Metody redukcji wymiaru dla danych funkcjonalnych

Mirosław Krzyśko i Łukasz Waszak

Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza, Poznań

STATYSTYKA – WIEDZA – ROZWÓJ. KONFERENCJA NAUKOWA
Z OKAZJI MIĘDZYNARODOWEGO ROKU STATYSTYKI
Łódź 17–18.10.2013

Dane funkcjonalne

Dana jest cecha statystyczna Y. Jej wartości obserwujemy w J wybranych momentach czasowych, które nie muszą być równoodległe. Momenty czasowe oznaczać będziemy przez t_j, a wartości cechy statystycznej Y zaobserwowane w tych momentach przez y_j, $j = 1, 2, ..., J$. Wówczas nasze dane składają się z J par $\{t_i, y_i\}$. Dane dyskretne chcemy przekształcić w funkcję ciągłą $x(t)$, gdzie $t \in I$. Zbiór I jest zwartym przedziałem takim, że $t_i \in I$, dla $i = 1, 2, ..., J$. Będziemy zakładać, że funkcja $x(t)$ jest postaci

$$x(t) = \sum_{k=0}^{K} c_k \varphi_k(t), t \in I,$$

gdzie $\{\varphi_k\}$ jest znanym układem bazowych funkcji ortonormalnych, natomiast $c_0, ..., c_K$ są nieznanymi współczynnikami, które należy wyestymować metodą najmniejszych kwadratów z danych $\{t_i, y_i\}$, i=1,2,...,J.

Dane funkcjonalne

Przypomnijmy, że układ $\{\varphi_k\}$ w przestrzeni $L_2(I)$ jest układem ortonormalnym, gdy iloczyn skalarny każdych dwóch funkcji z tej przestrzeni jest równy:

$$< \varphi_i(t), \varphi_j(t) >= \int_I \varphi_i(t)\varphi_j(t)dt = \delta_{ij}.$$

Przykładowymi układami bazowych funkcji ortonormalnych są: układ Fouriera oraz układ wielomianów Legendre'a. Ortonormalny układ Fouriera na przestrzeni $L_2([0, T])$ ma postać:

$$\varphi_0(t) = 1, \varphi_{2k-1}(t) = \sqrt{\frac{2}{T}} \sin \frac{2\pi kt}{T}, \varphi_{2k}(t) = \sqrt{\frac{2}{T}} \cos \frac{2\pi kt}{T},$$

$k = 1, 2, \ldots, t \in [0, T]$.

Natomiast ortonormalny układ wielomianów Legendre'a na przestrzeni $L_2([-1,1])$ ma postać:

$$\widetilde{p}_k(x) = \sqrt{\frac{2k+1}{2}} p_k(x),$$

gdzie
$p_{k+1}(x) = \frac{1}{k+1}[(2k+1)xp_k(x) - kp_{k-1}(x)], k \geq 1, p_1(x) = x, p_0(x) = 1.$
Każdy skończony przedział $[a,b]$ można przekształcić na przedział $[-1,1]$ podstawiając:

$$x = \frac{2}{b-a}t - \frac{b+a}{b-a}, t \in [a,b], x \in [-1,1].$$

Dane funkcjonalne

Niech $\boldsymbol{y} = (y_1, y_2, \ldots, y_J)'$, $\boldsymbol{c}_k = (c_0, \ldots, c_K)'$ oraz niech $\boldsymbol{\Phi}$ będzie macierzą $J \times (K+1)$ o elementach $\varphi_k(t_j), j = 1, \ldots, J, k = 0, \ldots, K$. Poszukujemy zatem wektora $\hat{\boldsymbol{c}}$, dla którego funkcja celu

$$S(\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})'(\boldsymbol{y} - \boldsymbol{\Phi c})$$

osiąga minimum. Poszukiwany wektor ma postać

$$\hat{\boldsymbol{c}} = \left(\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}'\boldsymbol{y}.$$

Stopień gładkości uzyskanej funkcji ciągłej $x(t)$ zależy od liczby wyrazów K kombinacji liniowej reprezentującej tę funkcję. Optymalna wartość K wybierana jest za pomocą bayesowskiego kryterium informacyjnego BIC postaci:

$$BIC(x(t)) = ln\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{2}\right) + K\left(\frac{lnJ}{J}\right),$$

gdzie $\boldsymbol{e} = (e_1, ..., e_J)'$,$e_j = y_j - \sum_{k=0}^{K} \hat{c}_k \varphi_k(t_j)$,$j = 1, 2, \ldots, J$. Za optymalną wartość K, przyjmujemy tę wartość, która minimalizuje BIC$(x(t))$. Kryterium BIC mierzy dokładność dopasowania przyjętego modelu funkcji $x(t)$ do danych $\{t_i, y_i\}, i = 1, 2, ..., J$ (patrz Shmueli (2010)).

