Wybór zmiennych w klasyfikacji dla wielowymiarowych danych funkcjonalnych

Tomasz Górecki, Mirosław Krzyśko, Waldemar Wołyński

Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

II Kongres Statystyki Polskiej
Warszawa 10-12.07.2018

Wstęp

Rozważamy zagadnienie klasyfikacji obiektu opisanego p-wymiarowym procesem stochastycznym $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top$ do jednej z q klas. Ponadto zakładamy, że dysponujemy n-elementową próbą uczącą

$$\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\},$$

gdzie $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ są niezależnymi realizacjami procesu losowego $\mathbf{X}$, a $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ etykietami takimi, że $\mathbf{y}_i \in \mathbb{R}^q$, $i = 1, 2, \ldots, n$.

W prezentacji przedstawimy algorytm pozwalający na redukcję wymiaru wektora losowego $\mathbf{X}$ z jednoczesnym zachowaniem efektywności procedur klasyfikacyjnych.

Plan

1. Model danych funkcjonalnych.
2. Funkcjonalna kowariancja odległościowa.
3. Funkcjonalny współczynnik niezależności HSIC.
4. Algorytm wyboru zmiennych.
5. Przykład.
6. Literatura.

Dane funkcjonalne

Załóżmy, że $\boldsymbol{X} \in L_2^p(I)$, gdzie $L_2(I)$ jest przestrzenią Hilberta funkcji całkowalnych z kwadratem na przedziale I. Ponadto załóżmy, że $\mathrm{E}(\boldsymbol{X}) = \mathbf{0}$.

Z powyższego wynika, że każda składowa procesu $\mathbf{X}$ może być przedstawiona w następującej postaci:

$$X_k(t) = \sum_{b=0}^{\infty} \alpha_{kb}\varphi_b(t), \ \ t \in I,$$

przy czym funkcje $\varphi_1, \varphi_2, \ldots$ tworzą bazę w przestrzeni $L_2(I)$.

Dane funkcjonalne

W praktyce posługujemy się przybliżoną reprezentacją wykorzystującą jedynie skończoną liczbę pierwszych funkcji bazowych.
Załóżmy zatem, że k-ta składowa procesu $\boldsymbol{X}$ ma następującą reprezentację:

$$X_k(t) = \sum_{b=0}^{B_k} \alpha_{kb}\varphi_b(t), \ \ t \in I,$$

gdzie liczba B_k decyduje o stopniu gładkości funkcji X_k (im mniejsza wartość B_k tym większy stopień wygładzania).

Dane funkcjonalne

Przyjmijmy następujące oznaczenia:

$$\boldsymbol{\alpha} = (\alpha_{10}, \ldots, \alpha_{1B_1}, \ldots, \alpha_{p0}, \ldots, \alpha_{pB_p})^\top,$$

oraz

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1^\top(t) & \mathbf{0}^\top & \ldots & \mathbf{0}^\top \\ \mathbf{0}^\top & \boldsymbol{\varphi}_2^\top(t) & \ldots & \mathbf{0}^\top \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0}^\top & \mathbf{0}^\top & \ldots & \boldsymbol{\varphi}_p^\top(t) \end{bmatrix},$$

gdzie $\boldsymbol{\varphi}_k(t) = (\varphi_0(t), \varphi_1(t), \ldots, \varphi_{B_k}(t))^\top$, $k = 1, 2, \ldots, p$.
Używając powyższej notacji macierzowej, proces $\mathbf{X}$ może być zapisany w następujący sposób:

$$\mathbf{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha},$$

gdzie $\boldsymbol{\alpha} \in \mathbb{R}^{K+p}$, $K = B_1 + B_2 + \cdots + B_p$.

Dane funkcjonalne

Współczynniki $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top, \ldots, \boldsymbol{\alpha}_p^\top)$, $\boldsymbol{\alpha}_k \in \mathbb{R}^{B_k+1}$ szacujemy na bazie n niezależnych realizacji $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ procesu losowego $\mathbf{X}$ za pomocą metody najmniejszych kwadratów, w taki sposób aby zminimalizować funkcję

$$f(\mathbf{a}_k) = \sum_{j=1}^{J_k} (x_{kj} - \boldsymbol{\varphi}_k^\top(t_{kj})\mathbf{a}_k)^2,$$

gdzie $\mathbf{a}_k = (a_{k0}, a_{k1}, \ldots, a_{kB_k})^\top$, $k = 1, 2, \ldots, p$.

