STRESZCZENIE - ROZPRAWA DOKTORSKA 'WYBRANE WIELOWYMIAROWE METODY STATYSTYCZNE DLA WIELOZMIENNYCH DANYCH FUNKCJONALNYCH'

ŁUKASZ WASZAK

W klasycznych metodach statystycznych obiekty podlegające badaniu charakteryzowane są za pomocą cech obserwowanych w ustalonym momencie czasu. W prezentowanej rozprawie zakładać będziemy, że obiekty charakteryzowane są za pomocą zmiennych funkcjonalnych. Czym są zmienne funkcjonalne? Zmienna funkcjonalna X jest zmienną losową przyjmującą wartości w pewnej przestrzeni funkcjonalnej E. Zbiór danych funkcjonalnych jest próbą $\{X_1, \ldots, X_n\}$ (oznaczaną również przez $\{X_1(t), \ldots, X_n(t)\}$, jeśli jest to wygodne) pobraną z rozkładu zmiennej funkcjonalnej X. Termin dane funkcjonalne pojawił się po raz pierwszy w pracy Ramsaya i Dalzella (1991). W dalszym ciągu zakładać będziemy, że E jest przestrzenią Hilberta wszystkich funkcji całkowalnych z kwadratem na pewnym przedziale $[a, b]$, tj przestrzenią $L_2([a, b])$.

W tym przypadku dane funkcjonalne mogą być reprezentowane w postaci

$$X(t) = \sum_{b=0}^{\infty} c_b \varphi_b(t),$$

gdzie $\varphi_b(t)$ są znanymi, ustalonymi funkcjami ortonormalnymi lub inaczej elementami ortonormalnej bazy $\{\varphi_0, \varphi_1, ...\}$. Zauważmy, że reprezentacja funkcji za pomocą nieskończonego szeregu ortonormalnego wymaga znajomości nieskończonej liczby współczynników c_b. Niestety nikt z nas nie potrafi radzić sobie z nieskończoną liczbą współczynników. W związku z tym do aproksymacji funkcji $X(t)$ wykorzystuje się ucięty (skończony) szereg ortonormalny, zwany inaczej sumą częściową, postaci

$$X_B(t) := \sum_{b=0}^{B} c_b \varphi_b(t) = \boldsymbol{c}'\boldsymbol{\varphi}(t) = \boldsymbol{\varphi}'(t)\boldsymbol{c}.$$

Parametr B, będący liczbą naturalną, nazywa się punktem ucięcia. Zazwyczaj tylko niewielka liczba współczynników rozwinięcia jest istotna, a wszystkie pozostałe są mało znaczące. Prowadzi to do istotnej redukcji danych.

Z grubsza rzecz biorąc, główny problem statystyczny polega na optymalnym wyborze punktu ucięcia B oraz optymalnym oszacowaniu współczynników c_b. Problemom tym jest poświęcony Rozdział 1 i 2 tej rozprawy.

W tym miejscu można postawić naturalne pytanie: czy w rzeczywistości istnieją dane funkcjonalne? Pytanie to ma istotne znaczenie, gdyż w praktyce wartości obserwowanego procesu losowego $X(t)$ są zawsze rejestrowane w dyskretnych momentach czasu $t_1, t_2, ..., t_J$, rzadziej lub gęściej rozmieszczonych w przedziale zmienności argumentu t. Tak więc ostatecznie mamy zawsze do czynienia z szeregiem czasowym $\{x(t_1), x(t_2), ..., x(t_J)\}$ lub inaczej z wysokowymiarowym wektorem obserwacji. Istnieją jednakże liczne powody, by szeregi takie modelować jako elementy przestrzeni funkcjonalnej, ponieważ dane funkcjonalne mają wiele zalet w porównaniu z innymi sposobami reprezentowania szeregów czasowych.

Po pierwsze, łatwo radzą sobie z problemem brakujących obserwacji, nieuniknionym problemem w wielu dziedzinach badań. Niestety, większość metod analizy danych wymaga kompletnych szeregów czasowych. Jednym z rozwiązań jest po prostu usunięcie szeregu czasowego mającego brakujące wartości ze zbioru danych, ale działanie takie może prowadzić, i na ogół prowadzi, do utraty informacji. Inna możliwość, to posłużenie się jedną z wielu metod statystycznych predykcji brakujących danych, ale wówczas wyniki będą zależały od metody interpolacji. W przeciwieństwie do tego typu działań, w przypadku danych funkcjonalnych, problem brakujących obserwacji jest rozwiązany poprzez wyrażenie szeregów czasowych w postaci zbioru krzywych ciągłych.