Dane funkcjonalne

Załóżmy teraz, że mamy n niezależnych par wartości $\{t_{ij}, y_{ij}\}$, $i = 1, ..., J, j = 1, ..., n$. Te dane dyskretne przekształcamy w funkcje ciągłe postaci:

$$x_j(t) = \sum_{k=0}^{K_j} \hat{c}_{jk}\varphi_k(t), j = 1, ..., n, t \in I.$$

Spośród wszystkich $K_1, K_2, ..., K_n$ wybieramy wartość modalną (powiedzmy K) i przyjmujemy, że każda funkcja $x_j(t)$ ma postać

$$x_j(t) = \sum_{k=0}^{K} \hat{c}_{jk}\varphi_k(t), j = 1, ..., n, t \in I.$$

Dane funkcjonalne

Uzyskany w ten sposób zbiór funkcji $\{x_1(t), ..., x_n(t)\}$ nazywa się zbiorem danych funkcjonalnych (patrz Ramsay i Silverman (2005)).
Do tej pory zajmowaliśmy się danymi dotyczącymi pojedynczej cechy statystycznej. Rozważania nasze możemy uogólnić na przypadek $p \geq 2$ cech statystycznych. Wówczas nasze dane będą się składały z n niezależnych funkcji wektorowych postaci $\boldsymbol{x}_j(t) = [x_{1j}(t), x_{2j}(t), ..., x_{pj}(t)]'$, $j = 1, ..., n$. Dane $\{\boldsymbol{x}_1(t), ..., \boldsymbol{x}_n(t)\}$ nazywać będziemy wielozmiennymi danymi funkcjonalnymi .

PCA dla wielozmiennych danych funkcjonalnych

Wielozmienne dane funkcjonalne możemy traktować jako realizacje wielozmiennych procesów stochastycznych $\boldsymbol{X}(t)$ z ciągłym parametrem $t \in I$, gdzie $\boldsymbol{X}(t) = [X_1(t), X_2(t), ..., X_p(t)]'$, $p \geq 2$. Będziemy zakładać, że $E(\boldsymbol{X}(t)) = \boldsymbol{0}$ oraz, że $\boldsymbol{X}(t) \in L_2(I)^p$, gdzie $L_2(I)$ jest przestrzenią Hilberta funkcji całkowalnych z kwadratem na przedziale I wyposażoną w iloczyn skalarny:

$$< u, v >= \int_I u(t)v(t)dt.$$

Weźmy pod uwagę przypadek gdy i-ta składowa procesu stochastycznego $\boldsymbol{X}(t)$ może być reprezentowana za pomocą skończonej liczby ortonormalnych funkcji bazowych $\{\varphi_k\}$

$$X_i(t) = \sum_{k=0}^{K_i} c_{ik}\varphi_k(t), t \in I, i = 1, 2, ..., p,$$

gdzie c_{ik} są zmiennymi losowymi takimi, że $E(c_{ik}) = 0, Var(c_{ik}) < \infty$, $i = 1, 2, ..., p, k = 0, ..., K_i$.

Niech

$$\boldsymbol{c} = (\boldsymbol{c}_1', \boldsymbol{c}_2', ..., \boldsymbol{c}_p'),$$

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}'_{K_1}(t) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}'_{K_2}(t) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\varphi}'_{K_p}(t) \end{bmatrix},$$

gdzie $\boldsymbol{c}_i = (c_{i0}, ..., c_{iK_i})'$,$\boldsymbol{\varphi}_{K_i}(t) = (\varphi_0(t), ..., \varphi_{K_i}(t))'$, $i = 1, ..., p$.
Wówczas

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c}, \qquad t \in I, E(\boldsymbol{c}) = \mathbf{0}.$$

Naszym celem jest znalezienie iloczynu skalarnego

$$U = < \boldsymbol{u}(t), \boldsymbol{X}(t) >$$

mającego maksymalną wariancję dla wszystkich $\boldsymbol{u}(t) \in L_2(I)^p$ takich, że $< \boldsymbol{u}(t), \boldsymbol{u}(t) >= 1$.

