Miary zależności między dwoma wielowymiarowymi procesami stochastycznymi

Mirosław Krzyśko

mkrzysko@amu.edu.pl

Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Plan prezentacji

Wstęp

W ostatnich latach wiele uwagi poświęcono danym reprezentującym funkcje lub krzywe. Dane tego typu znane są w literaturze pod nazwą danych funkcjonalnych (Ramsay & Silverman, 2005). Przykłady danych funkcjonalnych można znaleźć w wielu dziedzinach takich, jak medycyna, ekonomia, meteorologia itd. W wielu zastosowaniach używamy metod statystycznych dla obiektów scharakteryzowanych wieloma cechami obserwowanymi w wielu momentach czasowych (dane podwójnie wielowymiarowe). Dane tego typu nazywamy wielowymiarowymi danymi funkcjonalnymi. Autorem pionierskiej pracy teoretycznej z tego zakresu był Besse (1979). W jego pracy zmienne losowe przyjmowały wartości w przestrzeni Hilberta. Saporta (1981) rozważał metody analizy czynnikowej (analiza składowych głównych i korelacji kanonicznych) w zastosowaniu do wielowymiarowych danych funkcjonalnych.

Wygładzanie

Załóżmy, że $\boldsymbol{X} \in L_2^p(I_1)$ i $\boldsymbol{Y} \in L_2^q(I_2)$ są procesami losowymi, gdzie $L_2(I)$ jest przestrzenią Hilberta funkcji całkowalnych z kwadratem na przedziale I.

Ponadto załóżmy, że

$$\mathrm{E}(\boldsymbol{X}(s)) = \boldsymbol{0}, \;\; s \in I_1,$$

$$\mathrm{E}(\boldsymbol{Y}(t)) = \boldsymbol{0}, \;\; t \in I_2.$$

Założenie to nie powoduje utraty ogólności rozważań, ponieważ funkcjonalne współczynniki korelacji są wyznaczane jedynie na podstawie macierzy kowariancji procesów $\boldsymbol{X}$ i $\boldsymbol{Y}$ postaci

$$\boldsymbol{\Sigma}(s,t) = \begin{bmatrix} \boldsymbol{\Sigma}_{XX}(s,t) & \boldsymbol{\Sigma}_{XY}(s,t) \\ \boldsymbol{\Sigma}_{YX}(s,t) & \boldsymbol{\Sigma}_{YY}(s,t) \end{bmatrix}, \;\; s \in I_1, \;\; t \in I_2.$$

Wygładzanie

W dalszym ciągu zakładamy, że każda składowa X_g procesu $\boldsymbol{X}$ i każda składowa Y_h procesu $\boldsymbol{Y}$ może być reprezentowana przez skończoną liczbę funkcji bazowych $\{\varphi_e\}$ i $\{\varphi_f\}$:

$$X_g(s) = \sum_{e=0}^{E_g} \alpha_{ge}\varphi_e(s), s \in I_1, g = 1, 2, ..., p,$$

$$Y_h(t) = \sum_{f=0}^{F_h} \beta_{hf}\varphi_f(t), t \in I_2, h = 1, 2, ..., q.$$

Stopień gładkości funkcji X_g i Y_h zależy od wartości E_g i F_h (małe wartości dają mniejsze wygładzenie funkcji).

Wygładzanie

Wprowadźmy następujące oznaczenia:

$$\boldsymbol{\alpha} = (\alpha_{10}, ..., \alpha_{1E_1}, ..., \alpha_{p0}, ..., \alpha_{pE_p})',$$

$$\boldsymbol{\beta} = (\beta_{10}, ..., \beta_{1F_1}, ..., \beta_{q0}, ..., \beta_{qF_q})',$$

$$\boldsymbol{\Phi}_1(s) = \begin{bmatrix} \varphi'_{E_1}(s) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \varphi'_{E_2}(s) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \varphi'_{E_p}(s) \end{bmatrix},$$

$$\boldsymbol{\Phi}_2(t) = \begin{bmatrix} \varphi'_{F_1}(t) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \varphi'_{F_2}(t) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \varphi'_{F_q}(t) \end{bmatrix},$$

gdzie $\varphi_{E_1}, ..., \varphi_{E_p}$ i $\varphi_{F_1}, ..., \varphi_{F_q}$ są ortonormalnymi funkcjami bazowymi przestrzeni $L_2(I_1)$ i $L_2(I_2)$ oraz $K_1 = E_1 + ... + E_p$, $K_2 = F_1 + ... + F_q$.

Wygładzanie

Używając notacji macierzowej procesy $\boldsymbol{X}(s)$ i $\boldsymbol{Y}(t)$ mają następującą reprezentację

$$\boldsymbol{X}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\alpha}, \quad \boldsymbol{Y}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\beta}.$$

Oznacza to, że realizacje procesów $\boldsymbol{X}$ i $\boldsymbol{Y}$ zawarte są w skończenie wymiarowych podprzestrzeniach przestrzeni $L_2^p(I_1)$ i $L_2^q(I_2)$ odpowiednio. Te podprzestrzenie będziemy oznaczali przez $\mathcal{L}_2^p(I_1)$ i $\mathcal{L}_2^q(I_2)$.

Ponadto, dla wektorów $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$ mamy:

$$\mathrm{E}(\boldsymbol{\alpha}) = \mathbf{0}, \ \mathrm{E}(\boldsymbol{\beta}) = \mathbf{0}$$

oraz

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\alpha\alpha} & \boldsymbol{\Sigma}_{\alpha\beta} \\ \boldsymbol{\Sigma}_{\beta\alpha} & \boldsymbol{\Sigma}_{\beta\beta} \end{bmatrix}.$$

Wygładzanie

Zazwyczaj proces jest obserwowany w skończonej liczbie momentów czasowych. Proces transformacji danych dyskretnych do danych funkcjonalnych jest wykonywany oddzielnie dla każdej realizacji każdej składowej procesu. Niech x_{gj} oznacza obserwowaną wartość składowej X_g, $g = 1, 2, \ldots p$ w j-tym momencie czasowym s_j, gdzie $j = 1, 2, ..., J$. Podobnie, niech y_{hj} oznacza obserwowaną wartość składowej Y_h, $h = 1, 2, \ldots q$ w j-tym momencie czasowym t_j, gdzie $j = 1, 2, ..., J$. Wtedy nasze dane składają się z pJ par (s_j, x_{gj}) oraz z qJ par (t_j, y_{hj}). Współczynniki $\boldsymbol{\alpha}_i$ i $\boldsymbol{\beta}_i$ są estymowane metodą najmniejszych kwadratów. Oznaczmy $\hat{\boldsymbol{\alpha}}_i$ i $\hat{\boldsymbol{\beta}}_i$ przez $\boldsymbol{a}_i$ i $\boldsymbol{b}_i$, $i = 1, 2, \ldots, n$.

