Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Matematyczny
specjalność: analiza danych

Marta Morawiec

Test normalności i niezależności

Praca licencjacka
napisana pod kierunkiem
dr Wiktora Ejsmonta

Wrocław 2020

Spis treści

Streszczenie

W tej pracy zaproponujemy wspólny test niezależności i normalności, który może być zaimplementowany w dowolnym wymiarze. Bazuje on na pomyśle charakteryzacji rozkładu normalnego według dr Ejsmonta[4]. Test ten używa całki z kwadratu modułu różnicy pomiędzy produktem funkcji charakterystycznej próby i pewnej stałej. Szczególną uwagę przywiązujemy do przypadku dwóch jednowymiarowych wektorów losowych, kiedy statystyka testowa może być wyrażana funkcjami Bassela. Przeprowadzamy dla tego wypadku symulacje wartości krytycznych oraz mocy testu przy różnych hipotezach alternatywnych w środowisku R.

Rozdział 1

Wstęp

Jednym z najczęściej występujących i najważniejszym problemem w statystyce jest testowanie niezależności pomiędzy dwoma lub większą ilością wektorów losowych. Tradycyjne podejście do tego problemu bazuje na korelacji parametrów Pearsona. Metoda ta ma jednak wady, a największą z nich jest jej brak wrażliwości na obserwacje odstające. Skłania to nas do poszukiwania innych metod, które będą nieparametryczne. Istnieją procedury testowania, które opierają się na statystykach liniowych. Są to na przykład procedura Savage'a, Spearmana i van den Waerdena. W tej pracy opisujemy metodę testowania niezależności i normalności, która bazuje na różnicy pomiędzy stałą $\exp(-\frac{1}{2})$ a funkcją empiryczną, którą opiszemy w dalszej części pracy. W statystyce wiele badań zajmuje się badaniem zależności pomiędzy wektorami $X = (X_1, \cdots, X_m)$ oraz $Y = (Y_1, \cdots, Y_n)$. Podstawowym pytaniem na które szukamy wówczas odpowiedzi jest to czy elementy wektorów X_i i Y_j są niezależne i mają ten sam rozkład normalny. Możemy założyć wielowymiarowy rozkład normalny złączenia wektorów X i Y. Wiemy, że jeśli wektor losowy ma wielowymiarowy rozkład normalny, to wszystkie jego składniki, które nie są ze sobą skorelowane są niezależne. Z tego wynika, że jeśli każde dwa lub większa ilość elementów takiego wektora są parami niezależne to są niezależne. Wówczas dla wektora $(X_1, ..., X_m, Y_1, ..., Y_n)$ problem sprowadza się do badania hipotezy zerowej H_0 : parametry korelacji są równe 0.
W rozdziale 2 omówimy krótko najważniejsze podstawy teoretyczne do wyprowadzenia statystyki testowej. Skupimy się na teorii związanej z funkcjami Bessela. W następnym rozdziale zajmiemy się wyprowadzeniem i opisaniem postaci statystyki testowej, którą w tej pracy proponujemy. W pierwszej kolejności w wersji dla wielowymiarowych wektorów, a następnie jej postacią w wypadku dwóch wektorów jednowymiarowych. W rozdziale 4 przeprowadzimy symulacje wartości krytycznych oraz mocy testu.
Poniższa praca opiera się na pracy oraz pomyśle dr Wiktora Ejsmonta[1],

któremu bardzo dziękuję za jej udostępnienie i pomoc w przygotowaniu tej pracy.

Notacja

W poniższej pracy będziemy używać następującej notacji. Iloczyn skalarny wektorów $t, s \in \mathbb{R}^p$ jest oznaczany przez $\langle t, s \rangle$, a norma euklidesowa (L_2) z t to $||t|| = \sqrt{\langle t, t \rangle}$. Mamy również wektory losowe $X := (X_1, ..., X_m) \in \mathbb{R}^m$, $Y := (Y_1, ..., Y_n) \in \mathbb{R}^n$, gdzie n, m są liczbami całkowitymi dodatnimi. Funkcje charakterystyczne X i Y oznaczamy kolejno ϕ_X i ϕ_Y. Dla ułatwienia oznaczeń wprowadzamy również $[n] = \{1, ..., n\}$ oraz $[m] = \{1, ..., m\}$. Symbol $\mathbf{0}$ oznacza wektor samych zer. Dla funkcji o wartościach zespolonych $f(\cdot)$ ich sprzężenie zespolone oznaczamy przez $\overline{f}$, a $|f|^2 = f\overline{f}$.

