione są nierówności*

$$\limsup_{N\to\infty} N\mathbb{E}\,\|\upsilon_j - \widehat{c}_j\widehat{\upsilon}_j\|^2 < \infty, \quad \limsup_{N\to\infty} N\mathbb{E}\Big[\Big|\lambda_j - \widehat{\lambda}_j\Big|^2\Big] < \infty,$$

dla $j \leqslant p$.

Zarysujemy teraz funkcjonalny odpowiednik analizy głównych składowych (ang. *principal components analysis, PCA*). Funkcje własne operatora kowariancji z próby $\widehat{C}$ nazywamy **empirycznymi funkcjonalnymi głównymi składowymi** (ang. *empirical functional principal components, EFPC's*) danych funkcjonalnych $X_1, ..., X_N$. Jeśli $X_1, ..., X_N$ mają taki sam rozkład co L^2-element losowy X (całkowalny z kwadratem, tj. spełniający $\mathbb{E}\|X\|^2 < \infty$), to funkcje własne operatora C nazywamy **funkcjonalnymi głównymi składowymi** (*FPC's*), które estymujemy przez EFPC z dokładnością do znaku.
Jak w przypadku zmiennej losowej i wektorów własnych macierzy kowariancji, tak w przypadku zmiennej funkcjonalnej funkcje własne operatora kowariancji, FPC's i EFPC's, tworzą bazę ortonormalną optymalną do rozwinięcia odpowiednio zmiennej funkcjonalnej X i danych funkcjonalnych $X_1, ...X_N$. Iloczyn skalarny $\langle X_i, \widehat{\upsilon}_j\rangle = \int X_i(t)\widehat{\upsilon}_j(t)dt$ odpowiada zmienności X_i opisanej przez υ_j i nazywany jest w anglojęzycznej literaturze ***score***. Istotnie, przy założeniu $\mathbb{E}X_i = 0$, statystyka

$$\frac{1}{N}\sum_{i=1}^{N}\langle X_i, x\rangle^2 = \langle\widehat{C}(x), x\rangle$$

może być traktowana jako wariancja z próby "w kierunku" funkcji x. ...

Rozdział 2

Test istotności w funkcjonalnym modelu liniowym

W tym rozdziale opiszemy funkcjonalny odpowiednik mieszanego modelu liniowego, a następnie test na jego efektywność, tj. sprawdzenie, czy operator stojący przy zmiennej objaśniającej jest istotnie różny od zera.

2.1 Funkcjonalny model liniowy

Załóżmy, że mamy dane scentrowane L^2-elementy losowe Y oraz X takie, że $\mathbb{E}\|X\|^2 < \infty$, $\mathbb{E}\|Y\|^2 < \infty$ dla których chcemy zbudować funkcjonalny model liniowy w którym Y jest zmienną objaśnianą a X zmienną objaśniającą. Obserwujemy próbę długości N, tzn. mamy dane ciągi $\{Y_n\}_{n=1}^N, \{X_n\}_{n=1}^N, \{\varepsilon_n\}_{n=1}^N$ takie że kolejne trójki $(Y_n, X_n, \varepsilon_n)$ są niezależne, X_n oraz ε_n są niezależne dla każdego $n \in \{1, ..., N\}$, $\mathbb{E}\varepsilon_n = 0$
Pełen model funkcjonalny (ang. *fully functional model*) przyjmuje postać

$$Y_n = \Psi X_n + \varepsilon_n, \quad n = 1, 2, ..., N, \tag{2.1}$$

gdzie błąd ε_n również należy do przestrzeni Hilberta $L^2(T)$. Operator $\Psi : L^2 \to L^2$ jest całkowym operatorem Hilberta-Schmidta, a zatem jądro całkowe $\psi(t, s)$ jest funkcją całkowalną z kwadratem na $T \times T$. Równość (2.1) rozumiemy zatem następująco

$$Y_n(t) = \int \psi(t, s) X_n(s) ds + \varepsilon_n(t), \quad n = 1, 2, ..., N. \tag{2.2}$$

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*) postaci

$$Y_n = \int \psi(s) X_n(s) ds + \varepsilon_n, \quad n = 1, 2, ..., N,$$

  w którym tylko zmienne objaśniające X_n są zmiennymi funkcjonalnymi,

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

$$Y_n(t) = \psi(t) x_n + \varepsilon_n(t), \quad n = 1, 2, ..., N,$$

  w którym zmienne objaśniające x_n są deterministycznymi skalarami.

Naturalnym problemem pojawiającym się przy funkcjonalnym modelu liniowym jest estymacja operatora Ψ należącego do nieskończenie wymiarowej przestrzeni na podstawie skończonej próbki danych. Niech $\{\eta_k\}_{k=1}^{\infty}$ i $\{\theta_l\}_{l=1}^{\infty}$ będą pewnymi ustalonymi bazami, niekoniecznie ortonormalnymi, np. bazami Fouriera lub splajnowymi. Ponadto, niech funkcje η_k dobrze przybliżają funkcje X_n, a θ_l dobrze przybliżają Y_n. Wtedy nieznane jądro ψ estymujemy według postaci

$$\widehat{\psi}(t,s) = \sum_{k=1}^{K}\sum_{l=1}^{L} p_{kl}\eta_k(s)\theta_l(t),$$

gdzie K i L są odpowiednio małymi liczbami wybranymi do wygładzenia przybliżenia X_n i Y_n. Możliwym jest znalezienie operatora, który daje idealne dopasowanie do danych (dla którego wszystkie różnice od próbki są równe zero), nie narzucając dodatkowych założeń. Wykorzystany w dalszej części pracy pakiet *fda*, do programu *R-project*, do znalezienia operatora Ψ stosuje metodę najmniejszych kwadratów, tj. przez minimalizację sumy kwadratów reszt

$$\sum_{n=1}^{N}\left\|Y_n - \int X_n(s)\widehat{\psi}(s,\cdot)\right\|^2,$$

ale przypomina on biały szum i jego interpretacja jest często problematyczna i niepraktyczna. Jednym ze sposobów na rozwiązanie tego problemu jest poszukiwanie operatora należącego do podprzestrzeni generowanej przez funkcje własne operatora kowariancji danych z próby, nazywane empirycznymi funkcjonalnymi głównymi składowymi (*EFPC's*), które zostały opisane w podrozdziale 1.3. Główne składowe odpowiadają istotnym czynnikom zmienności zmiennych, dobrze służą zatem do przybliżania ich wartości.
Niech X i Y będą scentrowanymi L^2-elementami losowymi o rozwinięciach

$$X(s) = \sum_{i=1}^{\infty}\xi_i\upsilon_i(s), \quad Y(t) = \sum_{j=1}^{\infty}\zeta_j u_j(t), \tag{2.3}$$

gdzie υ_i są funkcjonalnymi głównymi składowymi X, zaś u_j są funkcjonalnymi głównymi składowymi Y i

$$\xi_i = \langle X, \upsilon_i\rangle, \quad \zeta_j = \langle Y, u_j\rangle.$$

Lemat 2.1 *[Horváth, Kokoszka]*
Niech X, Y, ε będą scentrowanymi L^2-elementami losowymi. Załóżmy, że ε będzie niezależne od X i niech spełnione będzie równanie liniowe

$$Y(t) = \int \psi(t,s)X(s)ds + \varepsilon(t) \tag{2.4}$$

z jądrem $\psi(\cdot,\cdot)$ takim, że

$$\iint \psi^2(t,s)dtds < \infty. \tag{2.5}$$

Wtedy

$$\psi(t,s) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\frac{\mathbb{E}[\xi_l\zeta_k]}{\mathbb{E}[\xi_l^2]}u_k(t)\upsilon_l(s),$$

gdzie zbieżność jest w $L^2(T\times T)$.

Dowód. Skoro $\{\upsilon_i\}_{i\geqslant 1}$ i $\{u_j\}_{j\geqslant 1}$ są bazami w L^2, to ciąg funkcji $\{\upsilon_i(s)u_j(t),\ s,t\in T\}_{i,j\geqslant 1}$ tworzy bazę $L^2(T\times T)$. Korzystając z założenia (2.5) oraz z Uwagi 1.6 zauważmy, że operator Ψ jest operatorem Hilberta-Schmidta. Zatem jądro ψ posiada jednoznaczną reprezentację

$$\psi(t,s) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\psi_{kl}u_k(t)\upsilon_l(s), \tag{2.6}$$

a na mocy założenia (2.5) współczynniki ψ_{kl} spełniają

$$\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\psi_{kl}^2 = \iint \psi^2(t,s)dtds < \infty.$$

Podstawiając (2.3) i (2.6) do (2.4) otrzymujemy

$$\begin{aligned}
\sum_{j=1}^{\infty}\zeta_{jn}u_j(t) &= \int\Big[\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\psi_{kl}u_k(t)v_l(s)\sum_{i=1}^{\infty}\xi_i v_i(s)\Big]ds + \varepsilon(t)\\
&= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\psi_{kl}u_k(t)\sum_{i=1}^{\infty}\xi_i\int v_l(s)v_i(s)ds + \varepsilon(t)\\
&= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\psi_{kl}u_k(t)\sum_{i=1}^{\infty}\xi_i\langle v_l(s), v_i(s)\rangle + \varepsilon(t)\\
&= \sum_{k=1}^{\infty}\sum_{i=1}^{\infty}\psi_{ki}\xi_i u_k(t) + \varepsilon(t).
\end{aligned}$$