Wygładzanie

W rezultacie procesu transformacji otrzymujemy dane funkcjonalne postaci:

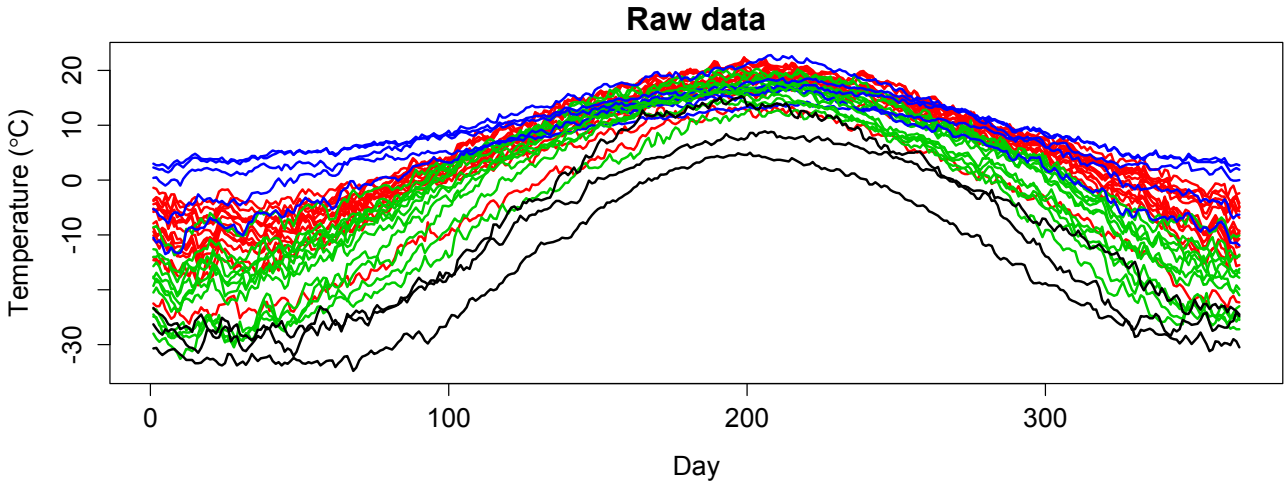
$$\boldsymbol{x}_i(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{a}_i,\ \boldsymbol{y}_i(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{b}_i,$$

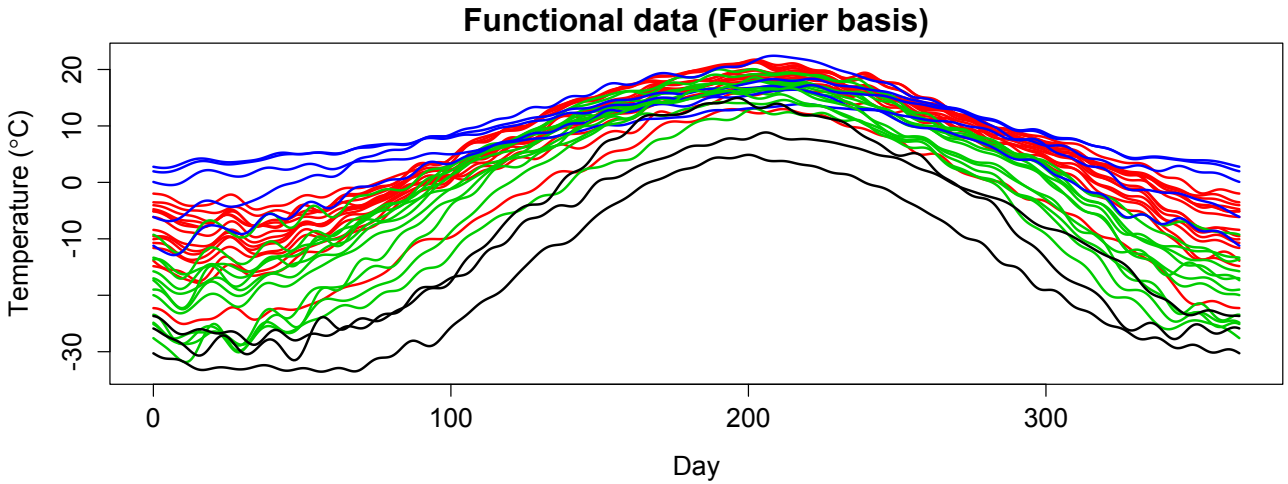
gdzie $s \in I_1,\ t \in I_2,\ i = 1, 2, \ldots, n$. Niech $\boldsymbol{A} = (\boldsymbol{a}_1', \boldsymbol{a}_2', \ldots, \boldsymbol{a}_n')$ oraz $\boldsymbol{B} = (\boldsymbol{b}_1', \boldsymbol{b}_2', \ldots, \boldsymbol{b}_n')$. Wtedy

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},\boldsymbol{\alpha}} = \frac{1}{n}\boldsymbol{A}'\boldsymbol{A}, \quad \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},\boldsymbol{\beta}} = \frac{1}{n}\boldsymbol{B}'\boldsymbol{B}, \quad \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \frac{1}{n}\boldsymbol{A}'\boldsymbol{B}.$$

Wygładzanie

Średnie dzienne temperatury dla 35 stacji meteorologicznych w Kanadzie w latach 1960 – 1994. Dane pochodzą z książki Ramsay'a i Silvermana (2005), Functional Data Analysis, Springer, New York.





Scałkowany współczynnik ρV

Kwadrat scałkowanego współczynnika ρV (Robert i Escoufier (1976)) definiujemy jako

$$(I\rho V)^2 = \frac{\|\mathbf{\Sigma}_{XY}\|^2}{\|\mathbf{\Sigma}_{XX}\|\|\mathbf{\Sigma}_{YY}\|},$$

gdzie

$$\|\mathbf{\Sigma}_{XY}\| = \sqrt{\int_{I_1}\int_{I_2} \mathrm{tr}(\mathbf{\Sigma}'_{XY}(s,t)\mathbf{\Sigma}_{XY}(s,t))dsdt}.$$

Własności:

- $0 \leq I\rho V \leq 1$.
- $I\rho V = 0$ wtedy i tylko wtedy, gdy dla każdych $g = 1, 2, ..., p$ i $h = 1, 2, ..., q$ takich, że $g \neq h$, składowa X_g procesu $\boldsymbol{X}$ i składowa Y_h procesu $\boldsymbol{Y}$ są nieskorelowane.

