prof. dr hab. Jan Mielniczuk
Wydział MiNI PW
i Instytut Podstaw Informatyki PAN

8.01.2016

Recenzja rozprawy doktorskiej mgr Łukasza Waszaka

Wybrane wielowymiarowe metody statystyczne dla wielozmiennych danych funkcjonalnych

(wersja poprawiona)

W recenzji pierwszej przesłanej mi wersji rozprawy doktorskiej mgr Łukasza Waszaka usytuowałem jej tematykę w obrębie Analizy Danych Funkcjonalnych, z tą istotną różnicą w stosunku do klasycznego obszaru jej zainteresowań, że rozpatrywane obiekty są wektorami funkcji, a nie pojedyńczymi funkcjami. W pracy bada się w tym przypadku adaptacje trzech podstawowych metod statystycznej analizy wielowymiarowej: analizy składowych głównych PCA, liniowej analizy klasyfikacyjnej LDA i analizy kanonicznej.
Podstawowym tutaj narzędziem jest reprezentacja pojedyńczej losowej współrzędnej funkcyjnej jako skończonego rozwinięcia $x(t) = \sum_{b=0}^{B} c_b \phi_b(t)$, gdzie $(\phi_b)_0^B$ są funkcjami ortonormalnymi na pewnym odcinku I, a $(c_b)_0^B$ zmiennymi losowymi. Przy wykorzystaniu powyższej reprezentacji jednowymiarowej, cały wektor można zapisać jako liniową funkcję wszystkich współczynników rozwinięcia, przy czym dopuszcza się, że rzędy rozwinięć dla poszczególnych współrzędnych mogą być różne (str. 7 pracy).
Konsekwencją takiego podejścia dla analizy danych jest to, że mając p krzywych opisujących obiekt, całe postępowanie sprowadza się do zastąpienia k-tej krzywej B_k+1 współczynnikami jej rozwinięcia, zestawienia tych współczynników w jeden wektor o wymiarze $\sum_{k=1}^{p}(B_k+1)$ i zastosowania do tego wektora klasycznych metod analizy wielowymiarowej.
Okazuje się, że wyniki uzyskane dla trzech problemów rozpatrywanych w pracy są jednorodne w tym sensie, że składowe główne, zmienne dyskryminacyjne Fishera i zmienne kanoniczne dla problemu funkcyjnego są takie same jak w odpowiednim problemie zredukowanym. Mówią o tym główne twierdzenia pracy: twierdzenie 2.1, 3.1 i 4.1. Powodem tej niezmienniczości jest prosty fakt, mówiący, że:
Przekształcenie $R^p \ni \mathbf{w} \to \mathbf{\Phi}(t)\mathbf{w} \in L^2(I^p)$ zachowuje iloczyn skalarny:

$$< \mathbf{s}, \mathbf{w} >=< \mathbf{\Phi}(t)\mathbf{s}, \mathbf{\Phi}(t)\mathbf{w} >_1,$$

gdzie $\mathbf{\Phi}(t)$ *jest macierzą skonstruowaną z wartości funkcji ortonormalnych i zdefiniowaną w (1.3) w pracy,* $a < \mathbf{f}_1, \mathbf{f}_2 >_1 = \int_I \mathbf{f}_1^T(t)\mathbf{f}_2(t)\,dt$, $\mathbf{f}_1(t), \mathbf{f}_2(t) \in \mathbb{R}^p$ *dla każdego* $t \in I$.

Ocena pracy Analiza danych funkcjonalnych jest silnie rozwijającą się gałezią współczesnej statystyki; po cześci również dlatego, że daje alternatywne narzędzie do analizy wielowymiarowych szeregów czasowych. W takim podejściu równania strukturalne dotyczą reprezentacji trajektorii, a nie generacji procesu losowego. Rozprawa doktorska mgr Łukasza Waszaka dotyczy zatem istotnej tematyki. Główną zaletą pracy jest pokazanie, że trzy podstawowe problemy analizy danych funkcjonalnych (analiza składowych głównych, liniowa analiza klasyfikacyjna i analiza kanoniczna) mogą być *w taki sam sposób* sprowadzone do analogicznych problemów klasycznej analizy wielowymiarowej. Ta prosta, nie poczyniona wcześniej obserwacja, oparta na omówionym powyżej istnieniu izometrii między odpowiednimi przestrzeniami rozwiązań znacznie upraszcza procedury postępowania dla wielowymiarowych danych funkcjonalnych. Do innych nowych elementów w pracy zaliczam analizę zmiennych kanonicznych w przypadku wielowymiarowych zmiennych funkcjonalnych wymagającej adaptacji funkcji kary na przypadek wielowymiarowy oraz naturalną propozycję ważności oryginalnej współrzędnej do na przykład konstrukcji składowej głównej omówioną w (2.5).