Mnożąc powyższe wyrażenie obustronnie przez $u_l(t)$, a następnie całkując po t otrzymujemy kolejne równości

$$\begin{aligned}
\int u_l(t)\sum_{j=1}^{\infty}\zeta_j u_j(t)dt &= \int u_l(t)\Big(\sum_{k=1}^{\infty}\sum_{i=1}^{\infty}\psi_{ki}\xi_i u_k(t) + \varepsilon(t)\Big)dt\\
\sum_{j=1}^{\infty}\zeta_j\int u_j(t)u_l(t)dt &= \sum_{k=1}^{\infty}\sum_{i=1}^{\infty}\psi_{ki}\xi_i\int u_k(t)u_l(t)dt + \int u_l(t)\varepsilon(t)dt\\
\sum_{j=1}^{\infty}\zeta_j\langle u_j, u_l\rangle &= \sum_{k=1}^{\infty}\sum_{i=1}^{\infty}\psi_{ki}\xi_i\langle u_k, u_l\rangle + \langle u_l, \varepsilon\rangle\\
\zeta_l &= \sum_{i=1}^{\infty}\psi_{li}\xi_i + \langle u_l, \varepsilon\rangle.
\end{aligned}$$

Wystarczy teraz pomnożyć obustronnie przez ξ_k, nałożyć wartość oczekiwaną na obie strony powyższej równości i skorzystać z Lematu 1.5, żeby otrzymać

$$\begin{aligned}
\mathbb{E}[\zeta_l\xi_k] &= \mathbb{E}\sum_{i=1}^{\infty}\psi_{li}\xi_i\xi_k + \mathbb{E}\xi_k\langle u_l, \varepsilon\rangle = \sum_{i=1}^{\infty}\psi_{li}\mathbb{E}\xi_i\xi_k = \sum_{i=1}^{\infty}\psi_{li}\mathbb{E}\langle X, v_i\rangle\langle X, v_k\rangle\\
&= \sum_{i=1}^{\infty}\psi_{li}\mathbb{E}\langle\langle X, v_i\rangle X, v_k\rangle = \sum_{i=1}^{\infty}\psi_{li}\langle\mathbb{E}\langle X, v_i\rangle X, v_k\rangle = \sum_{i=1}^{\infty}\psi_{li}\langle Cv_i, v_k\rangle\\
&= \sum_{i=1}^{\infty}\psi_{li}\lambda_i\langle v_i, v_k\rangle = \psi_{lk}\lambda_k\langle v_k, v_k\rangle = \psi_{lk}\mathbb{E}[\xi_k^2]
\end{aligned}$$

co kończy dowód. □

Zauważmy, że $\mathbb{E}[\xi_l^2] = \lambda_l$, gdzie λ_l jest wartością własną odpowiadającą funkcji własnej v_l. Bez zmiany reprezentacji (2.3) możemy pominąć z niej funkcje własne odpowiadające zerowym wartościom własnym i dzięki temu założyć, że $\mathbb{E}[\xi_l^2] > 0$ dla każdego $l \geqslant 1$. Zatem w Lemacie 2.1 warunek (2.5) można zastąpić przez

$$\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\frac{(\mathbb{E}[\xi_l\zeta_k])^2}{\lambda_l^2} < \infty.$$

Powyższe rozważania prowadzą do estymatora

$$\widehat{\psi}_{KL}(t,s) = \sum_{k=1}^{K}\sum_{l=1}^{L} \widehat{\lambda}_l^{-1}\widehat{\sigma}_{lk}\widehat{u}_k(t)\widehat{v}_l(s),$$

gdzie $\widehat{\sigma}_{lk}$ jest estymatorem $\mathbb{E}[\xi_l\zeta_k]$, czyli np.

$$\widehat{\sigma}_{lk} = \frac{1}{N}\sum_{i=1}^{N}\langle X_i, \widehat{v}_l\rangle\langle Y_i, \widehat{u}_k\rangle.$$

2.2 Procedura testowa

Jednym z podstawowych testów na efektywność modelu jest test istotności zmiennych objaśniających. Badamy zatem zerowanie się operatora Ψ, tj. hipotezy

$$\mathrm{H}_0: \quad \Psi = 0 \quad \text{przeciw} \quad \mathrm{H}_A: \quad \Psi \neq 0.$$

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

Załóżmy, że mamy dane scentrowane L^2-elementy losowe Y oraz X takie, że $\mathbb{E}\|X\|^2$, $\mathbb{E}\|Y\|^2 < \infty$ dla których chcemy zbudować funkcjonalny model liniowy w którym Y jest zmienną objaśnianą a X zmienną objaśniającą. Obserwujemy próbę długości N, tzn. mamy dane ciągi $\{Y_n\}_{n=1}^N, \{X_n\}_{n=1}^N, \{\varepsilon_n\}_{n=1}^N$ takie że kolejne trójki $(Y_n, X_n, \varepsilon_n)$ są niezależne, X_n oraz ε_n są niezależne dla każdego $n \in \{1, ..., N\}$, $\mathbb{E}\varepsilon_n = 0$ oraz Y_n ma taki sam rozkład jak Y, zaś X_n ma taki sam rozkład jak X. Dla zmiennych Y oraz X mamy zadane operatory kowariancji:

$$C(x) = \mathbb{E}[\langle X, x\rangle\, X], \quad \Gamma(x) = \mathbb{E}[\langle Y, x\rangle\, Y], \quad \Delta(x) = \mathbb{E}[\langle X, x\rangle\, Y], \quad x \in L^2. \tag{2.7}$$

Przez $\widehat{C}$, $\widehat{\Gamma}$, $\widehat{\Delta}$ oznaczamy ich estymatory (zgodnie z (1.6)), tj.

$$\widehat{C}(x) = \frac{1}{N}\sum_{n=1}^{N}\langle X_n, x\rangle\, X_n, \quad \widehat{\Gamma}(x) = \frac{1}{N}\sum_{n=1}^{N}\langle Y_n, x\rangle\, Y_n, \quad \widehat{\Delta}(x) = \frac{1}{N}\sum_{n=1}^{N}\langle X_n, x\rangle\, Y_n, \quad x \in L^2.$$

Definiujemy również wartości i wektory własne C i Γ

$$C(\upsilon_k) = \lambda_k\upsilon_k, \quad \Gamma(u_j) = \gamma_j u_j, \tag{2.8}$$

których estymatory będziemy oznaczać $(\widehat{\lambda}_k, \widehat{\upsilon}_k)$, $(\widehat{\gamma}_j, \widehat{u}_j)$.
Test obejmuje obcięcie powyższych operatorów na podprzestrzenie skończenie wymiarowe. Podprzestrzeń $\mathcal{V}_p = \mathrm{span}\{\upsilon_1, ..., \upsilon_p\}$ 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 $\widehat{C}$ tak, że $\widehat{\mathcal{V}}_p = \mathrm{span}\{\widehat{\upsilon}_1, ..., \widehat{\upsilon}_p\}$ zawiera najlepsze przybliżenie X_n. Analogicznie $\mathcal{U}_q = \mathrm{span}\{u_1, ..., u_q\}$ zawiera przybliżenia $\mathrm{span}\{Y_1, ..., Y_N\}$.

Z ogólnej postaci funkcjonalnego modelu liniowego

$$Y = \Psi X + \varepsilon$$

możemy wyprowadzić kolejne równości

$$\begin{aligned}\langle X, x\rangle Y &= \langle X, x\rangle(\,\Psi X + \varepsilon) = \langle X, x\rangle\,\Psi X + \langle X, x\rangle\varepsilon\\ \mathbb{E}\left[\langle X, x\rangle Y\right] &= \mathbb{E}\left[\langle X, x\rangle\,\Psi X\right] + \mathbb{E}\left[\langle X, x\rangle\varepsilon\right].\end{aligned}$$

Korzystając z definicji operatorów C oraz Δ (2.7), założenia, że Ψ jest operatorem ograniczonym (więc komutuje z wartością oczekiwaną) oraz z założenia o niezależności między X a ε zachodzi (Lemat 1.5)

$$\Delta(x) = \mathbb{E}[\langle X, x\rangle\, Y] = \mathbb{E}\left[\langle X, x\rangle\, \Psi X\right] + \mathbb{E}\left[\langle X, x\rangle \varepsilon\right]$$
$$= \mathbb{E}\left[\Psi(\langle X, x\rangle X)\right] = \Psi\left(\mathbb{E}\left[\langle X, x\rangle X\right]\right) = \Psi C(x).$$

W szczególności, dla funkcji własnych υ_k, $k \leqslant p$ prawdziwa jest równość

$$\Delta(\upsilon_k) = \Psi C(\upsilon_k) = \Psi(\lambda_k \upsilon_k) = \lambda_k\, \Psi(\upsilon_k),$$

więc $\lambda_k \neq 0$?!

$$\Psi(\upsilon_k) = \lambda_k^{-1} \Delta(\upsilon_k). \tag{2.9}$$

Stąd, Ψ zeruje się na $\operatorname{span}\{\upsilon_1, ..., \upsilon_p\}$ wtedy i tylko wtedy, gdy $\Delta(\upsilon_k) = 0$ dla każdego $k = 1, ..., p$. Zauważmy, że

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

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

$$\left\langle \widehat{\Delta}(\upsilon_k), u_j \right\rangle = 0, \quad k = 1, ..., p, \quad j = 1, ..., q. \tag{2.10}$$

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, $n = 1, ..., N$. Statystyka testowa powinna zatem sumować kwadraty iloczynów skalarnych (2.10). Celem kolejnego podrozdziału jest udowodnienie Twierdzenia 2.1 stanowiącego, że (przy założonym sposobie estymacji wartości oraz wektorów własnych operatora kowariancji) statystyka

$$\widehat{T}_N(p, q) = N \sum_{k=1}^{p} \sum_{j=1}^{q} \widehat{\lambda}_k^{-1} \widehat{\gamma}_j^{-1} \left\langle \widehat{\Delta}(\widehat{\upsilon}_k), \widehat{u}_j \right\rangle^2, \tag{2.11}$$

zbiega według rozkładu do rozkładu χ^2 z pq stopniami swobody.
Zachodzi

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

oraz $\lambda_k = \mathbb{E}\langle X, \upsilon_k\rangle^2$ i $\gamma_j = \mathbb{E}\langle Y, u_j\rangle^2$ (por. Rozdział 1.2).