Scałkowany współczynnik ρV

Ponieważ

$$\mathbf{\Sigma}_{XY}(s,t) = \mathbf{\Phi}_1(s)\mathbf{\Sigma}_{\alpha\beta}\mathbf{\Phi}_2'(t)$$

to

$$\|\mathbf{\Sigma}_{XY}\|^2 = \mathrm{tr}(\mathbf{\Sigma}'_{\alpha\beta}\mathbf{\Sigma}_{\alpha\beta}).$$

Analogicznie

$$\|\mathbf{\Sigma}_{XX}\|^2 = \mathrm{tr}(\mathbf{\Sigma}'_{\alpha\alpha}\mathbf{\Sigma}_{\alpha\alpha}),$$
$$\|\mathbf{\Sigma}_{YY}\|^2 = \mathrm{tr}(\mathbf{\Sigma}'_{\beta\beta}\mathbf{\Sigma}_{\beta\beta}).$$

Stąd

$$(I\rho V)^2 = \frac{\mathrm{tr}(\mathbf{\Sigma}'_{\alpha\beta}\mathbf{\Sigma}_{\alpha\beta})}{\sqrt{\mathrm{tr}(\mathbf{\Sigma}'_{\alpha\alpha}\mathbf{\Sigma}_{\alpha\alpha})\,\mathrm{tr}(\mathbf{\Sigma}'_{\beta\beta}\mathbf{\Sigma}_{\beta\beta})}}.$$

Scałkowany współczynnik ρV

Twierdzenie

Scałkowany współczynnik korelacji $I\rho V$ pary procesów losowych $\boldsymbol{X} \in \mathcal{L}_2^p(I_1)$ i $\boldsymbol{Y} \in \mathcal{L}_2^q(I_2)$ jest równoważny współczynnikowi korelacji ρV pary wektorów losowych $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$.

Współczynnik ρV możemy estymować na podstawie próby. Niech

$$\boldsymbol{W}_{\alpha\alpha} = \boldsymbol{A}\boldsymbol{A}', \ \boldsymbol{W}_{\beta\beta} = \boldsymbol{B}\boldsymbol{B}'.$$

Wtedy

$$(\widehat{\rho V})^2 = (rV)^2 = \frac{\mathrm{tr}(\boldsymbol{W}_{\alpha\alpha}\boldsymbol{W}_{\beta\beta})}{\sqrt{\mathrm{tr}(\boldsymbol{W}_{\alpha\alpha}^2)\,\mathrm{tr}(\boldsymbol{W}_{\beta\beta}^2)}}.$$

Testowanie istotności współczynnika $I\rho V$

Zauważmy, że problem testowania hipotezy zerowej

$$H_0 : I\rho V = 0$$

przeciwko

$$H_1 : I\rho V \neq 0$$

dla pary procesów losowych $\boldsymbol{X} \in \mathcal{L}_2^p(I_1)$ i $\boldsymbol{Y} \in \mathcal{L}_2^q(I_2)$ jest równoważny z problemem testowania hipotezy zerowej

$$H_0 : \rho V = 0$$

przeciwko

$$H_1 : \rho V \neq 0$$

dla pary wektorów losowych $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Testowanie istotności współczynnika $l\rho V$

Przy prawdziwości hipotezy zerowej oraz przy założeniu, że łączny rozkład wektorów losowych należy do klasy rozkładów eliptycznych, graniczny rozkład współczynnika nrV jest równy rozkładowi zmiennej losowej

$$\frac{1+k}{\operatorname{tr}(\mathbf{\Sigma}_{\alpha\alpha}^2)\operatorname{tr}(\mathbf{\Sigma}_{\beta\beta}^2)} \sum_{l=1}^{K_1+p} \sum_{m=1}^{K_2+q} \lambda_l \gamma_m Z_{lm}^2,$$

gdzie k jest kurtozą rozkładu eliptycznego, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{K_1+p}$ są wartościami własnymi macierzy kowariancji $\mathbf{\Sigma}_{\alpha\alpha}$, $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_{K_2+q}$ są wartościami własnymi macierzy kowariancji $\mathbf{\Sigma}_{\beta\beta}$, Z_{lm} są niezależnymi zmiennymi losowymi o rozkładach $N(0,1)$.

Ponieważ testy asymptotyczne w wielu sytuacjach zawodzą, sensowne wydaje się użycie testu permutacyjnego (Josse i inni (2008)).

Korelacja odległościowa

Funkcja charakterystyczna łącznego rozkładu procesów losowych $\boldsymbol{X}$ i $\boldsymbol{Y}$ ma postać:

$$\varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = \mathrm{E}\{\exp[i < \boldsymbol{l},\boldsymbol{X} >_p + i < \boldsymbol{m},\boldsymbol{Y} >_q]\},$$

gdzie

$$< \boldsymbol{l},\boldsymbol{X} >_p = \int_{I_1} \boldsymbol{l}'(s)\boldsymbol{X}(s)ds, \quad < \boldsymbol{m},\boldsymbol{Y} >_q = \int_{I_2} \boldsymbol{m}'(t)\boldsymbol{Y}(t)dt.$$

Funkcje charakterystyczne rozkładów brzegowych procesów $\boldsymbol{X}$ i $\boldsymbol{Y}$ są postaci:

$$\varphi_{\boldsymbol{X}}(\boldsymbol{l}) = \varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{0}), \quad \varphi_{\boldsymbol{Y}}(\boldsymbol{m}) = \varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{0},\boldsymbol{m}).$$

Korelacja odległościowa

Załóżmy, że $\boldsymbol{X} \in \mathcal{L}_2^p(I_1)$ oraz $\boldsymbol{Y} \in \mathcal{L}_2^q(I_2)$. Wówczas procesy te mogą mieć reprezentację:

$$\boldsymbol{X}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\alpha}, \quad \boldsymbol{Y}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\beta},$$

gdzie $\boldsymbol{\alpha} \in \mathbb{R}^{K_1+p}$, $\boldsymbol{\beta} \in \mathbb{R}^{K_2+q}$, natomiast funkcje wektorowe $\boldsymbol{l}$ i $\boldsymbol{m}$ mogą być reprezentowane jako

$$\boldsymbol{l}(s) = \boldsymbol{\Phi}_1(s)\boldsymbol{\lambda}, \quad \boldsymbol{m}(t) = \boldsymbol{\Phi}_2(t)\boldsymbol{\mu},$$

gdzie $\boldsymbol{\lambda} \in \mathbb{R}^{K_1+p}$, $\boldsymbol{\mu} \in \mathbb{R}^{K_2+q}$. Stąd

$$< \boldsymbol{l}, \boldsymbol{X} >_p = \boldsymbol{\lambda}'\boldsymbol{\alpha}, \quad < \boldsymbol{m}, \boldsymbol{Y} >_q = \boldsymbol{\mu}'\boldsymbol{\beta},$$

oraz

$$\varphi_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l}, \boldsymbol{m}) = \mathrm{E}\{\exp[i\boldsymbol{\lambda}\boldsymbol{\alpha} + i\boldsymbol{\mu}\boldsymbol{\beta}]\} = \varphi_{\boldsymbol{\alpha},\boldsymbol{\beta}},$$

gdzie $\varphi_{\boldsymbol{\alpha},\boldsymbol{\beta}(\boldsymbol{\lambda},\boldsymbol{\mu})}$ jest łączną funkcją charakterystyczną wektorów losowych $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$.