W swojej poprzedniej recenzji wskazałem na kilka istotnych niedostatków pracy, w szczególności na (i) brak istotnych nowych własności badanych procedur i estymatorów (ii) rutynowość przedstawionych zastosowań, (iii) brak porównania z innymi podejściami estymacji wektora współczynników wektora $\mathbf{c}$ oraz (iv) brak analizy wpływu wyboru rzędu rozwinięcia B na własności rozwiązania. Nowa wersja w sposób przeważnie zadowalający odnosi się do tej krytyki. W szczególności na stronie 6 krótko omówiono inne metody estymacji wektora współczynników, wskazując na zadowalające działanie estymatora MNK, w rozdziale 2.3 przeprowadzono symulacyjną analizę wpływu wyboru rzędu rozwinięcia B przy użyciu kryteriów AIC, BIC i rozszerzonego kryterium BIC (eBIC). W rozdziale 5 rozbudowano analizę danych rzeczywistych dochodząc do kilku ciekawych wniosków praktycznych, m.in. w przykładzie pierwszym o zastępowalności grupy procesów 'ekologicznych' przez grupę procesów 'ekonomicznych' (str. 61).

Uwagi krytyczne

(i) Rozszerzona wersja pracy nie zawiera nowych teoretycznych własności zaproponowanych procedur, w szczególności na poziomie próbkowym.

(ii) Nie jest prawdziwe stwierdzenie (str. 11), że 'w dotychczasowych pracach badane obiekty charakteryzowane były tylko i wyłącznie za pomocą jednej cechy obserwowanej dynamicznie'; por. Ramsay i Silverman (2005), sekcja 8.5, czy cytowana i wykorzystywana przez autora w rozprawie doktorskiej praca Berrendo i inni (2011).
(iii) Jakkolwiek doceniam pomysłową metodę rangowania jakości różnych metod wyboru rzędu rozwinięcia przeprowadzoną w rozdziale 2.3, nie zmienia to faktu, że rozważane konstrukcje kryteriów informacyjnych nie miały na celu maksymalizacji części wyjaśnianej wariancji przez dwie pierwsze składowe i taki komentarz powinien znaleźć się w pracy.
(iv) W 2.3 nie znalazłem istotnej informacji, ile wynosił maksymalny rozpatrywany rząd rozwinięcia B_{max}.
(v) Choć związek między eBIC a BIC na dole str. 19 podany jest prawidłowo, postać BIC jest błędna: mianownik w członie lnJ/J jest zbędny.
(vi) Warunki ograniczające w konstrukcji zmiennych klasyfikacyjnych (ich nieskorelowanie) są podane błędnie na str. 29 (dwa miejsca);
(vii) W pracy nadal sporo jest literówek i drobnych błędów, częstokroć pojawiających się w wersji pierwotnej, w szczególności: brak dwóch transpozycji we wzorze na r_{YY} (str. 40); oznaczenie $\mathrm{Var}(U^{(N)})$ jest mylące, gdyż nie odpowiada wariancji zmiennej $U^{(N)}$ (str. 39); B_1, B_2 oznaczają rzędy modeli, więc nie powinny występować w definicjach na str. 19; 'nastomiast z równowści' (str. 35); itp.
(viii) Policzenie kowariancji procesów w dowodzie twierdzenia 4.1 jest zbędne.

Mimo tych niedostatków uważam, że mgr Łukasz Waszak wykonał sporą pracę poprawienia oryginalnej rozprawy doktorskiej i skutek tych działań jest w większości przypadków zadowalający.

Konkluzja W mojej opinii rozprawa doktorska mgr Łukasza Waszaka *Wybrane wielowymiarowe metody statystyczne dla wielozmiennych danych funkcjonalnych* spełnia ustawowe wymagania stawiane rozprawom doktorskim w dziedzinie nauk matematycznych, dyscyplina matematyka. Wnoszę o dopuszczenie mgr Ł. Waszaka do dalszych etapów przewodu doktorskiego.

Jan Mielniczuk