W rezultacie procesu transformacji n niezależnych realizacji procesu losowego $\mathbf{X}$ otrzymujemy dane funkcjonalne postaci:

$$\mathbf{x}_i(t) = \boldsymbol{\Phi}(t)\mathbf{a}_i,$$

gdzie $t \in I$, $i = 1, 2, \ldots, n$.

Kowariancja odległościowa

Funkcja charakterystyczna łącznego rozkładu procesu losowego $\boldsymbol{X} \in L_2^p(I)$ oraz wektora losowego $\boldsymbol{Y} \in \mathbb{R}^q$ ma postać

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

Przyjmijmy, że (Ramsay, Silverman (2005)) wektor funkcji wagowych $\boldsymbol{l}$ oraz proces losowy $\boldsymbol{X}$ należą do tej samej podprzestrzeni przestrzeni $L_2^p(I)$, tzn. funkcja $\boldsymbol{l}$ może być przedstawiona w postaci

$$\boldsymbol{l}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\lambda},$$

gdzie $\boldsymbol{\lambda} \in \mathbb{R}^{K+p}$.
Zatem

$$< \boldsymbol{l},\boldsymbol{X} >_p = \boldsymbol{\lambda}'\boldsymbol{\alpha}$$

oraz

$$f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{l},\boldsymbol{m}) = f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}),$$

gdzie $f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m})$ jest funkcją charakterystyczną łącznego rozkładu pary wektorów losowych $(\boldsymbol{\alpha},\boldsymbol{Y})$.

Kowariancja odległościowa

Bazując na pomyśle kowariancji odległościowej pomiędzy dwoma wektorami losowymi (Székely i inni (2007)), możemy zdefiniować funkcjonalną kowariancję odległościową pomiędzy procesem losowym $\boldsymbol{X}$ a wektorem losowym $\boldsymbol{Y}$ jako nieujemną liczbę $\mathrm{dCov}_{\boldsymbol{X},\boldsymbol{Y}}$ taką, że

$$\mathrm{dCov}^2_{\boldsymbol{X},\boldsymbol{Y}} = \frac{1}{C_{K+p}C_q}\int_{\mathbb{R}^{K+p+q}} \frac{|f_{\boldsymbol{\alpha},\boldsymbol{Y}}(\boldsymbol{\lambda},\boldsymbol{m}) - f_{\boldsymbol{\alpha}}(\boldsymbol{\lambda})f_{\boldsymbol{Y}}(\boldsymbol{m})|^2}{\|\boldsymbol{\lambda}\|_{K+p}^{K+p+1}\|\boldsymbol{m}\|_q^{q+1}}\,d\boldsymbol{\lambda}d\boldsymbol{m},$$

gdzie

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

oraz $f_{\boldsymbol{\alpha}}(\boldsymbol{\lambda})$, $f_{\boldsymbol{Y}}(\boldsymbol{m})$ są funkcjami charakterystycznymi rozkładów brzegowych.

Kowariancja odległościowa

Dla rozkładów o skończonych pierwszych momentach, funkcjonalna kowariancja odległościowa charakteryzuje niezależność w taki sposób, że $\mathrm{dCov}_{\boldsymbol{X},\boldsymbol{Y}} = 0$ wtedy i tylko wtedy, gdy $\boldsymbol{X}$ i $\boldsymbol{Y}$ są niezależne.

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

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

gdzie

$$A_{kl} = \|\tilde{\boldsymbol{a}}_k - \tilde{\boldsymbol{a}}_l\|_{K+p},\ B_{kl} = \|\tilde{\boldsymbol{y}}_k - \tilde{\boldsymbol{y}}_l\|_q,$$

$$\tilde{\boldsymbol{a}}_k = \boldsymbol{a} - \bar{\boldsymbol{a}},\ \tilde{\boldsymbol{y}}_k = \boldsymbol{y} - \bar{\boldsymbol{y}},\ k,l = 1,2,\ldots,n.$$

Wybór zmiennych

Zakładamy, że jedynie pewna liczba składowych procesu losowego $\boldsymbol{X}$ ma wpływ na wektor losowy Y. Wybieramy istotne składowe w taki sposób, aby funkcjonalna kowariancja odległościowa $\mathrm{dCov}_{\boldsymbol{X},\boldsymbol{Y}} = \mathrm{dCov}_{\boldsymbol{\alpha},\boldsymbol{Y}}$ była duża.