Korelacja odległościowa

Bazując na idei kowariancji odległościowej pomiędzy dwoma wektorami losowymi (Székely i inni (2007)), możemy zdefiniować kowariancję odległościową pomiędzy procesami losowymi $\boldsymbol{X}$ i $\boldsymbol{Y}$ jako

$$\nu_{\boldsymbol{X},\boldsymbol{Y}} = \nu_{\boldsymbol{\alpha},\boldsymbol{\beta}},$$

gdzie

$$\nu^2_{\boldsymbol{\alpha},\boldsymbol{\beta}} = \frac{1}{C_{K_1+p} C_{K_2+q}} \int_{\mathbb{R}^{K_1+K_2+p+q}} \frac{|\varphi_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{\lambda},\boldsymbol{\mu}) - \varphi_{\boldsymbol{\alpha}}(\boldsymbol{\lambda})\varphi_{\boldsymbol{\beta}}(\boldsymbol{\mu})|^2}{\|\boldsymbol{\lambda}\|^{K_1+p+1}_{K_1+p} \|\boldsymbol{\mu}\|^{K_2+q+1}_{K_2+q}} d\boldsymbol{\lambda} d\boldsymbol{\mu},$$

gdzie

$$C_r = \frac{\pi^{\frac{1}{2}(r+1)}}{\Gamma(\frac{1}{2}(r+1))}.$$

Korelacja odległościowa

Funkcjonalna korelacja odległościowa pomiędzy procesami losowymi $\boldsymbol{X}$ i $\boldsymbol{Y}$ jest liczbą nieujemną zdefiniowaną jako

$$\mathcal{R}_{\boldsymbol{X},\boldsymbol{Y}} = \mathcal{R}_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\nu_{\boldsymbol{X},\boldsymbol{Y}}}{\sqrt{\nu_{\boldsymbol{X},\boldsymbol{X}}\nu_{\boldsymbol{Y},\boldsymbol{Y}}}}$$

jeżeli $\nu_{\boldsymbol{X},\boldsymbol{X}}$ i $\nu_{\boldsymbol{Y},\boldsymbol{Y}}$ są dodatnie i zero w przeciwnym przypadku.
Dla rozkładów ze skończonymi pierwszymi momentami mamy

$$0 \leq \mathcal{R}_{\boldsymbol{X},\boldsymbol{Y}} \leq 1$$

i $\mathcal{R}_{\boldsymbol{X},\boldsymbol{Y}} = 0$ wtedy i tylko wtedy gdy $\boldsymbol{X}$ i $\boldsymbol{Y}$ są niezależne. $\mathcal{R}_{\boldsymbol{X},\boldsymbol{Y}}$ jest estymatorem zgodnym.

Korelacja odległościowa

Musimy teraz oszacować $\mathcal{R}_{\boldsymbol{X},\boldsymbol{Y}}$ na podstawie $(\boldsymbol{a}_k, \boldsymbol{b}_k)$, $k = 1, 2, \ldots, n$. Niech

$$\bar{\boldsymbol{a}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i, \qquad \bar{\boldsymbol{b}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{b}_i,$$

$$\tilde{\boldsymbol{a}}_k = \boldsymbol{a}_k - \bar{\boldsymbol{a}}, \qquad \tilde{\boldsymbol{b}}_k = \boldsymbol{b}_k - \bar{\boldsymbol{b}}, \qquad k = 1, 2, \ldots, n$$

oraz

$$\boldsymbol{A} = (a_{kl}), \qquad \boldsymbol{B} = (b_{kl}),$$
$$\tilde{\boldsymbol{A}} = (A_{kl}), \qquad \tilde{\boldsymbol{B}} = (B_{kl}),$$

gdzie

$$a_{kl} = \|\boldsymbol{a}_k - \boldsymbol{a}_l\|_{K_1+p}, \qquad b_{kl} = \|\boldsymbol{b}_k - \boldsymbol{b}_l\|_{K_2+q},$$
$$A_{kl} = \|\tilde{\boldsymbol{a}}_k - \tilde{\boldsymbol{a}}_l\|_{K_1+p}, \qquad B_{kl} = \|\tilde{\boldsymbol{b}}_k - \tilde{\boldsymbol{b}}_l\|_{K_2+q}, \qquad k, l = 1, 2, \ldots, n.$$

Korelacja odległościowa

Zachodzą związki

$$\tilde{\boldsymbol{A}} = \boldsymbol{HAH} \text{ oraz } \tilde{\boldsymbol{B}} = \boldsymbol{HBH},$$

gdzie

$$\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$$

jest macierzą centrującą.

Niech $\tilde{\boldsymbol{A}} \circ \tilde{\boldsymbol{B}} = (A_{kl}B_{kl})$ będzie iloczynem Hadamarda macierzy $\tilde{\boldsymbol{A}}$ oraz $\tilde{\boldsymbol{B}}$.

Korelacja odległościowa

Na podstawie wyników Székely'ego i innych (2007), mamy

$$\hat{\nu}^2_{\boldsymbol{X},\boldsymbol{Y}} = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

Funkcjonalna korelacja odległościowa z próby jest zatem zdefiniowana jako

$$\hat{\mathcal{R}}_{\boldsymbol{X},\boldsymbol{Y}} = \hat{\mathcal{R}}_{\boldsymbol{a},\boldsymbol{b}} = \frac{\hat{\nu}_{\boldsymbol{a},\boldsymbol{b}}}{\sqrt{\hat{\nu}_{\boldsymbol{a},\boldsymbol{a}}\hat{\nu}_{\boldsymbol{b},\boldsymbol{b}}}}$$

jeżeli $\hat{\nu}_{\boldsymbol{a},\boldsymbol{a}}$ oraz $\hat{\nu}_{\boldsymbol{b},\boldsymbol{b}}$ są dodatnie i zero w przeciwnym przypadku.

Testowanie istotności korelacji odległościowej

Problem testowania niezależności pomiędzy procesami losowymi $\boldsymbol{X} \in \mathcal{L}_2^p(I_1)$ oraz $\boldsymbol{Y} \in \mathcal{L}_2^q(I_2)$ jest równoważny problemowi testowania $H_0 : \mathcal{R}_{\boldsymbol{X},\boldsymbol{Y}} = 0$. Székely i inni (2007) pokazali, że przy prawdziwości H_0, $n\hat{\mathcal{R}}_{\boldsymbol{X},\boldsymbol{Y}}$ zbiega do

$$\sum_{j=1}^{\infty} \nu_j Z_j^2,$$

gdzie Z_j są niezależnymi zmiennymi losowymi o rozkładzie $N(0,1)$, oraz ν_j zależy od rozkładu $(\boldsymbol{X}, \boldsymbol{Y})$. W praktyce używamy jednak testów permutacyjnych w celu zbadania istotności korelacji odegłościowej (Josse & Holmes (2016)).