Uwaga 2.1 *Oczywistym jest, że jeśli odrzucamy H_0, to $\Psi(\upsilon_k) \neq 0$ dla pewnego $k \geqslant 1$. Jednak ograniczając się do p największych wartości własnych, test jest skuteczny tylko jeśli ψ nie zanika na którymś wektorze υ_k, $k = 1, ..., p$. 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 $\widehat{T}_N(p, q)$ (2.11).
4. Jeśli $\widehat{T}_N(p, q) > \chi^2_{pq}(1-\alpha)$, to odrzucamy hipotezę zerową o braku liniowej zależności. W przeciwnym razie nie mamy podstaw do odrzucenia H_0.

Przedstawiony test można stosować już do prób wielkości 40, co pokazują autorzy pozycji [Horváth, Kokoszka] w Rozdziale 9.3.

2.3 Rozkład statystyki testowej

Założenie 2.1 *Trójka $(Y_n, X_n, \varepsilon_n)$ tworzy ciąg niezależnych zmiennych funkcjonalnych o jednakowym rozkładzie, takich że ε_n jest niezależne od X_n oraz*

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

$$\mathbb{E}\|Y_n\|^2 < \infty, \quad \mathbb{E}\|X_n\|^4 < \infty \quad i \quad \mathbb{E}\|\varepsilon_n\|^4 < \infty.$$

Założenie 2.2 *Wartości własne operatorów C oraz Γ spełniają, dla pewnych $p > 0$ i $q > 0$*

$$\lambda_1 > \lambda_2 > ... > \lambda_p > \lambda_{p+1} \geqslant 0, \quad \gamma_1 > \gamma_2 > ... > \gamma_q > \gamma_{q+1} \geqslant 0.$$

Twierdzenie 2.1 *[Kokoszka et al.], [Horváth, Kokoszka]*
Jeśli spełnione są powyższe Założenia 2.1, 2.2 oraz H_0, to $\widehat{T}_N(p,q) \xrightarrow{d} \chi^2_{pq}$ przy $N \to \infty$.

Twierdzenie 2.2 *[Kokoszka et al.], [Horváth, Kokoszka]*
Przy Założeniach 2.1, 2.2 oraz jeśli $\langle \Psi(v_k), u_j \rangle \neq 0$ dla $k \leqslant p$ oraz $j \leqslant q$, to $\widehat{T}_N(p,q) \xrightarrow{P} \infty$ przy $N \to \infty$.

Dowody powyższych twierdzeń rozbijemy w krokach na kolejne lematy i wnioski. ...
Najpierw jednak zauważmy, że konsekwencją prawdziwości H_0 i przyjęcia modelu postaci $Y_n = \Psi X_n + \varepsilon_n$ jest równość $Y_n = \varepsilon_n$. ?
Będziemy korzystać z wielowymiarowej wersji Centralnego Twierdzenia Granicznego:

Twierdzenie 2.3 *[Billingsley]*
Niech $X_n = (X_{n,1}, X_{n,2}, ..., X_{n,k})$ będzie ciągiem niezależnych k-wymiarowych wektorów losowych o tym samym rozkładzie. Załóżmy, że $\mathbb{E}X^2_{n,j} < \infty$ dla każdego $j = 1, ..., k$ oraz oznaczmy wektor średnich jako $\mu = (\mu_1, ..., \mu_k) := (\mathbb{E}X_{n,1}, ..., \mathbb{E}X_{n,k})$, oraz macierz kowariancji jako $\Sigma = [\sigma_{ij}]^k_{i,j=1}$, $\sigma_{ij} = Cov(X_{n,i}, X_{n,j}) = \mathbb{E}(X_{n,i} - \mu_i)(X_{n,j} - \mu_j)$. Wówczas ciąg wektorów losowych

$$\frac{X_1 + ... + X_n - n\mu}{\sqrt{n}}$$

zbiega według rozkładu do k-wymiarowego rozkładu normalnego o wektorze średnich μ i macierzy kowariancji Σ.

Pierwszym krokiem będzie zbadanie asymptotyki składowych statystyki, gdy estymujemy tylko operator Δ, zaś wartości własne C oraz Γ przyjmujemy za znane.

Lemat 2.2 *[Kokoszka et al.], [Horváth, Kokoszka]*
Jeśli spełnione są Założenia 2.1, 2.2 i H_0, to zachodzi zbieżność według rozkładu pq-wymiarowych wektorów losowych

$$\{\sqrt{N}\langle \widehat{\Delta} v_k, u_j \rangle, \ 1 \leqslant k \leqslant p, \ 1 \leqslant j \leqslant q\} \xrightarrow{d} \{\eta_{kj}\sqrt{\lambda_k \gamma_j}, \ 1 \leqslant k \leqslant p, \ 1 \leqslant j \leqslant q\}, \quad (2.12)$$

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

Dowód. Przy H_0 mamy

$$\sqrt{N}\langle \widehat{\Delta} v_k, u_j \rangle = \sqrt{N}\Big\langle \frac{1}{N}\sum_{n=1}^{N} \langle X_n, v_k \rangle Y_n, u_j \Big\rangle = \frac{1}{\sqrt{N}}\sum_{n=1}^{N} \langle X_n, v_k \rangle \langle Y_n, u_j \rangle$$

$$= \frac{1}{\sqrt{N}}\sum_{n=1}^{N} \langle X_n, v_k \rangle \langle \varepsilon_n, u_j \rangle.$$

Skoro $(X_n)_{n\geqslant 1}$ oraz $(\varepsilon_n)_{n\geqslant 1}$ są niezależnymi po współrzędnych ciągami niezależnych L^2-elementów losowych składających się z elementów o tych samych rozkładach, to także ciągi wektorów losowych $(\{\langle X_n, \upsilon_k\rangle,\ 1 \leqslant k \leqslant p\})_{n\geqslant 1}$ oraz $(\{\langle \varepsilon_n, u_j\rangle,\ 1 \leqslant j \leqslant q\})_{n\geqslant 1}$ są niezależne po współrzędnych i składają się z niezależnych wektorów losowych o tych samych rozkładach. W takim razie $(\{\langle X_n, \upsilon_k\rangle\langle \varepsilon_n, u_j\rangle,\ 1 \leqslant k \leqslant p,\ 1 \leqslant j \leqslant q\})_{n\geqslant 1}$ także jest ciągiem niezależnych wektorów losowych o tym samym rozkładzie. Zauważmy, że powyższe zmienne losowe mają skończony drugi moment, gdyż (z nierówności Cauchy'ego-Schwarza oraz niezależności X_n i ε_n)

$$\mathbb{E}|\langle X_n, \upsilon_k\rangle\langle \varepsilon_n, u_j\rangle|^2 \leqslant \|\upsilon_k\|^2\|u_j\|^2 \cdot \mathbb{E}\|X_n\|^2\|\varepsilon_n\|^2$$

$$= \|\upsilon_k\|^2\|u_j\|^2 \cdot \mathbb{E}\|X_n\|^2\mathbb{E}\|\varepsilon_n\|^2.$$

Zachodzi

$$\mathbb{E}\langle X_n, \upsilon_k\rangle\langle \varepsilon_n, u_j\rangle = \mathbb{E}\langle X_n, \upsilon_k\rangle\mathbb{E}\langle \varepsilon_n, u_j\rangle = 0$$

oraz

$$\mathbb{E}\big(\langle X_n, \upsilon_k\rangle\langle \varepsilon_n, u_j\rangle\big)^2 = \mathbb{E}\langle X_n, \upsilon_k\rangle^2 \cdot \mathbb{E}\langle Y_n, u_j\rangle^2 = \lambda_k\gamma_j.$$