Niech

$$\lambda_1 = \sup_{\boldsymbol{u}(t)\in L_2(I)^p} Var(< \boldsymbol{u}(t), \boldsymbol{X}(t) >) = Var(< \boldsymbol{u}_1(t), \boldsymbol{X}(t) >),$$

gdzie $< \boldsymbol{u}_1(t), \boldsymbol{u}_1(t) >= 1$. Iloczyn skalarny $U_1 =< \boldsymbol{u}_1(t), \boldsymbol{X}(t) >$ nosi nazwę pierwszej funkcjonalnej składowej głównej, a funkcja wektorowa $\boldsymbol{u}_1(t)$ jest nazywana pierwszą wektorową funkcją wagową. Następnie poszukujemy drugiej funkcjonalnej składowej głównej $U_2 =< \boldsymbol{u}_2(t), \boldsymbol{X}(t) >$, która maksymalizuje $Var(< \boldsymbol{u}(t), \boldsymbol{X}(t) >)$, spełnia warunek $< \boldsymbol{u}_2(t), \boldsymbol{u}_2(t) >= 1$ i jest nieskorelowana z pierwszą funkcjonalną składową główną U_1, tj. spełnia warunek $< \boldsymbol{u}_1(t), \boldsymbol{u}_2(t) >= 0$.

Ogólnie k-ta funkcjonalna składowa główna $U_k =< \boldsymbol{u}_k(t), \boldsymbol{X}(t) >$ spełnia warunki

$$\lambda_k = \sup_{\boldsymbol{u}(t) \in L_2(I)^p} Var(< \boldsymbol{u}(t), \boldsymbol{X}(t) >) = Var(< \boldsymbol{u}_k(t), \boldsymbol{X}(t) >),$$

$$< \boldsymbol{u}_i(t), \boldsymbol{u}_j(t) >= \delta_{ij}, \qquad i, j = 1, ..., k.$$

Parę $(\lambda_k, \boldsymbol{u}_k(t))$ nazywać będziemy k-tym układem głównym procesu losowego $\boldsymbol{X}(t)$.

Wcześniej pokazaliśmy, że proces $\boldsymbol{X}(t)$ ma reprezentację $\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c}, t \in I$. Weźmy teraz pod uwagę składowe główne wektora losowego $\boldsymbol{c}$. k-ta składowa główna $U_k^* =< \boldsymbol{u}_k, \boldsymbol{c} >$ tego wektora spełnia warunki:

$$\gamma_k = \sup_{\boldsymbol{u} \in \mathbb{R}^{K+p}} Var(< \boldsymbol{u}, \boldsymbol{c} >) = \sup_{\boldsymbol{u} \in \mathbb{R}^{K+p}} \boldsymbol{u}' Var(\boldsymbol{c}) \boldsymbol{u}$$

$$= \sup_{\boldsymbol{u} \in \mathbb{R}^{K+p}} \boldsymbol{u}' \boldsymbol{\Sigma} \boldsymbol{u},$$

$$\boldsymbol{u}_i' \boldsymbol{u}_j = \delta_{ij},$$

gdzie $i, j = 1, ..., k, K = K_1 + K_2 + ... + K_p$. Parę $(\gamma_k, \boldsymbol{u}_k)$ nazywać będziemy k-tym układem głównym wektora losowego $\boldsymbol{c}$.

Wyznaczenie k-tego układu głównego wektora losowego $\boldsymbol{c}$ sprowadza się zatem do znalezienia wartości własnych i odpowiadających im wektorów własnych macierzy kowariancji $\boldsymbol{\Sigma}$ tego wektora normowanych tak, by $\boldsymbol{u}_i'\boldsymbol{u}_j = \delta_{ij}$. Zauważmy, że $tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{K} Var(c_i) = \sum_{i=1}^{K} \gamma_i$, czyli suma wariancji zmiennych pierwotnych $c_1, ..., c_K$ jest równa sumie wariancji składowych głównych. Zatem wskaźnik

$$\frac{\gamma_1 + ... + \gamma_r}{\gamma_1 + ... + \gamma_K} \cdot 100\%$$

jest miarą wyjaśnienia całkowitej zmienności składowych wektora $\boldsymbol{c}$ przez r pierwszych składowych głównych.

Twierdzenie

k-ty układ główny $(\lambda_k, \boldsymbol{u}_k(t))$ *procesu losowego* $\boldsymbol{X}(t)$ *oraz k-ty układ główny* $(\gamma_k, \boldsymbol{u})$ *wektora losowego* $\boldsymbol{c}$ *związane są zależnościami:*

$$\lambda_k = \gamma_k, \qquad \boldsymbol{u}_k(t) = \boldsymbol{\Phi}(t)\boldsymbol{u}_k, \qquad t \in I, k = 1, ..., K + p,$$

gdzie $K = K_1 + K_2 + ... + K_p$.