Jądrowy współczynnik zgodności wektorów losowych

Niech $\boldsymbol{X} \in \mathbb{R}^p$ i $\boldsymbol{Y} \in \mathbb{R}^q$ będą wektorami losowymi. Oznaczmy przez $P_{\boldsymbol{X}}$, $P_{\boldsymbol{Y}}$ i $P_{\boldsymbol{X},\boldsymbol{Y}}$ miary probabilistyczne wektorów $\boldsymbol{X}$, $\boldsymbol{Y}$ i $(\boldsymbol{X}, \boldsymbol{Y})$ odpowiednio określonych na przestrzeniach $\mathbb{R}^p$, $\mathbb{R}^q$ i $\mathbb{R}^p \times \mathbb{R}^q$. Niech $S_{\boldsymbol{x}} = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_n\}$ będzie próbą z rozkładu $P_{\boldsymbol{X}}$. Niech

$$k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$$

będzie rzeczywistą funkcją ciągłą, zwaną jądrem i niech $\boldsymbol{K} = (k_{ij})$, gdzie $k_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ będzie macierzą jądrową, $i, j = 1, \ldots, n$. Dalej będziemy przyjmować, że k jest jądrem gaussowskim, tj.

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\lambda ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2),$$

$\lambda > 0$.

Jądrowy współczynnik zgodności wektorów losowych

Dla każdej macierzy jądrowej $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, przez jądrową macierz scentrowaną rozumiemy macierz postaci

$$\tilde{\boldsymbol{K}} = \left(\boldsymbol{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \boldsymbol{K} \left(\boldsymbol{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right),$$

gdzie $\mathbf{1}_n \in \mathbb{R}^{n \times 1}$ oznacza wektor złożony z samych jedynek. Macierz $\tilde{\boldsymbol{K}}$ jest macierzą nieujemnie określoną.

Jądrowy współczynnik zgodności wektorów losowych

Niech

$$< \boldsymbol{A}, \boldsymbol{B} >_F = \mathrm{tr}(\boldsymbol{A}^T \boldsymbol{B})$$

$$||\boldsymbol{A}||_F = (< \boldsymbol{A}, \boldsymbol{A} >_F)^{1/2}$$

będą iloczynem i normą Frobeniusa macierzy $\boldsymbol{A},\ \boldsymbol{B} \in \mathbb{R}^{n \times n}$. Dla każdych macierzy jądrowych opartych na próbach $S_{\boldsymbol{x}} = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_n\}$ oraz $S_{\boldsymbol{y}} = \{\boldsymbol{y}_1, \ldots \boldsymbol{y}_n\}$ zachodzą równości:

$$< \tilde{\boldsymbol{K}}_{\boldsymbol{x}}, \tilde{\boldsymbol{K}}_{\boldsymbol{y}} >_F = < \boldsymbol{K}_{\boldsymbol{x}}, \tilde{\boldsymbol{K}}_{\boldsymbol{y}} >_F = < \tilde{\boldsymbol{K}}_{\boldsymbol{x}}, \boldsymbol{K}_{\boldsymbol{y}} >_F .$$

Jądrowy współczynnik zgodności wektorów losowych

Gretton i inni (2005) zdefiniowali współczynnik HSIC (ang. Hilbert-Schmidt Independence Criterion) postaci

$$\mathrm{HSIC}(P_{\boldsymbol{X},\boldsymbol{Y}}) = ||\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}||^2_{HS},$$

gdzie

$$\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}} = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[< \varphi(\boldsymbol{X}) - \mathrm{E}_{\boldsymbol{X}}[\varphi(\boldsymbol{X})], \psi(\boldsymbol{Y}) - \mathrm{E}_{\boldsymbol{Y}}[\psi(\boldsymbol{Y})] >_{\mathcal{H}}]$$

jest operatorem kowariancji między wektorami $\boldsymbol{X}$ i $\boldsymbol{Y}$, będącym uogólnieniem macierzy kowariancji dla wektorów losowych oraz $||\boldsymbol{A}||_{HS} = \mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A})$.

W przestrzeniach rzeczywistych norma Hilberta-Schmidta nazywana jest normą Frobeniusa.

Jądrowy współczynnik zgodności wektorów losowych

Dla jądra gaussowskiego $||\boldsymbol{C}_{\boldsymbol{X},\boldsymbol{Y}}||_{HS} = 0$ wtedy i tylko wtedy, gdy $\boldsymbol{X}$ i $\boldsymbol{Y}$ są niezależne.

Estymator z próby $S_{\boldsymbol{x},\boldsymbol{y}}$ pobranej z rozkładu $P_{\boldsymbol{X},\boldsymbol{Y}}$ współczynnika HSIC ma postać

$$\mathrm{HSIC}(S_{\boldsymbol{x},\boldsymbol{y}}) = \frac{1}{n^2} < \tilde{\boldsymbol{K}}_{\boldsymbol{x}}, \tilde{\boldsymbol{K}}_{\boldsymbol{y}} >_F .$$

Jądrowy współczynnik zgodności wektorów losowych

Cortes i inni (2012) zdefiniowali współczynnik zgodności między macierzami jądrowymi (KTA – ang. Kernel Target Alignment) $\boldsymbol{K_x}$ i $\boldsymbol{K_y}$ postaci

$$\rho(\boldsymbol{K_x}, \boldsymbol{K_y}) = \frac{< \tilde{\boldsymbol{K}}_{\boldsymbol{x}}, \tilde{\boldsymbol{K}}_{\boldsymbol{y}} >_F}{||\tilde{\boldsymbol{K}}_{\boldsymbol{x}}||_F \cdot ||\tilde{\boldsymbol{K}}_{\boldsymbol{y}}||_F},$$

gdzie $\rho(\boldsymbol{K_x}, \boldsymbol{K_y}) \in [0, 1]$. Porównując HSIC$(S_{\boldsymbol{x},\boldsymbol{y}})$ z $\rho(\boldsymbol{K_x}, \boldsymbol{K_y})$ widzimy, że współczynnik zgodności $\rho(\boldsymbol{K_x}, \boldsymbol{K_y})$ jest po prostu znormalizowaną wersją współczynnika HSIC$(S_{\boldsymbol{x},\boldsymbol{y}})$.