Wielowymiarowe Centralne Twierdzenie Graniczne (twierdzenie 2.3) pokazuje teraz, że wektor losowy

$$\{\sqrt{N}\langle \widehat{\Delta}\upsilon_k, u_j\rangle,\ 1 \leqslant k \leqslant p,\ 1 \leqslant j \leqslant q\} = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}\{\langle X_n, \upsilon_k\rangle\langle \varepsilon_n, u_j\rangle,\ 1 \leqslant k \leqslant p,\ 1 \leqslant j \leqslant q\}$$

zbiega do pq-wymiarowego rozkładu normalnego, w którym współrzędne mają rozkład normalny o średniej 0 i wariancji $\lambda_k\gamma_j$. Aby zakończyć dowód twierdzenia należy wykazać, że współrzędne wektora granicznego są niezależne, czyli (wobec normalności rozkładu łącznego) nieskorelowane. Wystarczy pokazać, że dla $(k, j) \neq (k', j')$, zmienne losowe $\langle X_n, \upsilon_k\rangle\langle \varepsilon_n, u_j\rangle$ i $\langle X_n, \upsilon_{k'}\rangle\langle \varepsilon_n, u_{j'}\rangle$ są nieskorelowane. Wynika to z faktu że mają one średnią zero oraz z poniższych przekształceń:

$$\mathbb{E}\langle X_n, \upsilon_k\rangle\langle \varepsilon_n, u_j\rangle\langle X_n, \upsilon_{k'}\rangle\langle \varepsilon_n, u_{j'}\rangle = \mathbb{E}\langle X_n, \upsilon_k\rangle\langle X_n, \upsilon_{k'}\rangle\mathbb{E}\langle \varepsilon_n, u_j\rangle\langle \varepsilon_n, u_{j'}\rangle =$$

$$= \mathbb{E}\langle X_n, \upsilon_k\rangle\langle X_n, \upsilon_{k'}\rangle\mathbb{E}\langle Y_n, u_j\rangle\langle Y_n, u_{j'}\rangle = \mathbb{E}\langle\langle X_n, \upsilon_k\rangle X_n, \upsilon_{k'}\rangle\mathbb{E}\langle\langle Y_n, u_j\rangle Y_n, u_{j'}\rangle$$

$$= \langle\mathbb{E}\langle X_n, \upsilon_k\rangle X_n, \upsilon_{k'}\rangle\langle\mathbb{E}\langle Y_n, u_j\rangle Y_n, u_{j'}\rangle = \langle C\upsilon_k, \upsilon_{k'}\rangle\langle\Gamma\upsilon_j, \upsilon_{j'}\rangle$$

$$= \lambda_k\langle\upsilon_k, \upsilon_{k'}\rangle\gamma_j\langle\upsilon_j, \upsilon_{j'}\rangle = \lambda_k\gamma_j\delta_{k,k'}\delta_{j,j'}.$$

□

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

Lemat 2.3 *[Kokoszka et al.], [Horváth, Kokoszka]*
Przy założeniach Twierdzenia 2.1 mamy

$$\mathbb{E}\left\|\widehat{\Delta}\right\|_{\mathcal{S}}^2 = N^{-1}\mathbb{E}\,\|X\|^2\,\mathbb{E}\,\|\varepsilon_1\|^2\,.$$

Dowód. Zauważmy, że

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

$$= N^{-2} \sum_{n,n'=1}^{N} \langle X_n, e_j\rangle\langle X_{n'}, e_j\rangle\langle Y_n, Y_{n'}\rangle.$$

Stąd, przy założeniu H_0, mamy

$$\mathbb{E}\left\|\widehat{\Delta}\right\|_{\mathcal{S}}^2 = \mathbb{E}\sum_{j=1}^{\infty}\left\|\widehat{\Delta}(e_j)\right\|^2 = \sum_{j=1}^{\infty}\mathbb{E}\left\|\widehat{\Delta}(e_j)\right\|^2 = \sum_{j=1}^{\infty}\mathbb{E}N^{-2}\sum_{n,n'=1}^{N}\langle X_n, e_j\rangle\langle X_{n'}, e_j\rangle\langle Y_n, Y_{n'}\rangle$$

$$= N^{-2}\sum_{j=1}^{\infty}\sum_{n,n'=1}^{N}\mathbb{E}[\langle X_n, e_j\rangle\langle X_{n'}, e_j\rangle\langle Y_n, Y_{n'}\rangle] = N^{-2}\sum_{j=1}^{\infty}\sum_{n,n'=1}^{N}\mathbb{E}[\langle X_n, e_j\rangle\langle X_{n'}, e_j\rangle\langle \varepsilon_n, \varepsilon_{n'}\rangle]$$

$$= N^{-2}\sum_{j=1}^{\infty}\sum_{n,n'=1}^{N}\mathbb{E}[\langle X_n, e_j\rangle\langle X_{n'}, e_j\rangle]\mathbb{E}\langle \varepsilon_n, \varepsilon_{n'}\rangle = (*).$$

Ponieważ dla $n \neq n'$ elementy losowe ε_n oraz $\varepsilon_{n'}$ są niezależne oraz $\mathbb{E}\varepsilon_n = 0$, to z Lematu 1.5 mamy $\mathbb{E}\langle \varepsilon_n, \varepsilon_{n'}\rangle = 0$, więc

$$(*) = N^{-2}\sum_{j=1}^{\infty}\sum_{n=1}^{N}\mathbb{E}[\langle X_n, e_j\rangle\langle X_n, e_j\rangle]\mathbb{E}\|\varepsilon_n\|^2 = N^{-1}\sum_{j=1}^{\infty}\mathbb{E}\langle X, e_j\rangle^2\mathbb{E}\|\varepsilon_1\|^2 =$$

$$= N^{-1}\mathbb{E}\sum_{j=1}^{\infty}\langle X, e_j\rangle^2\mathbb{E}\|\varepsilon_1\|^2 = N^{-1}\mathbb{E}\|X\|^2\mathbb{E}\|\varepsilon_1\|^2.$$

□

Lemat 2.4 *[Kokoszka et al.], [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.*

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

Wtedy zachodzi

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

Dowód. Ustalmy dowolnie $C > 0$. Należy wykazać, że $\lim_{n\to\infty} P(|\langle U_n, V_n\rangle| > C) = 0$. Weźmy dowolne $\varepsilon > 0$. Korzystając z nierówności Cauchy'ego-Schwarza mamy

$$\limsup_{n\to\infty} P(|\langle U_n, V_n\rangle| > C) \leqslant \limsup_{n\to\infty} P(\|U_n\|\|V_n\| > C)$$

$$= \limsup_{n\to\infty}\Big(P(\|U_n\|\|V_n\| > C,\ \|U_n\| > \varepsilon) + P(\|U_n\|\|V_n\| > C,\ \|U_n\| \leqslant \varepsilon)\Big)$$

$$\leqslant \limsup_{n\to\infty} P(\|U_n\| > \varepsilon) + \limsup_{n\to\infty} P(\varepsilon\|V_n\| > C,\ \|U_n\| \leqslant \varepsilon) \leqslant \limsup_{n\to\infty} P\left(\|V_n\| > \frac{C}{\varepsilon}\right).$$

Ponieważ nierówność jest prawdziwa dla dowolnego ε, to możemy przejść do granicy:

$$\limsup_{n\to\infty} P(|\langle U_n, V_n\rangle| > C) \leqslant \lim_{\varepsilon\to 0}\limsup_{n\to\infty} P\left(\|V_n\| > \frac{C}{\varepsilon}\right) = 0.$$

□

[nierówność Czebyszewa]
Pokażemy teraz, że zbieżność w Lemacie 2.2 zachodzi także, gdy wektory własne operatorów kowariancji zastąpimy ich estymatorami.

Lemat 2.5 *[Kokoszka et al.], [Horváth, Kokoszka]*
Jeśli spełnione są Założenia 2.1, 2.2 i H_0, to zachodzi zbieżność według rozkładu pq-wymiarowych wektorów losowych

$$\{\sqrt{N}\langle\widehat{\Delta}(\hat{v}_k),\hat{u}_j\rangle,\ 1\leqslant k\leqslant p,\ 1\leqslant j\leqslant q\}\xrightarrow{d}\{\eta_{kj}\sqrt{\lambda_k\gamma_j},\ 1\leqslant k\leqslant p,\ 1\leqslant j\leqslant q\},\quad(2.13)$$

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

Dowód. Na mocy Lematu 2.2, wystarczy pokazać, że dla dowolnych $1\leqslant k\leqslant p,\ 1\leqslant j\leqslant q$ zachodzi

$$\sqrt{N}\langle\widehat{\Delta}(\hat{v}_k),\hat{u}_j\rangle-\sqrt{N}\langle\widehat{\Delta}(v_k),u_j\rangle\xrightarrow{P}0,\quad(2.14)$$

gdyż zbieżność według prawdopodobieństwa jest mocniejsza od zbieżności według rozkładu. Ponieważ

$$\begin{gathered}\sqrt{N}\langle\widehat{\Delta}(\hat{v}_k),\hat{u}_j\rangle-\sqrt{N}\langle\widehat{\Delta}(v_k),u_j\rangle=\\ \sqrt{N}\langle\widehat{\Delta}(\hat{v}_k),\hat{u}_j\rangle-\sqrt{N}\langle\widehat{\Delta}(v_k),\hat{u}_j\rangle+\sqrt{N}\langle\widehat{\Delta}(v_k),\hat{u}_j\rangle-\sqrt{N}\langle\widehat{\Delta}(v_k),u_j\rangle=\\ \langle\widehat{\Delta}(v_k),\sqrt{N}(\hat{u}_j-u_j)\rangle+\sqrt{N}\langle\widehat{\Delta}(\hat{v}_k-v_k),\hat{u}_j\rangle,\end{gathered}$$

to wystarczy wykazać

$$\langle\widehat{\Delta}(v_k),\sqrt{N}(\hat{u}_j-u_j)\rangle\xrightarrow{P}0\quad(2.15)$$

i

$$\sqrt{N}\langle\widehat{\Delta}(\hat{v}_k-v_k),\hat{u}_j\rangle\xrightarrow{P}0.\quad(2.16)$$