Prawdziwe jest następujące twierdzenie

Twierdzenie (Kong i inni (2015))

Niech $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{R}^p$ i $\boldsymbol{Y} \in \mathbb{R}^q$ oraz załóżmy, że wektor losowy $\boldsymbol{Z}$ jest niezależny od wektorów losowych $(\boldsymbol{X}, \boldsymbol{Y})$. Wtedy

$$\mathrm{dCov}_{(\boldsymbol{X},\boldsymbol{Z}),\boldsymbol{Y}} \leqslant \mathrm{dCov}_{\boldsymbol{X},\boldsymbol{Y}} .$$

Wybór zmiennych

Zastosowaliśmy twierdzenie Konga jako regułę stopu w procedurze wyboru zmiennych. Procedura składa się z następujących kroków:

1. Wyznaczamy brzegową kowariancję odległościową dla X_k, $k = 1, \ldots, p$ i $\boldsymbol{Y}$.
2. Porządkujemy zmienne w kolejności malejących kowariancji odległościowych. Oznaczmy uporządkowane zmienne jako $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$. Zaczynamy od $\boldsymbol{X}_S = \{X_{(1)}\}$.
3. Dla k od 2 do p, dodajemy $X_{(k)}$ do $\boldsymbol{X}_S$ jeżeli $\mathrm{dCov}_{\boldsymbol{X}_S, \boldsymbol{Y}}$ rośnie o więcej niż zadana wartość progowa (ε). Zatrzymujemy procedurę w przeciwnym razie.

Macierz jądrowa

Niech $\boldsymbol{Z} \in \mathbb{R}^p$ będzie wektorem losowym. Ponadto, niech

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

będzie rzeczywistą funkcją ciągłą, zwaną jądrem oraz $\boldsymbol{K} = (k_{ij})$, gdzie $k_{ij} = k(\boldsymbol{z}_i, \boldsymbol{z}_j)$ będzie macierzą jądrową, $i, j = 1, 2, \ldots, n$.

W dalszym ciągu wszystkie jądra będą jądrami gaussowskimi, tzn.

$$k(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp(-\lambda \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2), \ \lambda > 0.$$

Jądrowy współczynnik zgodności

Niech $\boldsymbol{X} \in L_2^p(I)$ będzie procesem losowym, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ jego niezależnymi realizacjami postaci:

$$\boldsymbol{x}_i(t) = \boldsymbol{\Phi}(t)\boldsymbol{a}_i, \;\; t \in I, \;\; i = 1, 2, \ldots, n.$$

Wtedy

$$k_{\boldsymbol{X}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\lambda_1 \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) = \exp(-\lambda_1 \|\boldsymbol{a}_i - \boldsymbol{a}_j\|^2) = k_{\boldsymbol{a}}(\boldsymbol{a}_i, \boldsymbol{a}_j),$$

dla $i, j = 1, 2, \ldots, n$.

Analogicznie, niech $\boldsymbol{Y} \in \mathbb{R}^q$ będzie wektorem losowym, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ jego niezależnymi realizacjami.
Wtedy

$$k_{\boldsymbol{Y}}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \exp(-\lambda_2 \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2),$$

dla $i, j = 1, 2, \ldots, n$.

Jądrowy współczynnik zgodności

Niech $\phi\colon L_2^p(I) \to \mathcal{H}$, gdzie $\mathcal{H}$ jest przestrzenią Hilberta, będzie przekształceniem takim, że dla dowolnych $\boldsymbol{x}, \boldsymbol{x}' \in L_2^p(I)$:

$$\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{H}} = k_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{x}').$$

Analogicznie, niech $\psi\colon \mathbb{R}^q \to \mathcal{G}$, gdzie $\mathcal{G}$ jest przestrzenią Hilberta, będzie przekształceniem takim, że dla dowolnych $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^q$:

$$\langle \psi(\boldsymbol{y}), \psi(\boldsymbol{y}') \rangle_{\mathcal{G}} = k_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{y}').$$