Test istotności współczynnika zgodności dla wektorów losowych

Współczynnik zgodności ρ jest podstawą do weryfikacji hipotezy o niezależności wektorów $\boldsymbol{X}$ i $\boldsymbol{Y}$. Wprowadzimy dwie hipotezy:
$H_0 \colon P_{\boldsymbol{X},\boldsymbol{Y}} = P_{\boldsymbol{X}} P_{\boldsymbol{Y}}$, że $\boldsymbol{X}$ oraz $\boldsymbol{Y}$ są niezależne i $H_1 \colon \neg H_0$ są zależne.

Test istotności współczynnika zgodności dla wektorów losowych

Zhang i inni (2011) pokazali, że przy prawdziwości H_0 statystyka HSIC($S_{\boldsymbol{x},\boldsymbol{y}}$) ma ten sam rozkład asymptotyczny co statystyka

$$Z = \frac{1}{n^2} \sum_{i,j=1}^{n} \lambda_{\boldsymbol{X},i} \lambda_{\boldsymbol{Y},j} Z_{ij}^2,$$

gdzie Z_{ij}^2 są niezależnymi zmiennymi losowymi o rozkładzie χ_1^2, $\lambda_{\boldsymbol{X},i}$ jest i-tą nieujemną wartością własną macierzy $\tilde{\boldsymbol{K}}_{\boldsymbol{x}}$ oraz $\lambda_{\boldsymbol{y},j}$ jest j-tą nieujemną wartością własną macierzy $\tilde{\boldsymbol{K}}_{\boldsymbol{y}}$. Na tej podstawie został opracowany test hipotezy H_0 przeciwko H_1.

Test istotności współczynnika zgodności dla wektorów losowych

Z drugiej strony Doran i inni (2014) opracowali test permutacyjny badania niezależności macierzy jądrowych. Dokonaliśmy jego adaptacji do naszej sytuacji.

Jądrowy współczynnik zgodności dla procesów losowych

Niech $\boldsymbol{X}_s$ i $\boldsymbol{Y}_t$ będą procesami losowymi o wcześniej opisanej reprezentacji. Dla procesu $\boldsymbol{X}_s$ jądro gaussowskie jest równe

$$k_{\boldsymbol{X}_s}(\boldsymbol{x}(s), \boldsymbol{x}'(s)) = \exp(-\lambda_1 \|\boldsymbol{x}(s) - \boldsymbol{x}'(s)\|^2).$$

Ale

$$\begin{aligned}
\|\boldsymbol{x}(s) - \boldsymbol{x}'(s)\|^2 &= \int_{I_1} (\boldsymbol{x}(s) - \boldsymbol{x}'(s))^T (\boldsymbol{x}(s) - \boldsymbol{x}'(s)) ds \\
&= (\boldsymbol{\alpha} - \boldsymbol{\alpha}')^T \left(\int_{I_1} \boldsymbol{\Phi}_1^T(s) \boldsymbol{\Phi}_1(s) ds \right) (\boldsymbol{\alpha} - \boldsymbol{\alpha}') \\
&= (\boldsymbol{\alpha} - \boldsymbol{\alpha}')^T (\boldsymbol{\alpha} - \boldsymbol{\alpha}') = \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|^2.
\end{aligned}$$

Stąd

$$k_{\boldsymbol{X}_s}(\boldsymbol{x}(s), \boldsymbol{x}'(s)) = k_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\alpha}').$$

Podobnie

$$k_{\boldsymbol{Y}_t}(\boldsymbol{y}(t), \boldsymbol{y}'(t)) = k_{\beta}(\beta, \beta').$$

A to oznacza, że

$$\rho(K_{\boldsymbol{X}_s}, K_{\boldsymbol{Y}_t}) = \rho(K_{\boldsymbol{\alpha}}, K_{\beta}) = \frac{< \tilde{\boldsymbol{K}}_{\boldsymbol{\alpha}}, \tilde{\boldsymbol{K}}_{\beta} >_F}{\|\tilde{\boldsymbol{K}}_{\boldsymbol{\alpha}}\|_F \|\tilde{\boldsymbol{K}}_{\beta}\|_F}.$$

Zatem współczynnik zgodności procesów losowych $\boldsymbol{Y}_s$ i $\boldsymbol{Y}_t$ można wyrazić przez współczynnik zgodności wektorów losowych $\boldsymbol{\alpha}$ i β występujących w ich reprezentacji.

Test istotności współczynnika zgodności dla procesów losowych

Zauważmy również, że hipoteza zerowa $H_0 : \boldsymbol{X}_s \perp \boldsymbol{Y}_t$ niezależności procesów losowych $\boldsymbol{Y}_s$ i $\boldsymbol{Y}_t$ jest równoważna hipotezie $H_0 : \boldsymbol{\alpha} \perp \boldsymbol{\beta}$ niezależności wektorów losowych $\boldsymbol{\alpha}$ i $\boldsymbol{\beta}$ występujących w reprezentacji tych procesów. Możemy zatem posłużyć się testami przedstawionymi wcześniej.

Przykład

Jako przykład posłużyły nam dane Światowego Forum Ekonomicznego (`http://www.weforum.org`) dotyczące wskaźników socjo-ekonomicznych 38 europejskich krajów w latach 2008-2015.

1	Albania (AL)	14	Greece (GR)	27	Poland (PL)
2	Austria (AT)	15	Hungary (HU)	28	Portugal (PT)
3	Belgium (BE)	16	Iceland (IS)	29	Romania (RO)
4	Bosnia and Herzegovina (BA)	17	Ireland (IE)	30	Russian Federation (RU)
5	Bulgaria (BG)	18	Italy (IT)	31	Serbia (XS)
6	Croatia (HR)	19	Latvia (LV)	32	Slovak Republic (SK)
7	Cyprus (CY)	20	Lithuania(LT)	33	Slovenia (SI)
8	Czech Republic (CZ)	21	Luxembourg (LU)	34	Spain (ES)
9	Denmark (DK)	22	Macedonia FYR (MK)	35	Sweden (SE)
10	Estonia (EE)	23	Malta (MT)	36	Switzerland (CH)
11	Finland (FI)	24	Montenegro (ME)	37	Ukraine (UA)
12	France (FR)	25	Netherlands (NL)	38	United Kingdom (GB)
13	Germany (DE)	26	Norway (NO)		

Kraje użyte w analizie.