Macierz kowariancji $\boldsymbol{\Sigma}$ nie jest znana i musi być zastąpiona estymatorem z próby losowej. Niech $\hat{\boldsymbol{x}}_1(t), \hat{\boldsymbol{x}}_2(t), ..., \hat{\boldsymbol{x}}_n(t)$ będą n niezależnymi realizacjami postaci $\hat{\boldsymbol{x}}_i(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{c}}_i$ procesu losowego $\boldsymbol{X}(t)$, gdzie wektory $\hat{\boldsymbol{c}}_i$ są scentrowane, $i = 1, 2, ..., n$. Niech $\hat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_1, \hat{\boldsymbol{c}}_2, ..., \hat{\boldsymbol{c}}_n)'$. Wówczas

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\boldsymbol{C}}'\hat{\boldsymbol{C}}.$$

Jeżeli $n > K$, to macierz $\hat{\boldsymbol{\Sigma}}$ z prawdopodobieństwem 1 jest dodatnio określona. Niech $\hat{\gamma}_1 \geq \hat{\gamma}_2 \geq ... \geq \hat{\gamma}_s$ będą niezerowymi wartościami własnymi macierzy $\hat{\boldsymbol{\Sigma}}$, a $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_s$ odpowiadającymi im wektorami własnymi, gdzie $s =$rząd$(\hat{\boldsymbol{\Sigma}})$.

Ponadto k-ty układ główny procesu losowego $\boldsymbol{X}(t)$ wyznaczony z próby ma postać:

$$(\hat{\lambda}_k = \hat{\gamma}_k, \hat{\boldsymbol{u}}_k(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{u}}_k), \qquad k = 1, ..., s.$$

Współrzędna rzutu i-tej realizacji $\hat{\boldsymbol{x}}_i(t)$ procesu $\boldsymbol{X}(t)$ na k-tą funkcjonalną składową główną z próby jest równa:

$$\hat{U}_{ik} =< \hat{\boldsymbol{u}}_k(t), \hat{\boldsymbol{x}}_i(t) >=< \boldsymbol{\Phi}(t)\hat{\boldsymbol{u}}_k, \boldsymbol{\Phi}(t)\hat{\boldsymbol{c}}_i >$$

$$= \hat{\boldsymbol{u}}'_k < \boldsymbol{\Phi}(t), \boldsymbol{\Phi}(t) > \hat{\boldsymbol{c}}_i = \hat{\boldsymbol{u}}'_k \hat{\boldsymbol{c}}_i,$$

dla $i = 1, 2, ..., n, k = 1, 2, ..., s$. Zatem współrzędne rzutu i-tej realizacji $\hat{\boldsymbol{x}}_i(t)$ procesu $\boldsymbol{X}(t)$ na płaszczyznę dwóch pierwszych funkcjonalnych składowych głównych z próby są równe $(\hat{\boldsymbol{u}}'_1\hat{\boldsymbol{c}}_i, \hat{\boldsymbol{u}}'_2\hat{\boldsymbol{c}}_i), i = 1, 2, ..., n$.

Przykład - Ochrona środowiska i rozwój zrównoważony

Przykład. Dane udostępnione przez prof. Waldemara Ratajczaka z Wydziału Nauk Geograficznych i Geologicznych UAM.
Liczba cech: $p = 6$
Liczba województw: $n = 16$
Liczba lat obserwacji: $J = 10$
Przyjmujemy:
Przedział czasowy: $[0, T] = [0, 10]$
Momenty czasowe: $t_1 = 0,5(2002), t_2 = 1,5(2003), ..., t_{10} = 9,5(2011)$

Tabela : Województwa.

Nr	Województwo
1	ŁÓDZKIE
2	MAZOWIECKIE
3	MAŁOPOLSKIE
4	ŚLĄSKIE
5	LUBELSKIE
6	PODKARPACKIE
7	PODLASKIE
8	ŚWIĘTOKRZYSKIE
9	LUBUSKIE
10	WIELKOPOLSKIE
11	ZACHODNIOPOMORSKIE
12	DOLNOŚLĄSKIE
13	OPOLSKIE
14	KUJAWSKO-POMORSKIE
15	POMORSKIE
16	WARMIŃSKO-MAZURSKIE

Przykład - Ochrona środowiska i rozwój zrównoważony

Tabela : Cechy.