Aby udowodnić zbieżność (2.15), zauważmy, że z Twierdzenia 1.4 oraz nierówności Czebyszewa dla dowolnego $C>0$ mamy

$$\begin{aligned}&\lim_{C\to\infty}\limsup_{N\to\infty}P(\|\sqrt{N}(\hat{u}_j-u_j)\|>C)=\lim_{C\to\infty}\limsup_{N\to\infty}P(\|(\hat{u}_j-u_j)\|^2>\frac{C^2}{N})\\ &\leqslant\lim_{C\to\infty}\limsup_{N\to\infty}\frac{N}{C^2}\mathbb{E}\|(\hat{u}_j-u_j)\|^2\leqslant\lim_{C\to\infty}\frac{1}{C^2}\limsup_{N\to\infty}N\mathbb{E}\|(\hat{u}_j-u_j)\|^2=0,\end{aligned}\quad(2.17)$$

czyli $\|\sqrt{N}(\hat{u}_j-u_j)\|=O_P(1)$. Z kolei na mocy Lematu 2.3 mamy

$$\begin{aligned}P(\|\widehat{\Delta}(v_k)\|>C)&\leqslant\frac{1}{C}\mathbb{E}\|\widehat{\Delta}(v_k)\|\leqslant\frac{1}{C}\mathbb{E}\|\widehat{\Delta}\|\|v_k\|\leqslant\frac{1}{C}\mathbb{E}(\|\widehat{\Delta}\|^2)^{1/2}\mathbb{E}(\|v_k\|^2)^{1/2}\\ &=\frac{1}{C}(\mathbb{E}\|\widehat{\Delta}\|_{\mathcal{S}}^2)^{1/2}=\frac{1}{C\sqrt{N}}(\mathbb{E}\,\|X\|^2\,\mathbb{E}\,\|\varepsilon_1\|^2\,)^{1/2}\overset{N\to\infty}{\longrightarrow}0,\end{aligned}\quad(2.18)$$

czyli $\|\widehat{\Delta}(v_k)\|\xrightarrow{P}0$. Stąd zbieżność (2.15) wynika z Lematu 2.4.
Aby wykorzystać takie samo uzasadnienie dla (2.16) (skorzystać z Twierdzenia 1.4 oraz Lematu 2.4), zauważmy, że

$$\sqrt{N}\langle\widehat{\Delta}(\hat{v}_k-v_k),\hat{u}_j\rangle=\langle\sqrt{N}(\hat{v}_k-v_k),\widehat{\Delta}^*(\hat{u}_j)\rangle.$$

Fakt, że $\|\sqrt{N}(\hat{v}_j-v_j)\|=O_P(1)$ dowodzimy identycznie jak w (2.17). Ponieważ $\|\widehat{\Delta}^*\|=\|\widehat{\Delta}\|\leqslant\|\widehat{\Delta}\|_{\mathcal{S}}$ oraz $\sup_{N\geqslant1}\mathbb{E}\|\hat{u}_j\|^2<\infty$ (gdyż $\mathbb{E}\|\hat{u}_j-u_j\|^2\to0$) to możemy powtórzyć rozumowanie z (2.18) aby otrzymać $\|\widehat{\Delta}^*(\hat{u}_j)\|\xrightarrow{P}0$. □

Z Twierdzenia 1.4 oraz nierówności Czebyszewa mamy $\hat{\lambda}_k\xrightarrow{P}\lambda_k$ oraz $\hat{\gamma}_j\xrightarrow{P}\gamma_j$, gdyż

$$P(|\hat{\lambda}_k-\lambda_k|>C)\leqslant\frac{1}{C^2}\mathbb{E}|\hat{\lambda}_k-\lambda_k|^2=\frac{1}{NC^2}N\mathbb{E}|\hat{\lambda}_k-\lambda_k|^2\overset{N\to\infty}{\longrightarrow}0.$$

Lemat 2.5 daje teraz zbieżność także wtedy, gdy wartości własne zastąpimy ich estymatorami.

Wniosek 2.1 *[Kokoszka et al.], [Horváth, Kokoszka]*
Jeśli spełnione są Założenia 2.1, 2.2 i H_0, to zachodzi zbieżność według rozkładu pq-wymiarowych wektorów losowych

$$\{\sqrt{N}\hat{\lambda_k}^{-1/2}\hat{\gamma_j}^{-1/2}\ \langle\widehat{\Delta}(\hat{v}_k),\hat{u}_j\rangle,\ 1\leqslant k\leqslant p,\ 1\leqslant j\leqslant q\}\xrightarrow{d}\{\eta_{kj},\ 1\leqslant k\leqslant p,\ 1\leqslant j\leqslant q\},\tag{2.19}$$

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

Korzystając z tego wniosku jesteśmy w stanie łatwo udowodnić Twierdzenie 2.1.

Dowód Twierdzenia 2.1. Z Wniosku 2.1 mamy

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

$$=\sum_{k=1}^{p}\sum_{j=1}^{q}\left(\sqrt{N}\widehat{\lambda}_k^{-1/2}\widehat{\gamma}_j^{-1/2}\left\langle\widehat{\Delta}(\widehat{v}_k),\widehat{u}_j\right\rangle\right)^2\xrightarrow{d}\sum_{k=1}^{p}\sum_{j=1}^{q}\eta_{kj}^2.$$

Skoro η_{kj} są niezależne oraz mają rozkład $N(0,1)$, to $\sum\limits_{k=1}^{p}\sum\limits_{j=1}^{q}\eta_{kj}^2$ ma rozkład χ^2_{pq}. □

W celu udowodnienia Twierdzenia 2.2 potrzebujemy kolejnych lematów pomocniczych, analogicznych do lematów służących do dowodu Twierdzenia 2.1. Ponieważ jednak nie zakładamy hipotezy H_0, to pewne fragmenty rozumowania muszą być zmodyfikowane.

Lemat 2.6 *Jeśli spełnione są Założenia 2.1, 2.2, to zachodzi*

$$\mathbb{E}\|\widehat{\Delta}\|\leqslant\left(\mathbb{E}\|X\|^2\mathbb{E}\|Y\|^2\right)^{1/2}.$$

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

$$\|\widehat{\Delta}u\|=\|\frac{1}{N}\sum_{n=1}^{N}\langle X_n,u\rangle Y_n\|\leqslant\frac{1}{N}\sum_{n=1}^{N}|\langle X_n,u\rangle|\|Y_n\|\leqslant\frac{1}{N}\sum_{n=1}^{N}\|X_n\|\|Y_n\|,$$

więc

$$\mathbb{E}\|\widehat{\Delta}\|\leqslant\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|X_n\|\|Y_n\|\leqslant\frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}\|X_n\|^2\mathbb{E}\|Y_n\|^2\right)^{1/2}=\left(\mathbb{E}\|X\|^2\mathbb{E}\|Y\|^2\right)^{1/2}.$$

□

Twierdzenie 2.4 *[Billingsley] Mocne Prawo Wielkich Liczb*
Niech $\{X_n\}_{n\geqslant 1}$ będzie ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie takich, że $\mathbb{E}X_n=m$. Wtedy mamy

$$\frac{1}{N}\sum_{n=1}^{N}X_n\xrightarrow{p.n.}m.$$

Lemat 2.7 *[Kokoszka et al.], [Horváth, Kokoszka]*
Jeżeli spełnione jest Założenie 2.1, to dla dowolnych funkcji $\upsilon,u\in L^2$

$$\langle\widehat{\Delta}(\upsilon),u\rangle\xrightarrow{P}\langle\Delta(\upsilon),u\rangle.$$

Dowód. Tezę otrzymujemy korzystając z Prawa Wielkich Liczb zauważając

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

oraz

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

□

Lemat 2.8 *[Kokoszka et al.], [Horváth, Kokoszka]*
Jeżeli spełnione są Założenia 2.1 oraz 2.2, to

$$\langle \widehat{\Delta}(\hat{\upsilon}_k), \hat{u}_j\rangle \xrightarrow{P} \langle \Delta(\upsilon_k), u_j\rangle, \quad \textit{dla } k \leqslant p,\ j \leqslant q.$$

Dowód. Na mocy Lematu 2.7 wystarczy pokazać

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

W tym celu pokażemy, że

$$\langle \widehat{\Delta}(\upsilon_k), \hat{u}_j - u_j\rangle = \langle \frac{1}{\sqrt{N}}\widehat{\Delta}(\upsilon_k), \sqrt{N}(\hat{u}_j - u_j)\rangle \xrightarrow{P} 0$$

i

$$\langle \widehat{\Delta}(\hat{\upsilon}_k) - \widehat{\Delta}(\upsilon_k), \hat{u}_j\rangle = \langle \sqrt{N}(\hat{\upsilon}_k - \upsilon_k), \frac{1}{\sqrt{N}}\widehat{\Delta}^*(\hat{u}_j)\rangle \xrightarrow{P} 0,$$

korzystując z Lematu 2.4. $\|\sqrt{N}(\hat{u}_j - u_j)\| = O_P(1)$ oraz $\|\sqrt{N}(\hat{\upsilon}_k - \upsilon_k)\| = O_P(1)$ pokazaliśmy już w (2.17). Z Lematu 2.6 zachodzi

$$P(\|\frac{1}{\sqrt{N}}\widehat{\Delta}(\upsilon_k)\| > C) \leqslant \frac{1}{C\sqrt{N}}\mathbb{E}\|\widehat{\Delta}(\upsilon_k)\| \leqslant \frac{1}{C\sqrt{N}}\Big(\mathbb{E}\|X\|^2\mathbb{E}\|Y\|^2\Big)^{1/2} \overset{N\to\infty}{\longrightarrow} 0,$$

czyli $\|\frac{1}{\sqrt{N}}\widehat{\Delta}(\upsilon_k)\| \xrightarrow{P} 0$. Tak samo jak w dowodzie Lematu 2.5 korzystamy z faktów, że $\|\widehat{\Delta}^*\| = \|\widehat{\Delta}\|$ oraz $\sup\limits_{N\geqslant 1}\mathbb{E}\|\hat{u}_j\|^2 < \infty$, aby uzasadnić $\|\frac{1}{\sqrt{N}}\widehat{\Delta}^*(\hat{u}_j)\| \xrightarrow{P} 0$. □