Przykład

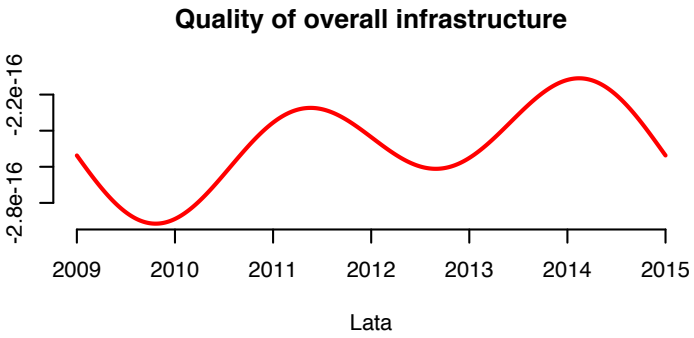
$\boldsymbol{X}$ (Infrastructure)	$\boldsymbol{Y}$ (Market size)
1. Quality of overall infrastructure	1. Domestic market size index
2. Quality of roads	2. Foreign market size index
3. Quality of port infrastructure	3. GDP valued at PPP
4. Quality of air transport infrastructure	4. Exports as a percentage of GDP
5. Available airline seat kilometers	
6. Quality of electricity supply	

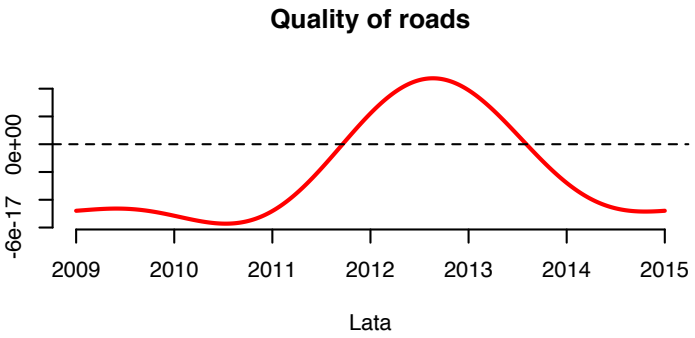
Zmienne użyte w analizie

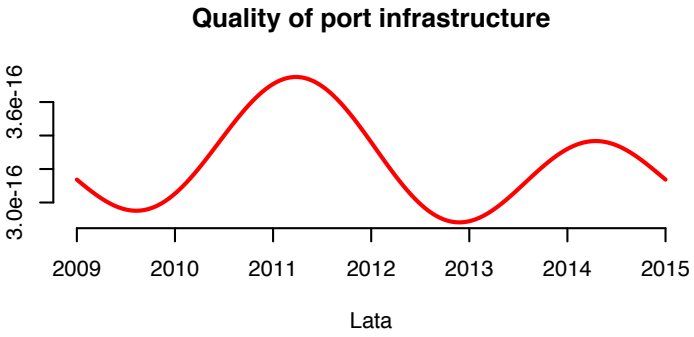
Przykład

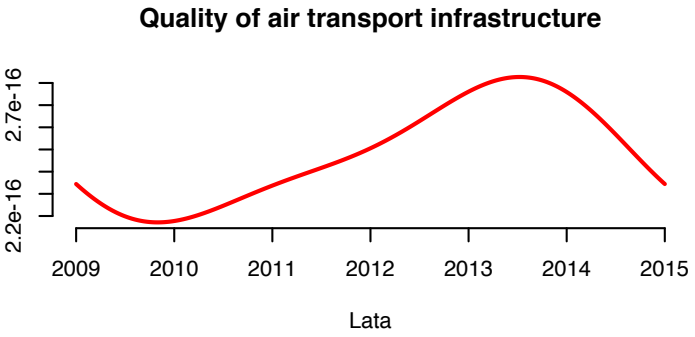
W pierwszym kroku wygładziliśmy dane. Została użyta baza Fouriera. Ponieważ liczba punktów czasowych była niewielka ($J = 7$), dla każdej zmiennej ustaliliśmy maksymalną wielkość bazy na 5. Na kolejnych dwóch wykresach przedstawione są średnie wartości analizowanych zmiennych dla danych funkcjonalnych.

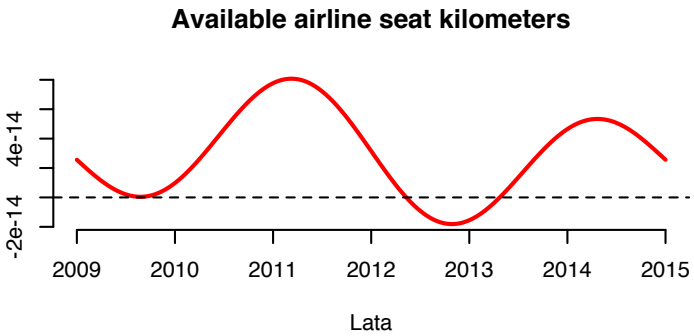
Przykład

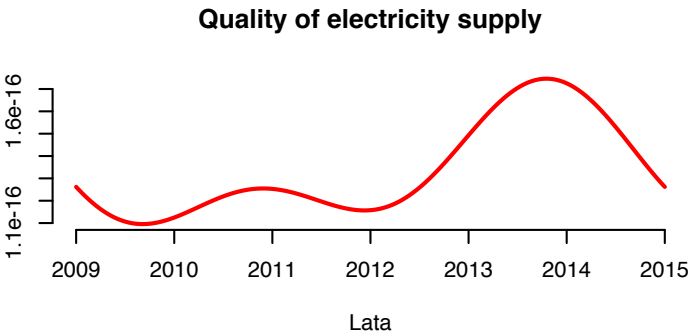

Quality of overall infrastructure
-2.8e-16
-2.2e-16
2009
2010
2011
2012
2013
2014
2015
Lata


Quality of roads
-6e-17
0e+00
2009
2010
2011
2012
2013
2014
2015
Lata

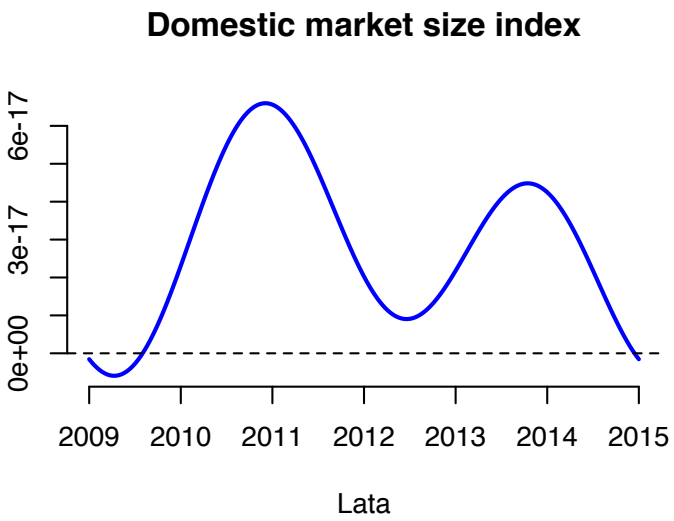

Quality of port infrastructure
3.0e-16
3.6e-16
2009
2010
2011
2012
2013
2014
2015
Lata