Nr	Cechy dotyczące ochrony środowiska
1	Emisja zanieczyszczeń gazowych (t/km^2)
2	Emisja zanieczyszczeń pyłowych (kg/km^2)
3	Odpady wytworzone w ciągu roku (t/km^2)
4	Ścieki odprowadzane ogółem $(^*dam^3/1000 mieszk.)$
5	Ścieki przemysłowe odprowadzane $(^*dam^3/1000 mieszk.)$
6	Zużycie wody na potrzeby gospodarki narodowej i ludności $(^*dam^3/1000 mieszk.)$

$^*dam^3$ (dekametr sześcienny) $= 1000 m^3$

Przykład - Ochrona środowiska i rozwój zrównoważony

Tabela : Wartości współczynników wagowej funkcji wektorowej pierwszej funkcjonalnej składowej głównej.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	Udział %
1	0,6947	-0,0007	-0,0179	-0,0056	0,0189	-0,0220	-0,0104	-0,0064	-0,0046	40,48
2	0,2927	0,0794	0,0094	0,04010	0,0138	0,0114	-0,0101	0,0299	0,0011	17,94
3	0,6443	0,0551	-0,0088	0,0086	0,0129	-0,0010	-0,0074	0,0147	-0,0002	37,65
4	0,0008	0,0001	0,0000	0,0001	0,0000	0,0001	0,0000	0,0001	0,0000	0,00
5	-0,0317	0,0001	0,0009	-0,0005	-0,0009	0,0003	0,0003	0,0009	-0,0004	1,86
6	-0,0355	0,0013	0,0011	0,0006	-0,0008	0,0003	0,0004	0,0013	-0,0004	2,07

$$u_i(t) = \boldsymbol{u}_i' \boldsymbol{\varphi}(t), \qquad i = 1, 2, \ldots, 6,$$

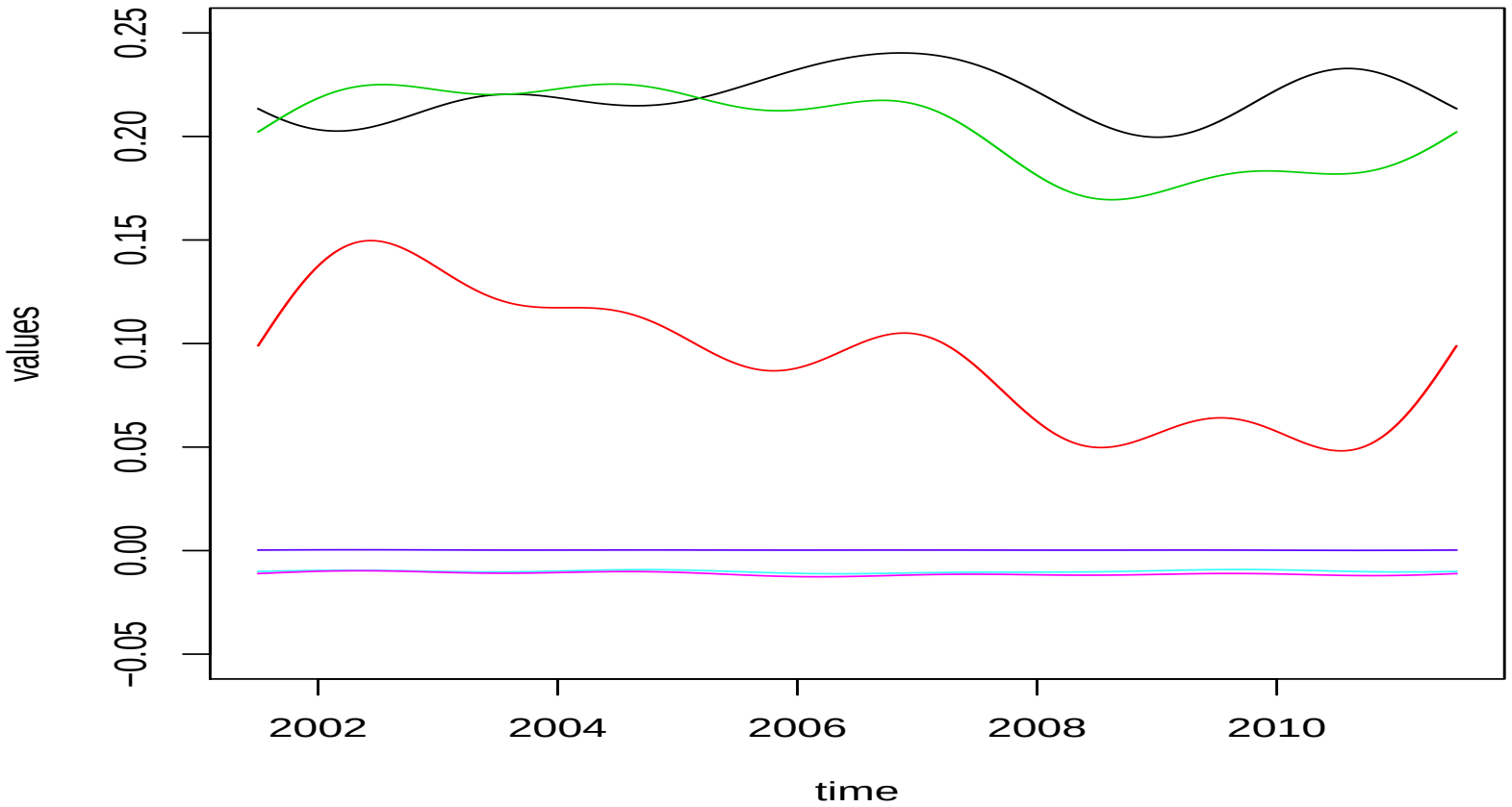
$$\boldsymbol{u}_i = (u_{i0}, u_{i1}, \ldots, u_{iK})', \qquad \boldsymbol{\varphi}(t) = (\varphi_0(t), \varphi_1(t), \ldots, \varphi_K(t))'.$$