Dowód Twierdzenia 2.2. Z założenia mamy $\langle \Psi(\upsilon_k), u_j\rangle \neq 0$ dla pewnych $1 \leqslant k \leqslant p,\ 1 \leqslant j \leqslant q$. Korzystając z równości (2.9) otrzymujemy

$$\langle \Delta(\upsilon_k), u_j\rangle = \lambda_k\langle \Psi(\upsilon_k), u_j\rangle \neq 0, \text{ więc } \langle \Delta(\upsilon_k), u_j\rangle^2 > 0.$$

Wprowadźmy oznaczenia

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

$$S(p,q) = \sum_{k=1}^{p}\sum_{j=1}^{q}\lambda_k^{-1}\gamma_j^{-1}\langle \Delta(\upsilon_k), u_j\rangle^2.$$

Pokazaliśmy już (korzystając z Twierdzenia 1.4), że

$$\hat{\lambda}_k^{-1}\hat{\gamma}_j^{-1} \xrightarrow{P} \lambda_k^{-1}\gamma_j^{-1}.$$

Z Lematu 2.8 mamy

$$\langle \widehat{\Delta}(\hat{\upsilon}_k), \hat{u}_j\rangle^2 \xrightarrow{P} \langle \Delta(\upsilon_k), u_j\rangle^2,$$

więc ostatecznie

$$\widehat{S}_N(p,q) \xrightarrow{P} S(p,q) > 0.$$

Stąd

$$\widehat{T}_N(p,q) = N\widehat{S}_N(p,q) \xrightarrow{P} \infty.$$

□

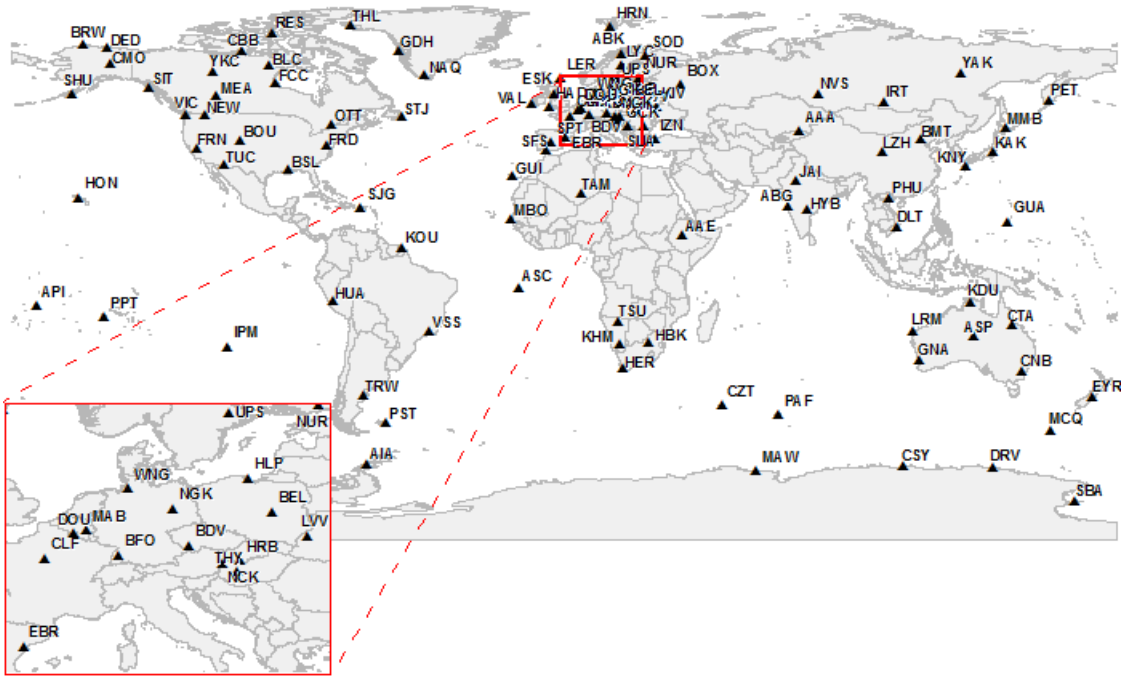
Rozdział 3

Przykład zastosowania

W tym rozdziale przedstawimy przykładowe zastosowanie testu przedstawionego w Rozdziale 2. Podobnie jak w artykule [Kokoszka et al.] oraz książce [Horváth, Kokoszka] stworzymy kilka funkcjonalnych modeli liniowych na podstawie danych opisujących natężenie pola magnetycznego Ziemi. Zmienną objaśniającą X będą obserwacje zanotowane w obserwatorium geofizycznych znajdującym się na wysokich szerokościach geograficznych. Zaś zmienna objaśniana Y będzie różna dla kolejnych modeli: zmienne Y_n będą obserwacjami z wybranego obserwatorium n znajdującego się na odpowiednio niższej szerokości geograficznej - wybrano obserwatoria ulokowane na średnich i niskich szerokościach geograficznych. Następnie każdy model zostanie przetestowany według zaprezentowanej wcześniej procedury.

3.1 Opis danych magnetometrycznych

Podobnie jak w artykule [Kokoszka et al.] oraz książce [Horváth, Kokoszka], zastosujemy przedstawiony test do modelu stworzonego na podstawie danych opisujących natężenie pola magnetycznego Ziemi. Takie dane zbierane są przez stacje geofizyczne i publikowane są w ramach międzynarodowego programu INTERMAGNET na stronie internetowej projektu [INTERMAGNET]. Do programu należy obecnie 129 naziemnych obserwatoriów, w tym dwie stacje znajdujące się w Polsce (mapa stacji na Rysunku 3.1).
Mianem **pogody kosmicznej** nazywamy charakteryzację zjawisk w przestrzeni międzyplanetarnej oddziałujących na atmosferę ziemską. Głównym źródłem jej zmian są wahania aktywności słonecznej. Słońce stale emituje naładowane cząsteczki, które docierają do Ziemi w postaci tzw. wiatrów słonecznych i mogą powodować pewne anomalie w magnetosferze i jonosferze ziemskiej.
Pogoda kosmiczna wpływa na działanie satelitów, promów kosmicznych, komunikację radiową i telefoniczną, loty samolotowe, na funkcjonowanie elektrowni, możliwe że także na klimat na Ziemi oraz na życie zwierząt oraz roślin. Zatem obserwacja i zrozumienie jej procesów, w tym subburz, jest niezwykle istotne do kontrolowania i przewidywania jej skutków.
Celem testu jest zbadanie, czy zmiany w polu magnetycznym na wysokich szerokościach geograficznych mają wpływ na pole na średnich szerokościach geograficznych, ...
Dane o polu magnetycznym, generowanym przez prąd elektryczny przepływający przez ziemską magnetosferę i jonosferę, rejestrowane są za pomocą tzw. magnetometru. To naziemne urządzenie odczytuje kilka składowych natężenia pola magnetycznego, nas interesować będzie składowa horyzontalna (H, *Horizontal*), która wskazuje na wielkość natężenia pola magnetycznego skierowanego w stronę magnetycznej północy. ...
[...]
Ze strony programu INTERMAGNET można pobrać dane dokładne: w odstępach jedno-