Rozdział 2

Teoretyczne podstawy

2.1 Funkcje Bessela

W wypadku, który będziemy dokładnie opisywać w 3.3 statystyka testowa ma postać, w której występuje funkcja Bessela. Funkcje te nazywane są również funkcjami cylindrycznymi. Definiuje się je jako rozwiązania $y(x)$ równania różniczkowego drugiego stopnia nazywanego równaniem Bessela:

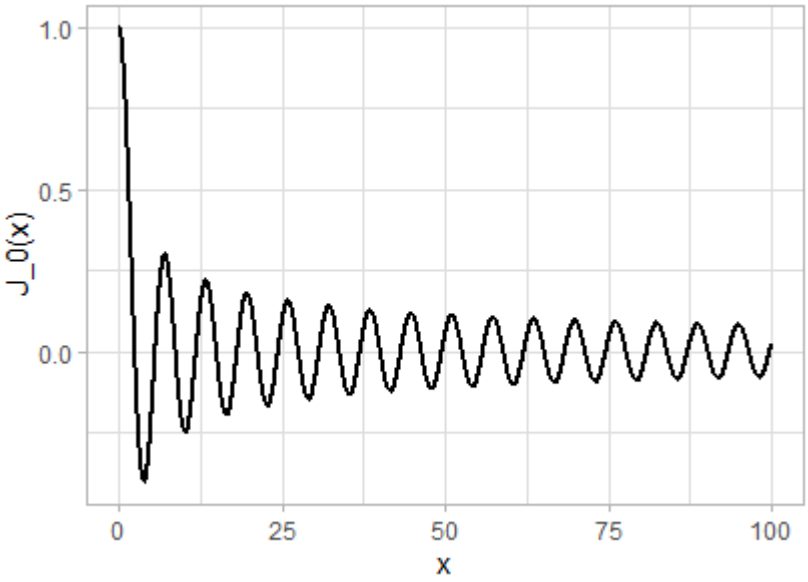
$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - \alpha^2)y = 0.$$

Człowiek, od którego wzięły nazwę, Friedrich Wilhelm Bessel, wyprowadził je około roku 1817 podczas prowadzenia badań nad rozwiązaniem jednego z równań ruchu planet Keplera. Mają one bardzo szerokie zastosowanie w fizyce, między innymi w przepływie ciepła lub energii w cylindrze stałym, dyfrakcji światła czy odkształcaniach ciał elastycznych.
Ogólne rozwiązanie równania Bessela daje funkcje Bessela pierwszego i drugiego rodzaju:

$$y = AJ_\alpha(x) + BY_\alpha(x),$$

gdzie A i B są ustalonymi stałymi, J_α to funkcja Bessela pierwszego rodzaju, a Y_α - drugiego rodzaju. Bardzo często funkcje te są przedstawiane z całkowitymi wartościami parametru α, podczas gdy może on przyjmować wartości z całego zakresu liczb rzeczywistych. My omówimy trochę dokładniej tylko pierwszy rodzaj, którego będziemy używać.
Funkcja Bessela pierwszego rodzaju jest skończona w $x = 0$ dla wszystkich rzeczywistych wartości parametru α. Możemy ją określić używając nieskończonego rozszerzenia szeregu mocy w następujący sposób:

$$J_\alpha(x) = \sum_{k=0}^{\infty} \frac{(-1)^k(x/2)^{\alpha+2k}}{k!(\alpha+k)!}.$$