Nierówność Cauchy'ego–Schwarza $| < \boldsymbol{x}, \boldsymbol{y} > | \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|, \qquad \|\boldsymbol{x}\| = \sqrt{< \boldsymbol{x}, \boldsymbol{x} >}.$

$$|u_i(t)| = |\boldsymbol{u}_i' \boldsymbol{\varphi}(t)| \leq \|\boldsymbol{u}_i\| \|\boldsymbol{\varphi}(t)\|, \qquad \|\boldsymbol{u}_i\| = \sqrt{\sum_{k=0}^{K} u_{ik}^2}.$$

Przykład - Ochrona środowiska i rozwój zrównoważony

Rysunek : Wykres 6 funkcji wagowych pierwszej funkcjonalnej składowej głównej.



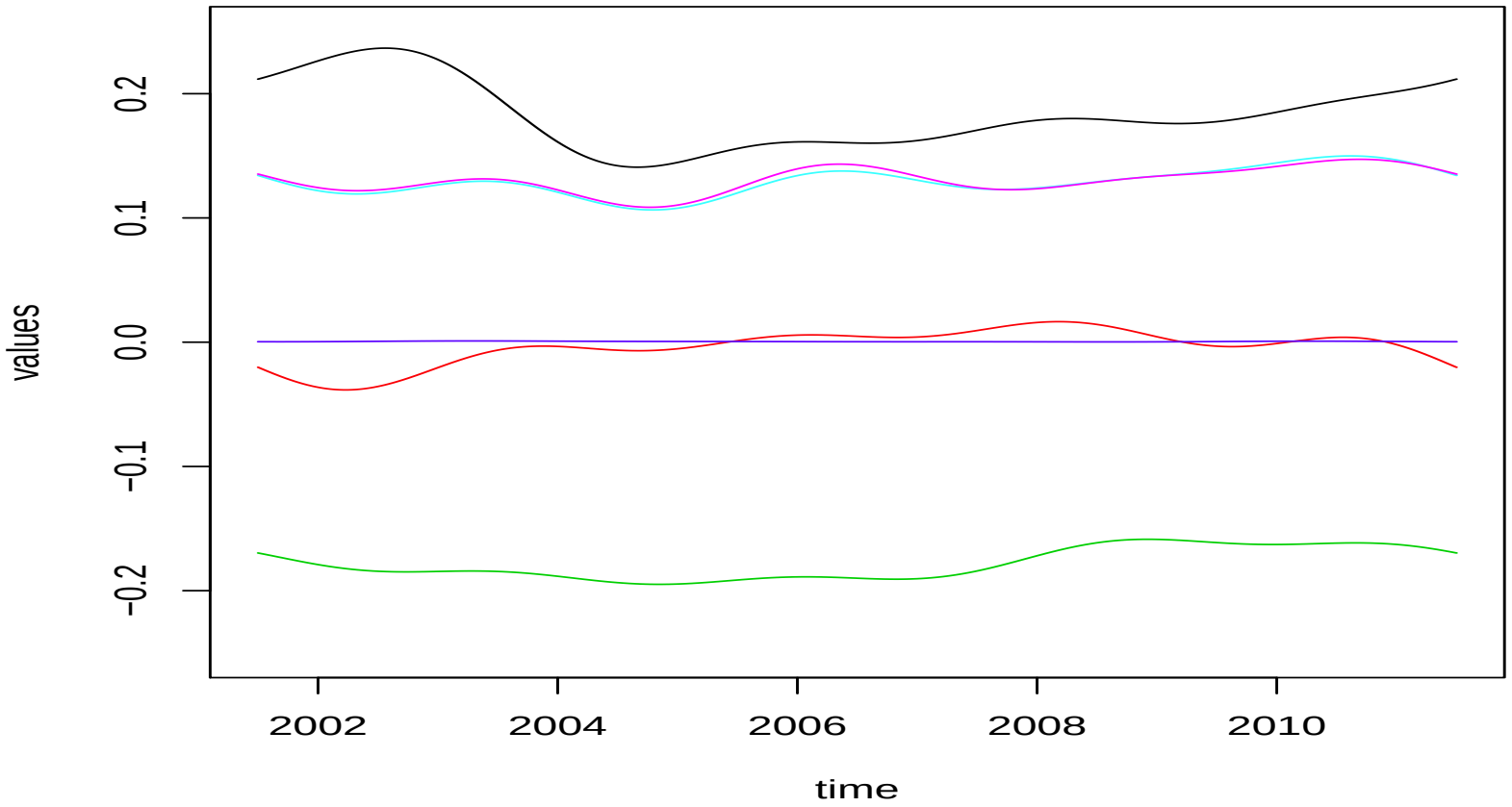
Przykład - Ochrona środowiska i rozwój zrównoważony