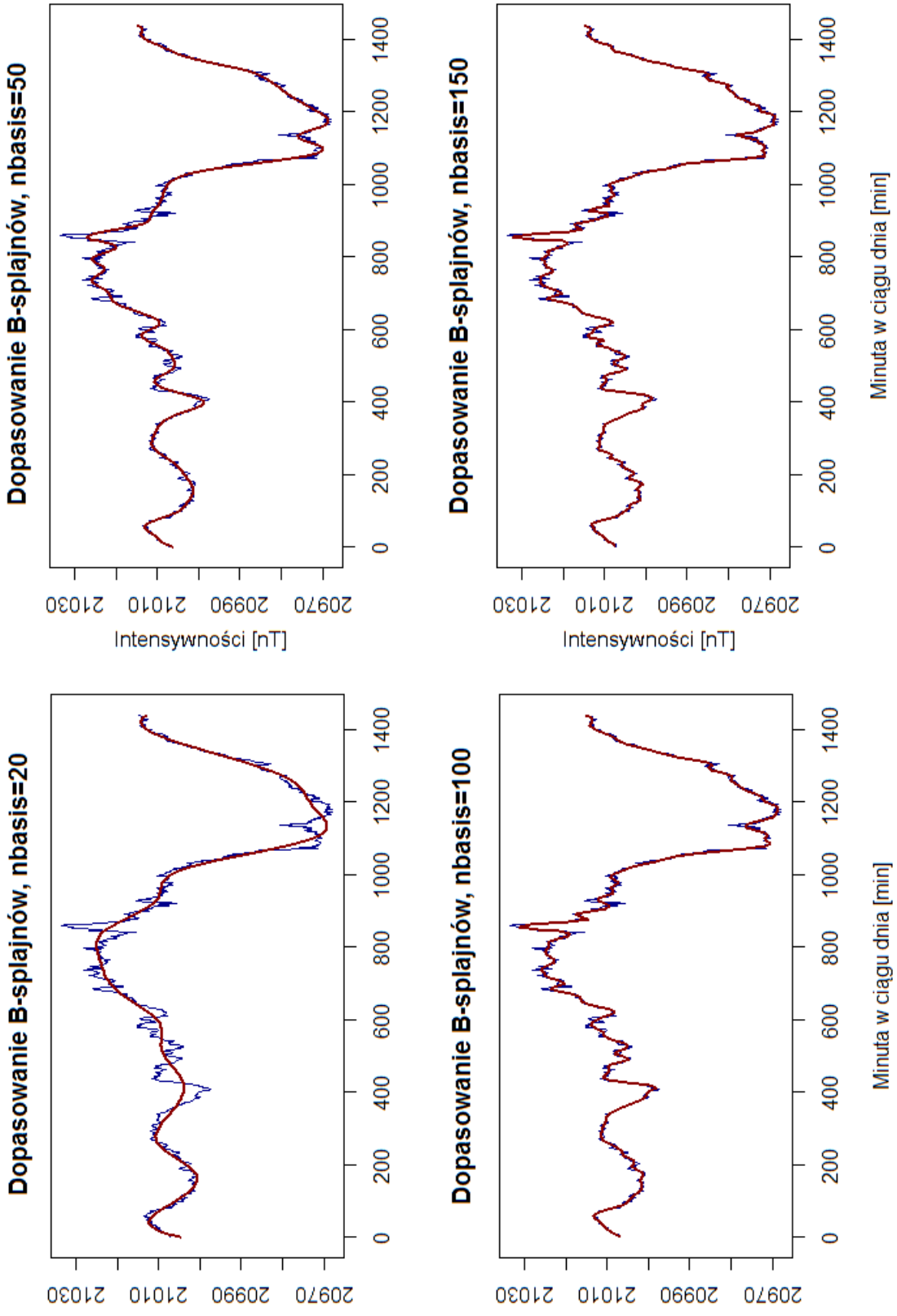
Rysunek 3.1: Mapa stacji geofizycznych należących do programu INTERMAGNET, źródło: strona internetowa projektu [INTERMAGNET]

sekundowych lub uproszczone: w odstępach jednominutowych (obserwacja jest średnią z 60 sekund). W pracy wykorzystano dane uproszczone, mamy zatem 1440 punktów każdego dnia, przypisanych według czasu centralnego, które posłużą nam do stworzenia danych funkcjonalnych. Tym sposobem jeden dzień stanie się jedną obserwacją.
Korzystając z dostępnego pakietu *fda* ([R: fda 1])...
Ze względu na częściowe braki danych w obserwacjach musieliśmy przyjąć pewne założenia odnośnie ich traktowania. W przypadku niektórych dni brakuje tylko jednej czy dwóch obserwacji, niekiedy jednak luki w zapisie danych dotyczą przynajmniej kilku godzin. Odsetek dni z brakami danych jest na tyle duży, że nie chcemy odrzucać bezwzględnie wszystkich dni z niedoborem danych. Przyjmujemy zatem następujące podejście: w przypadku braku więcej niż 10 wartości (10 minut) dzień zostanie odrzucony z analiz, jeśli jednak brakuje nie więcej niż 10 punktów w ciągu dnia obserwacje zostaną zachowane przy dopełnieniu braków danych ostatnią znaną wartością (w przypadku braku wartości początkowych bierzemy pierwszą znaną wartość).

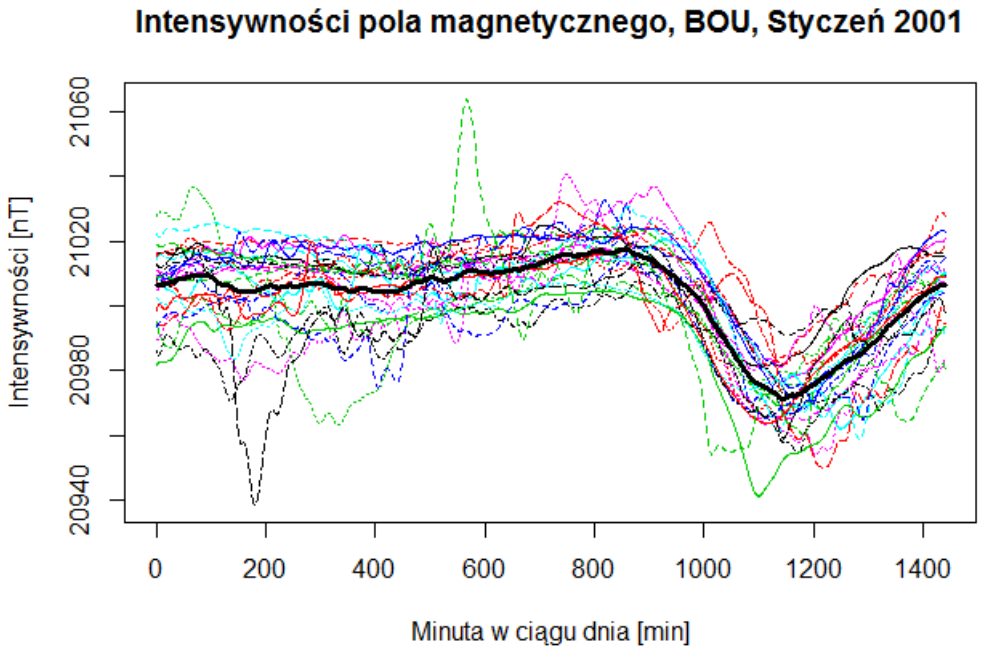
3.2 Ameryka Północna (Kanada)

W kręgu zainteresowań autorów artykułu [Kokoszka et al.] leżą dane pochodzące z obserwatoriów Ameryki Północnej, zaczniemy zatem od analizy podobnych danych.

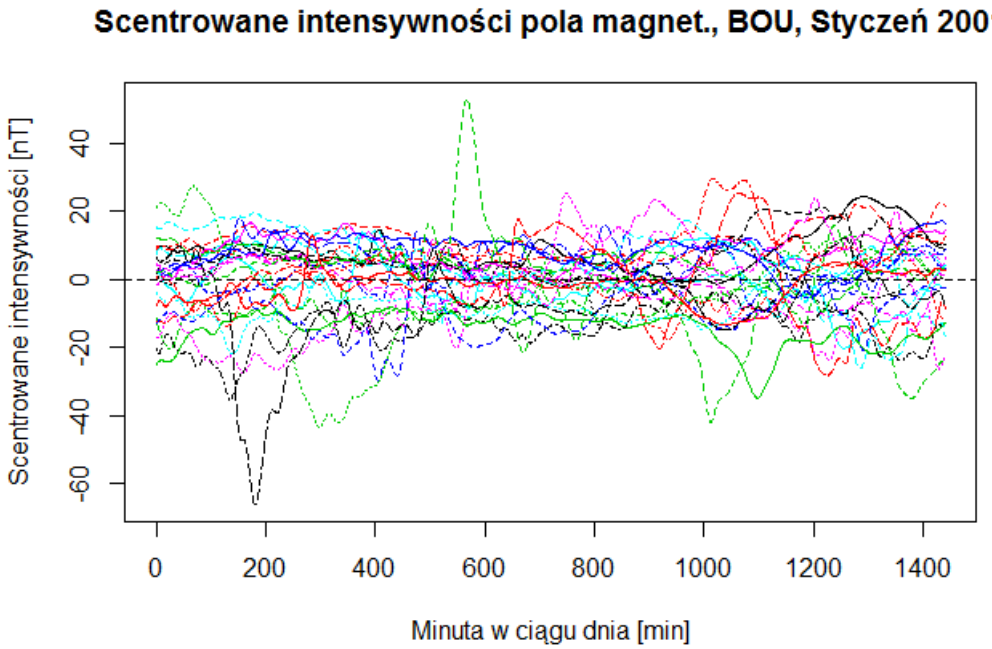
Rozważać będziemy okres od 1 stycznia do 30 czerwca 2001 roku... Wybór liczby funkcji bazowych nie ma istotnego wpływu na wynik testu, ważne żeby otrzymana dana funkcjonalna była dobrze dopasowana do oryginalnych punktów, ale z wyeliminowaniem szumu. Stopień dopasowania według wyboru liczby funkcji w bazie przedstawia Rysunek 3.2.
[...] Wymagamy, aby dane funkcjonalne były scentrowane, co można wykonać jednym poleceniem pakietu *fda* 'center.fd'. Dane przed i po scentrowaniu ilustrują Rysunki 3.3 i 3.4 [...]



Rysunek 3.2: Wykresy przedstawiają stopień dopasowania według wyboru liczby funkcji w bazie dla przykładowej danej funkcjonalnej z rozpatrywanego zbioru. czyli zmiany parametru 'nbasis' w funkcji 'create.bspline.basis' odpowiednio na 20, 50, 100 i 150. Kolor ciemno-niebieski przedstawia oryginalne dane (punkty połączono odcinkami), zaś kolor ciemno-czerwony to dopasowana krzywa.



Rysunek 3.3: Wykres przedstawia dane funkcjonalne dla stacji geofizycznej w Boulder ze stycznia 2001 (przed scentrowaniem).



Rysunek 3.4: Wykres przedstawia scentrowane dane funkcjonalne dla stacji geofizycznej w Boulder ze stycznia 2001.