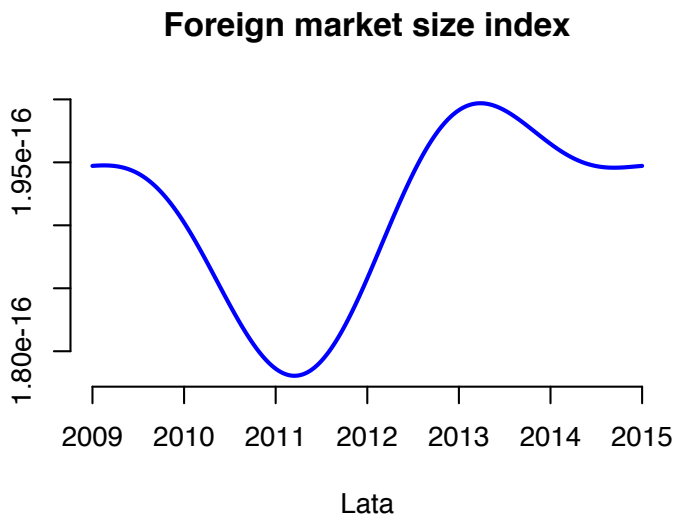

Quality of air transport infrastructure
2.2e-16
2.7e-16
2009
2010
2011
2012
2013
2014
2015
Lata

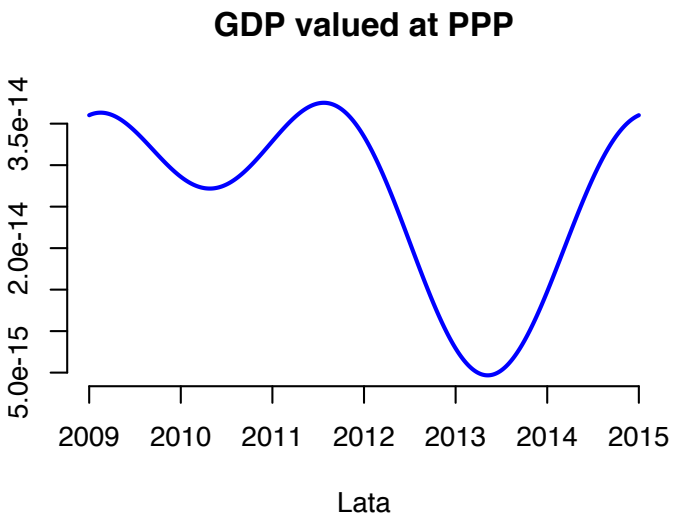

Available airline seat kilometers
-2e-14
4e-14
2009
2010
2011
2012
2013
2014
2015
Lata

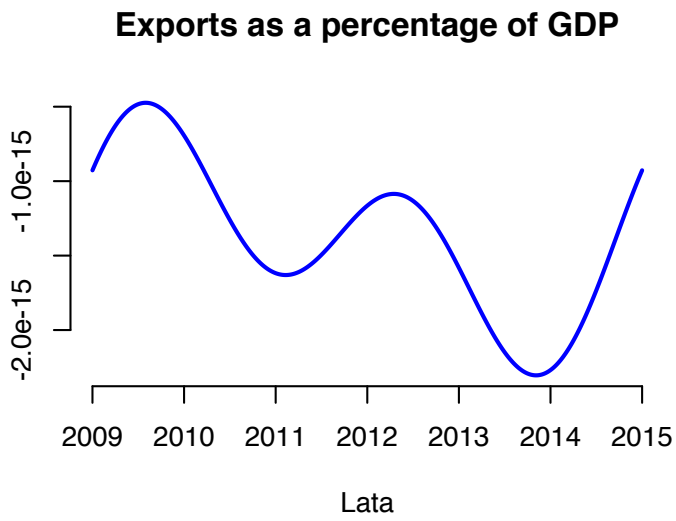

Quality of electricity supply
1.1e-16
1.6e-16
2009
2010
2011
2012
2013
2014
2015
Lata

Przykład


Domestic market size index
0e+00
3e-17
6e-17
2009
2010
2011
2012
2013
2014
2015
Lata


Foreign market size index
1.80e-16
1.95e-16
2009
2010
2011
2012
2013
2014
2015
Lata


GDP valued at PPP
5.0e-15
2.0e-14
3.5e-14
2009
2010
2011
2012
2013
2014
2015
Lata


Exports as a percentage of GDP
-2.0e-15
-1.0e-15
2009
2010
2011
2012
2013
2014
2015
Lata

Przykład

W następnym kroku zostały policzone opisane wcześniej współczynniki i p-wartości testów permutacyjnych dla ich istotności.

	Wartość współczynnika	p wartość
ρV	0.941	0.0001
dcor	0.965	0.0001
HSIC	0.974	0.0046
KTA	1.000	0.0046

Zgodnie z oczekiwaniami wszystkie współczynniki są bliskie 1. Jednakże ich wysokie wartości nie oznaczają jeszcze statystycznej istotności relacji pomiędzy grupami zmiennych. Dopiero małe p-wartości wskazują na ich istotność.

Literatura


CORTES C., MOHRI M., ROSTAMIZADEH A. (2012): Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research* 13, 795–828.


DORAN G., MUANDET K., ZHANG K., SCHÖLKOPF B. (2014): A Permutation-based kernel conditional independence test. Proceedings of the 30*th* Conference on Uncertainty in Artificial Intelligence 132–141.


GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W. (2017): Correlation analysis for multivariate functional data. *Studies in Classification Data Analysis, and Knowledge Organization: Data Science* 243–258.


GÓRECKI, T., KRZYŚKO, M., WOŁYŃSKI, W. (2017): Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data – submitted.


GRETTON A., BOUSQUET O., SMOLA A., SCHÖLKOPF B. (2005): Measuring statistical dependence with Hilbert-Schmidt norms. *Lecture Notes in Computer Science* 3734, 63–77.


HORVÁTH, L., KOKOSZKA, P. (2012): Inference for Functional Data with Applications, Springer.

Literatura

JOSSE, J., HOLMES, S. (2016): Measuring multivariate association and beyond. Statistics Surveys 10, 132–167.

JOSSE, J., PAGES, J., HUSSON, F. (2008): Testing the significance of the *RV* coefficient. *Computational Statistics and Data Analysis* 53, 82–91.

LEURGANS, S.E., MOYEED, R.A., SILVERMAN, B.W. (1993): Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society: Series B* 55, 725–740.

ROBERT, P., ESCOUFIER, Y. (1976): A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Applied Statistics* 25, 257–265.

SZÉKELY, G.J., RIZZO, M.L., BAKIROV, N.K. (2007): Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6), 2769–2794.

ZHANG K., PETERS J., JANZING D., SCHÖLKOPF B., (2011): Kernel-based conditional independence test and application in causal discovery. Proceedings of the 27*th* Conference on Uncertainty in Artificial Intelligence 804–813.