Tabela : Wartości współczynników wagowej funkcji wektorowej drugiej funkcjonalnej składowej głównej.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	Udział %
1	0,5771	-0,0005	0,0680	0,0392	0,0193	0,0165	-0,0105	-0,0091	-0,0114	28,89
2	-0,0122	-0,0227	-0,0289	-0,0087	-0,0092	-0,0140	0,0010	-0,0157	0,0005	2,27
3	-0,5636	-0,0314	0,0210	-0,0042	-0,0070	0,0018	0,0017	-0,0078	0,0034	27,99
4	0,0017	0,0004	0,0003	-0,0001	-0,0002	-0,0002	-0,0003	-0,0001	-0,0002	0,09
5	0,4080	-0,0210	0,0120	-0,0081	0,0096	-0,0065	-0,0155	-0,0117	0,0055	20,28
6	0,4120	-0,0164	0,0081	-0,0069	0,0122	-0,0041	-0,0163	-0,0115	0,0072	20,48

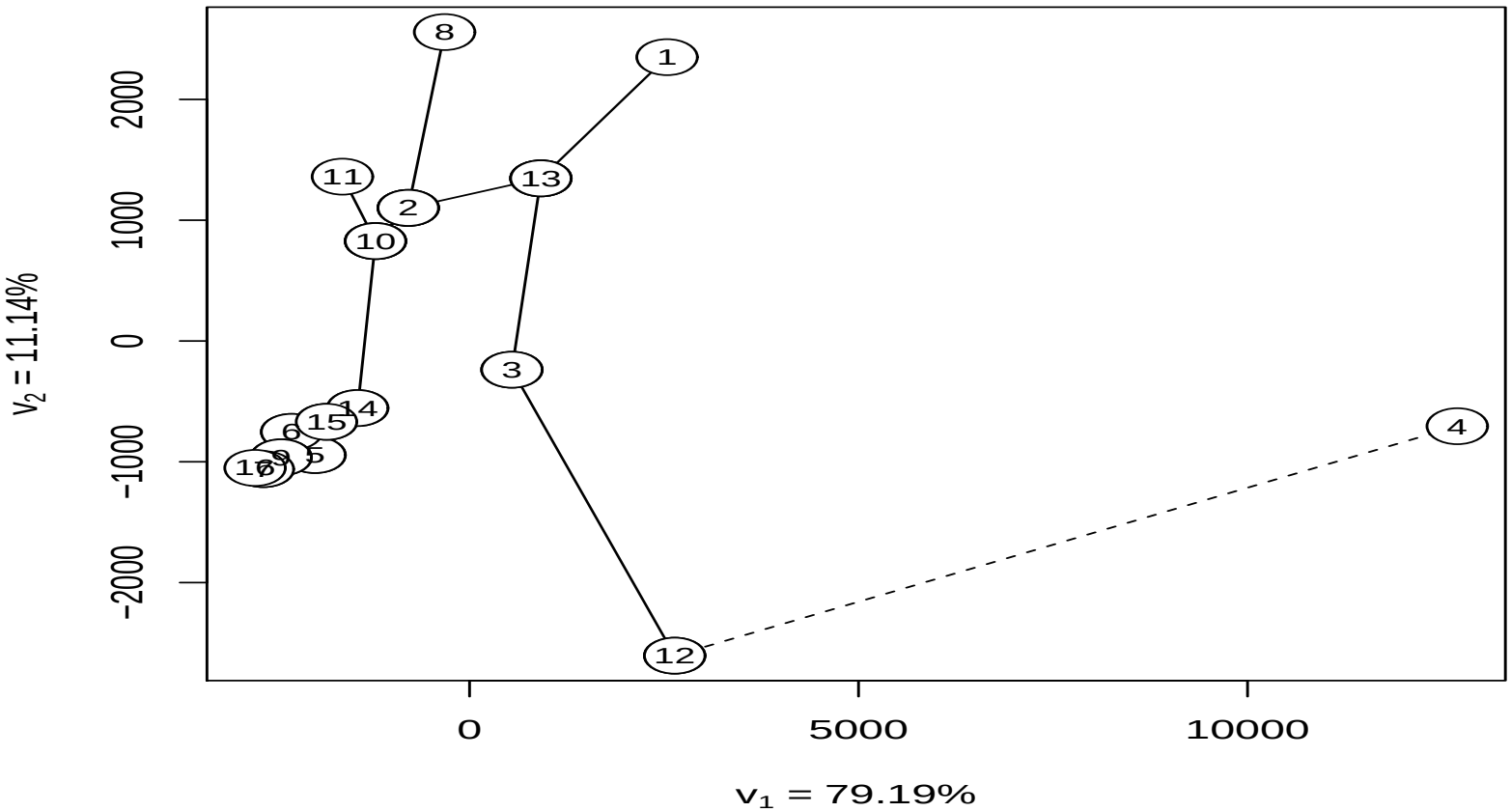
Przykład - Ochrona środowiska i rozwój zrównoważony