Operator kowariancji pomiędzy procesem losowym $\boldsymbol{X}$, a wektorem losowym $\boldsymbol{Y}$ definiujemy następująco:

$$\mathrm{Cov}_{\boldsymbol{X},\boldsymbol{Y}} = \mathrm{E}_{\boldsymbol{X},\boldsymbol{Y}}[(\phi(\boldsymbol{X}) - \mathrm{E}_{\boldsymbol{X}}(\phi(\boldsymbol{X}))) \otimes (\psi(\boldsymbol{Y}) - \mathrm{E}_{\boldsymbol{Y}}(\psi(\boldsymbol{Y})))].$$

Jądrowy współczynnik zgodności

Bazując na definicji współczynnika niezależności HSIC pomiędzy dwoma wektorami losowymi (Gretton i inni (2005)), możemy zdefiniować funkcjonalny współczynnik niezależności HSIC pomiędzy procesem losowym $\boldsymbol{X}$ a wektorem losowym $\boldsymbol{Y}$ jako

$$\mathrm{HSIC}_{\boldsymbol{X},\boldsymbol{Y}} = \| \mathrm{Cov}_{\boldsymbol{X},\boldsymbol{Y}} \|^2_{HS}.$$

Dla jądra gaussowskiego, $\mathrm{HSIC}_{\boldsymbol{X},\boldsymbol{Y}} = 0$ wtedy i tylko wtedy, gdy $\boldsymbol{X}$ i $\boldsymbol{Y}$ są niezależne.
Ponadto dla jądra gaussowskiego prawdziwe jest następujące twierdzenie:

Twierdzenie

Niech $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{R}^p$ i $\boldsymbol{Y} \in \mathbb{R}^q$ oraz załóżmy, że wektor losowy $\boldsymbol{Z}$ jest niezależny od wektorów losowych $(\boldsymbol{X}, \boldsymbol{Y})$. Wtedy

$$\mathrm{HSIC}_{(\boldsymbol{X},\boldsymbol{Z}),\boldsymbol{Y}} \leqslant \mathrm{HSIC}_{\boldsymbol{X},\boldsymbol{Y}}.$$

Jądrowy współczynnik zgodności

Estymator współczynnika zgodności HSIC ma postać:

$$\text{HSIC}_{\boldsymbol{X},\boldsymbol{Y}} = \frac{1}{n^2}\langle \tilde{\boldsymbol{K}}_{\boldsymbol{X}}, \tilde{\boldsymbol{K}}_{\boldsymbol{Y}} \rangle_F = \text{tr}(\tilde{\boldsymbol{K}}_{\boldsymbol{X}} \tilde{\boldsymbol{K}}_{\boldsymbol{Y}}),$$

gdzie $\tilde{\boldsymbol{K}}_{\boldsymbol{X}} = \boldsymbol{H}\boldsymbol{K}_{\boldsymbol{X}}\boldsymbol{H}$, $\tilde{\boldsymbol{K}}_{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{K}_{\boldsymbol{Y}}\boldsymbol{H}$ oraz $\boldsymbol{H} = (\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top)$.

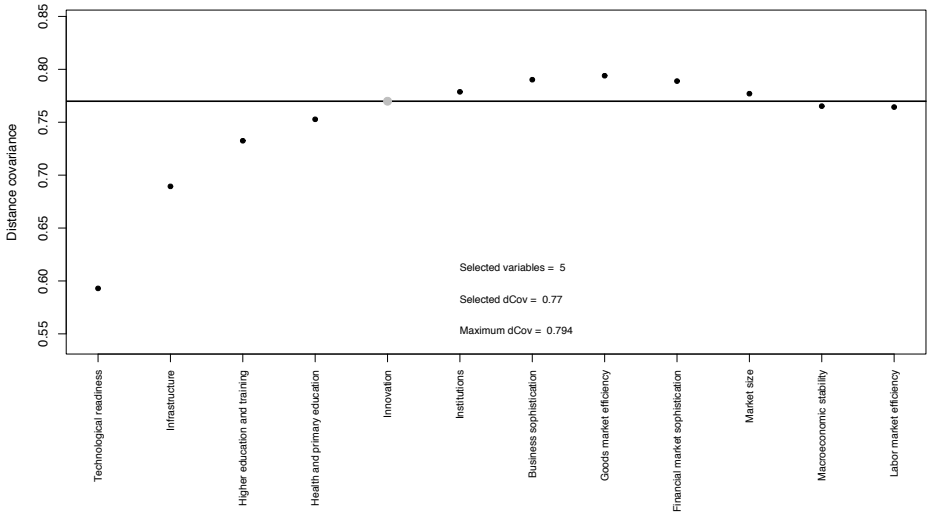
Zatem, w algorytmie wyboru zmiennych, zamiast kowariancją odległościową możemy posłużyć się współczynnikiem HSIC.