Po drugie, dane funkcjonalne w sposób naturalny zachowują strukturę obserwacji, tj. zachowują zależność czasową obserwacji i biorą pod uwagę informację o każdym pomiarze.

Po trzecie, momenty obserwacji nie muszą być równomiernie rozmieszczone w poszczególnych szeregach czasowych.

Po czwarte, dane funkcjonalne unikają „przekleństwa" nadmiernej wymiarowości. Gdy całkowita liczba punktów czasowych, w których dokonuje się obserwacji przekracza liczbę rozpatrywanych szeregów czasowych, większość metod statystycznych nie daje zadowalających wyników ze względu na przeparametryzowanie. Aby uniknąć tego problemu, najczęściej stosuje się techniki redukcji wymiaru, takie jak analiza składowych głównych. Jednakże w tym przypadku pewne informacje o strukturze przestrzennej i czasowej danych mogą zostać utracone. W przypadku danych funkcjonalnych można uniknąć tego problemu, ponieważ szeregi czasowe zostają zastąpione zbiorem krzywych ciągłych

niezależnych od całkowitej liczby punktów czasowych, w których dokonuje się obserwacji.

Chociaż prace dotyczące danych funkcjonalnych pojawiały się już wcześniej, to za symboliczny moment startu metod statystycznych dla danych funkcjonalnych należy przyjąć ukazanie się monografii Ramsaya i Silvermana (1997), monografii skierowanej do szerokiego kręgu odbiorców i z przewagą zagadnień praktycznych nad teorią. Uzupełnieniem tej pięknej monografii o dalsze aspekty praktyczne była książka Ramsaya i Silvermana (2002). W roku 2005 ukazało się wydanie drugie monografii z roku 1997. Wydanie to pociągnęło za sobą prawdziwy wysyp prac związanych z funkcjonalną analizą danych. Kolejnymi znaczącymi pozycjami w dziedzinie analizowania danych funkcjonalnych były książki: Clarkson i inni (2005), Ferraty i Vieu (2006) – książka o nastawieniu teoretycznym, Bosq i Blanke (2007), Dabo-Niang i Ferraty (2008), Ramsay i inni (2009) oraz Ferraty i Romain (Eds) (2011). Najnowszą pozycją uzupełniającą literaturę dotyczącą danych funkcjonalnych jest książka Horvatha i Kokoszki (2012). Zawiera ona wyważoną mieszankę aspektów teoretycznych i praktycznych funkcjonalnej analizy danych.

Wśród prac przeglądowych dotyczących tej tematyki wymienić należy następujące prace: Rice (2004), Muller (2005), Gonzalez-Manteiga i Vieu (2011) - zawiera ona bogatą bibliografię, Delsol i inni (2011), Febrero-Bande i Oviedo de la Fuente (2012) oraz Cuevas (2014).

Wśród wielu metod statystycznych skonstruowanych dla danych funkcjonalnych poczesne miejsce zajmują trzy metody określane wspólnym mianem - metody redukcji wymiaru. Są to: analiza składowych głównych, analiza zmiennych dyskryminacyjnych oraz analiza korelacji i zmiennych kanonicznych. Klasyczne wersje tych metod zakładają, że rozpatrywane obiekty charakteryzowane są wieloma cechami. Tymczasem, w przypadku danych funkcjonalnych, dotychczas istniejące prace przyjmują, że obiekty charakteryzowane są za pomocą jednowymiarowych danych funkcjonalnych. Jest tu widoczna rozbieżność między założeniami w przypadku metod klasycznych i metod dla danych funkcjonalnych. W celu usunięcia tej rozbieżności, w prezentowanej rozprawie, skonstruowane zostały metody redukcji wymiaru dla wielowymiarowych danych funkcjonalnych. Metody te opisane są w Rozdziałach 2-4. Rozdział 5 zawiera konkretne przykłady stosowania tych metod.