3.3 Europa (Polska)

Do programu INTERMAGNET należą także dwie polskie stacje geofizyczne: obserwatorium w Belsku oraz obserwatorium na Helu. Przeprowadzimy zatem podobna j.w. analizę dla Europy. Wybraliśmy ? obserwatoriów:
Do analiz wykorzystamy najświeższe dane: od 1 stycznia do ? 2015 roku...
[...]

Dodatek A

Kod w R

Poniżej załączony jest kod napisany w języku R wykorzystany w przedstawionym wyżej przykładzie.

```
#-------------------------------------------------------------------------
# WCZYTYWANIE DANYCH BEZPOŚREDNIO Z PLIKÓW .min
#-------------------------------------------------------------------------
# BOU - STYCZEŃ
BOU.1.1<-t(matrix(as.numeric(array(scan(file="D:/.../bou20010101dmin.min",
what="list", skip=26), dim=c(7,1440))[3:4,]),nrow=2,ncol=1440))
BOU.1.2<-t(matrix(as.numeric(array(scan(file="D:/.../bou20010102dmin.min",
what="list", skip=26), dim=c(7,1440))[3:4,]),nrow=2,ncol=1440))
...
#
BOU.1<-cbind(BOU.1.1[,2],BOU.1.2[,2],...,BOU.1.30[,2],BOU.1.31[,2])
...
#
#-------------------------------------------------------------------------
# USUNIĘCIE BRAKÓW DANYCH
#-------------------------------------------------------------------------
# braki danych = 99999 lub 88888
#
# ZLICZANIE BRAKÓW DANYCH
zlicz.braki<-function(zbior){
braki<<-c()
n<-dim(zbior)
for (i in 1:n[2]){
braki<<-c(braki,length(which(zbior[,i]>80000)))
}
braki
}
zlicz.braki(BOU.1)
length(which(braki>0))

# ZAMIANA ZBIORU - USUNIĘCIE/PODMIANA BRAKÓW DANYCH
zmien.braki<-function(zbior){
n<-dim(zbior)
temp<<-zbior
braki<<-c()
for (i in 1:30){
```

```
b1<<-which(zbior[,i]>80000)
b2<<-length(b1)
braki<<-c(braki,b2)
if (b2>0 & b2<11){
if(b1[1]==1 & b2==1){
temp[1,i]<<-temp[2,i]
}else if(b1[1]==1 & b2>1 & b1[2]!=2){
temp[1,i]<<-temp[2,i]
for (j in b1[-1]) temp[j,i]<-temp[j-1,i]
}else if(b1[1]==1 & b1[2]==2){
pierwsza<<-which(b1[-b2]!=b1[-1]-1)
if (length(pierwsza)<1){ pierwsza<<-b1[b2]
temp[1:pierwsza,i]<<-temp[pierwsza+1,i]
}else{ temp[1:pierwsza[1],i]<-temp[pierwsza[1]+1,i]
for (j in b1[pierwsza[1]+1:(b2-pierwsza[1])]){ temp[j,i]<-temp[j-1,i]} }
}else if(b1[1]>1){ for (j in b1) temp[j,i]<-temp[j-1,i]
}
}
}
temp<<-temp[,-which(braki>10)]
}
#
zmien.braki(BOU.1)
BOU.1b<-usun
BOU.1bb<-zbior2
dim(BOU.1bb)
...
#
#-------------------------------------------------------------------------
# PREZENTACJA DANYCH
#-------------------------------------------------------------------------
# INSTALACJA I ZAŁADOWANIE PAKIETU 'fda'
install.packages("fda", dependencies =TRUE)
library(fda)
#
# TWORZENIE BAZY B-SPLAJNÓW DLA WYBRANEJ LICZBY FUNKCJI = PARAMETRU 'nbasis'
bspline.basis<-create.bspline.basis(c(0,1440),50)
# TWORZENIE DANEJ FUNKCJONALNEJ NA PODSTAWIE ZBIORU DANYCH = ESTYMACJA
# WSPÓŁCZYNNIKÓW W KOMBINACJI Z WYBRANĄ BAZĄ
BOU.1.fd<-Data2fd(seq(1,1440),BOU.1bb,bspline.basis)
#
# WYKRES DANYCH WEJCIOWYCH
plot(x=1:1440,y=BOU.1.4[,2],type="l",col="dark blue",
xlab="Minuta w ciągu dnia [min]",ylab="Intensywności [nT]",
main="Dopasowanie B-splajnów, nbasis=50")
# DODANIE PRZYBLIŻONEJ DANEJ FUNKCJONALNEJ
lines(BOU.1.fd[4],col="dark red",lw=2)
...
#
#-------------------------------------------------------------------------
bspline.basis<-create.bspline.basis(c(0,1440),100)
```

```
BOU.1.fd<-Data2fd(seq(1,1440),BOU.1bb,bspline.basis)
plot.fd(BOU.1.fd, xlab="Minuta w ciągu dnia [min]",ylab="Intensywności [nT]",
main="Intensywności pola magnetycznego, BOU, Styczeń 2001")
# TWORZENIE FUNKCJI ŚREDNIEJ DANYCH FUNKCJONALNYCH
mean.BOU.1.fd<-mean.fd(BOU.1.fd)
lines(mean.BOU.1.fd,lw=3)
#
# SCENTROWANIE DANYCH FUNKCJONALNYCH
BOU.1.fdc<-center.fd(BOU.1.fd)
plot.fd(BOU.1.fdc)
...
#
#------------------------------------------------------------------------

# WAŻNE FUNKCJE
#cca.fd
#cor.fd
#var.fd
#basisfd.product
#Eigen
##eval.fd
#Fperm.fd
##fRegress
##fRegress.CV
##fRegress.stderr
##Fstat.fd   #qchisq(.95,df=7)
#inprod
#inprod.bspline
#linmod
#pca.fd
#plot.pca.fd
##sd.fd

# Magnetic Local Time (MLT), Magnetic Longitude (MLON), Magnetic Latitude (MLAT)
```

Bibliografia

[Beśka] M. Beśka, *Wstęp do teorii miary (skrypt do zajęć dydaktycznych).*

[Billingsley] P. Billingsley, *Prawdopodobieństwo i miara*, Wydawnictwo Naukowe PWN 2009, s. 231, 288, 383.

[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.

[Hsing, Eubank] T. Hsing, R. Eubank, *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, Wiley 2015.

[INTERMAGNET] INTERMAGNET `http://www.intermagnet.org/index-eng.php`

[Johnson, Wichern] R.D. Johnson, D.W. Wichern, *Applied Multivatiate Statistical Analysis (6th edition)*, Pearson 2007.

[Kokoszka et al.] P. Kokoszka, I. Maslova, J. Sojka, L. Zhu, *Testing for lack of dependence in the functional linear model*, Canadian Journal of Statistics, 36 (2008), s. 207-222.

[Maslova et al.] I. Maslova, P. Kokoszk, J. Sojka and L. Zhu, *Statistical significance testing for the association of magnetometer records at high–, mid– and low latitudes during substorm days.* Planetary and Space Science, 58 (2010), s. 437–445.

[R: fda 1] J.O. Ramsay, H. Wickham, S. Graves, G. Hooker, *Package 'fda'*, wersja 2.4.4. On-line: `https://cran.r-project.org/web/packages/fda/fda.pdf`

[R: fda 2] J.O. Ramsay, G. Hooker and S. Graves, *Functional Data Analysis with R and Matlab*, Springer 2009.

[Ramsay, Silverman] J.O. Ramsay, B.W. Silverman, *Functional Data Analysis*, Springer 2005.

[SuperMAG1] IMAGE Chain: Tanskanen, E.I. (2009), A comprehensive high-throughput analysis of substorms observed by IMAGE magnetometer network: Years 1993-2003 examined, 114, A05204, doi:10.1029/2008JA013682.

[SuperMAG2] MACCS: Engebretson, M. J., W. J. Hughes, J. L. Alford, E. Zesta, L. J. Cahill, Jr., R. L. Arnoldy, and G. D. Reeves (1995), Magnetometer array for cusp and cleft studies observations of the spatial extent of broadband ULF magnetic pulsations at cusp/cleft latitudes , J. Geophys. Res., 100, 19371-19386, doi:10.1029/95JA00768.

[SuperMAG3] MAGDAS / 210 Chain: Yumoto, K,. and the CPMN Group (2001), Characteristics of Pi 2 magnetic pulsations observed at the CPMN stations: A review of the STEP results, Earth Planets Space, 53, s. 981-992.

[SuperMAG4] SuperMAG: Gjerloev, J. W. (2012), The SuperMAG data processing technique, J. Geophys. Res., 117 , A09213, doi:10.1029/2012JA017683.

[SuperMAG5] McMAC Chain: Chi, P. J., M. J. Engebretson, M. B. Moldwin, C. T. Russell, I. R. Mann, M. R. Hairston, M. Reno, J. Goldstein, L. I. Winkler, J. L. Cruz-Abeyro, D.-H. Lee, K.Yumoto, R. Dalrymple, B. Chen, and J. P. Gibson (2013), Sounding of the plasmasphere by Mid-continent MAgnetoseismic Chain magnetometers, J. Geophys. Res. Space Physics, 118, doi:10.1002/jgra.50274.

[SuperMAG6] EMMA: Lichtenberger J., M. Clilverd, B. Heilig, M. Vellante, J. Manninen, C. Rodger, A. Collier, A. Jørgensen, J. Reda, R. Holzworth, and R. Friedel (2013), The plasmasphere during a space weather event: first results from the PLASMON project, J. Space Weather Space Clim., 3, A23 (`www.swsc-journal.org/articles/swsc/pdf/2013/01/swsc120062.pdf`).