Rysunek : Wykres 6 funkcji wagowych drugiej funkcjonalnej składowej głównej.



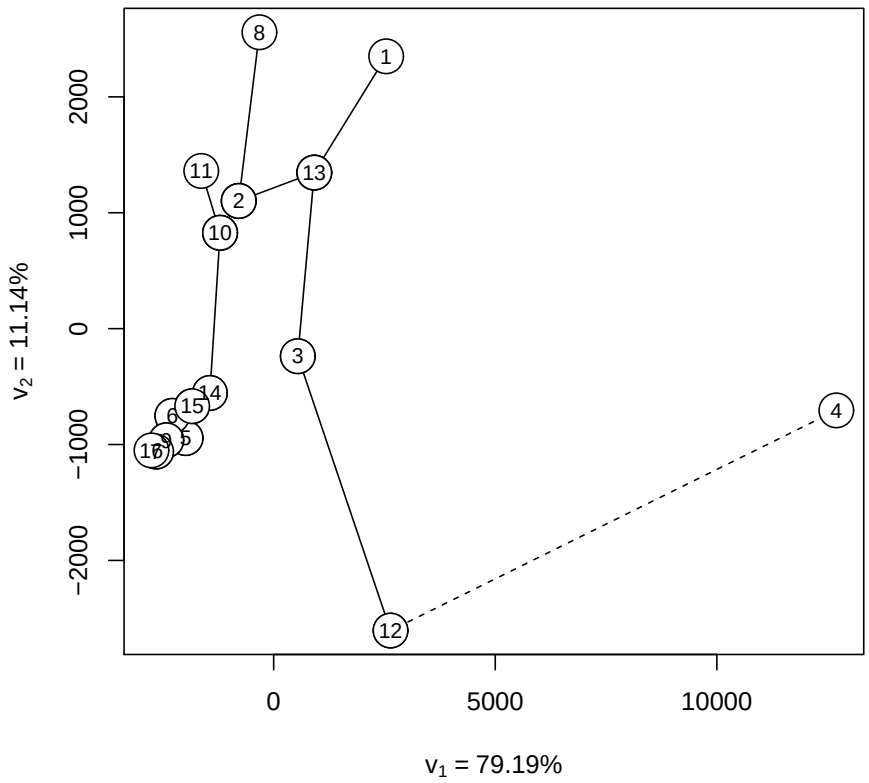
Przykład - Ochrona środowiska i rozwój zrównoważony

Rysunek : Położenie 16 województw w układzie dwóch pierwszych funkcjonalnych składowych głównych wraz z rozpiętym dendrytem.



Przykład - Ochrona środowiska i rozwój zrównoważony

4 - ŚLĄSKIE
12 - DOLNOŚLĄSKIE
3 - MAŁOPOLSKIE
13 - OPOLSKIE
1 - ŁÓDZKIE
2 - MAZOWIECKIE
10 - WIELKOPOLSKIE
11 - ZACHODNIOPOMORSKIE
8 - ŚWIĘTOKRZYSKIE



Rysunek : Położenie 16 województw w układzie dwóch pierwszych funkcjonalnych składowych głównych wraz z rozpiętym dendrytem.

Literatura

- Ramsay, J.O., Silverman, B.W. (2005). Functional Data Analysis, Second Edition, Springer.
- Shmueli, G. (2010): To explain or to predict?, Statistical Science 25(3), 289-310.