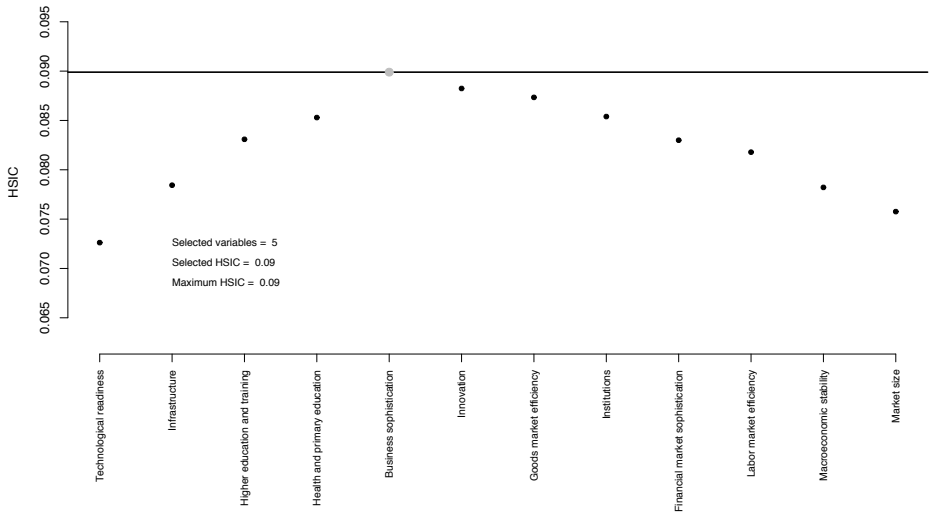
Przykład

Rozważmy dane pochodzące ze stron Światowego Forum Ekonomicznego (WEF). Dane dotyczące wskaźników socjo-ekonomicznych 115 krajów z lat 2008-2017. Eksperci WEF podzielili analizowane kraje na pięć grup.
W naszych rozważaniach wzięliśmy pod uwagę 12 zmiennych, tzw. filarów:

1. Institutions
2. Infrastructure
3. Macroeconomic stability
4. Health and primary education
5. Higher education and training
6. Goods market efficiency
7. Labor market efficiency
8. Financial market sophistication
9. Technological readiness
10. Market size
11. Business sophistication
12. Innovation

Przykład





Wybrane zmienne:

2. Infrastructure
4. Health and primary education
5. Higher education and training
9. Technological readiness
12. Innovation

2. Infrastructure
4. Health and primary education
5. Higher education and training
9. Technological readiness
11. Business sophistication

Ocena poprawności klasyfikacji

Baza Fouriera: $B_k = 5$.
Metoda oceny: LOO CV.
Reguła stopu: $\varepsilon = 0.05$.

Klasyfikator	Wybrane zmienne (5)	Wszystkie zmienne (12)
LDA	71,30%	66,09%
kNN ($k = 1, \ldots, 8$)	77,39%	71,30%
Naiwny Bayes (normal)	69,57%	65,22%
Naiwny Bayes (kernel)	67,83%	62,61%
Multinom	60,87%	56,52%

Bibliography


GÓRECKI, T., KRZYŚKO, M., WASZAK, Ł., WOŁYŃSKI, W. (2014): Methods of reducing dimension for functional data. *Statistics in Transition new series* 15, 231–242.


GRETTON, A., BOUSQUET, O., SMOLA, A., SCHÖLKOPF, B., (2005): Measuring statistical dependence with Hilbert-Schmidt norms. In: Algorithmic Learning Theory (S. Jain, H.U. Simon and E. Tomita, eds.). Lecture Notes in Computer Science 3734, 63–77. Springer, Berlin.


KONG, J., WANG, S., WAHBA G. (2015): Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine* 34, 1708–1720.


RAMSAY, J.O., SILVERMAN, B.W. (2005): *Functional Data Analysis*. Springer, New York.


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.