Rysunek 2.1: Wykres funkcji Bessela $J_0(x)$ na odcinku $x \in [0, 100]$

Dla parametru $\alpha = 0$, który najbardziej nas interesuje wzór ten przyjmuje formę

$$J_0(x) = \sum_{k=0}^{\infty} \frac{(-1)^k (x/2)^{2k}}{(k!)^2},$$

a wykres tej funkcji przedstawia rysunek 2.1. Możemy zauważyć, że przypomina on tłumioną cosinusoidę.
Funkcje Bessela pierwszego rodzaju możemy również przedstawić w postaci całki. Wówczas dla parametru $\alpha \in \mathbb{Z}$ wygląda ona tak:

$$J_\alpha(x) = \frac{1}{\pi} \int_0^\pi \cos(\alpha\theta - x \sin\theta) d\theta = \frac{1}{\pi} \int_0^\pi \cos(x \sin\theta - \alpha\theta) d\theta.$$

Istnieją różne przybliżenia tych funkcji ułatwiające pracę z nimi oraz szereg użytecznych własności. W wyznaczaniu naszej statystyki testowej będziemy używać tylko funkcji Bessela pierwszego rodzaju z parametrem $\alpha = 0$, której wykres widzimy wyżej. Pierwiastki $J_0(x) = 0$ są liczbami dodatnimi i jest ich nieskończenie wiele. Przy $n \to \infty$ różnica pomiędzy dwoma kolejnymi pierwiastkami dąży do liczby π. Rozwiązania te możemy przybliżać aproksymacją Stokes'a dla dużych wartości n.

2.2 Wprowadzenie do statystyki

Podstawy do badania opisywanego w tej pracy testu są zawarte w pracy doktora Ejsmonta [4]. W pracy tej autor formułuje nową charakteryzację rozkładu normalnego, która jest wzorowana na charakteryzacji Cooka[9] oraz Kagana i Shalaevskiego[10] oraz jest dana przez pewną inwersję niecentralnego rozkładu χ^2. Zakłada ona dodatkowe istnienie wszystkich momentów. Zostało pokazane, że jeśli wektory $X = (X_1, ..., X_m)$ oraz $Y = (Y_1, ..., Y_n)$ są niezależne oraz mają wszystkie momenty i rozkład $\sum_{i=1}^m X_i a_i + A + \sum_{j=1}^n Y_j b_j + B$ zależy tylko od $\sum_{i=1}^m a_i^2 + \sum_{j=1}^n b_j^2$ to $X_1, ..., X_m, Y_1, ..., Y_n$ są niezależne i pochodzą z jednego rozkładu normalnego. Nasza konstrukcja testu opiera się na zmodyfikowanej wersji tego twierdzenia. Mianowicie pomijamy założenie o istnieniu wszystkich momentów zmiennych losowych. Ze względu na to przeprowadzimy poniżej dowód zmodyfikowanego twierdzenia.

Twierdzenie 2.2.1 *Niech $(X_1, ..., X_m, A)$ i $(Y_1, ..., Y_n, B)$ będą niezależnymi wektorami losowymi, w których X_i i Y_j są niezdegenerowane dla $i \in [m]$, $j \in [m]$ i niech statystyka*

$$\langle a, X\rangle + \langle b, Y\rangle + A + B = \sum_{i=1}^m X_i a_i + A + \sum_{j=1}^n Y_j b_j + B,$$

ma rozkład zależny tylko od $||a||^2 + ||b||^2$, gdzie $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$. Wówczas zmienne losowe $X_1, ..., X_m, Y_1, ..., Y_n$ są niezależne i mają ten sam rozkład normalny o średniej $\mu = 0$.

DOWÓD

Dowód ten opieramy na analizie funkcji charakterystycznej rozkładu normalnego. Niech $\tilde{a} = \frac{a}{\sqrt{||a||^2+||b||^2}}$ i $\tilde{b} = \frac{b}{\sqrt{||a||^2+||b||^2}}$. Wtedy dla $t \in \mathbb{R}$ otrzymujemy

$$\mathbb{E}\exp\left(\frac{i(\langle a, X\rangle + \langle b, Y\rangle + A + B)t}{\sqrt{||a||^2 + ||b||^2}}\right) =$$

$$\mathbb{E}\exp\left(i(\langle \tilde{a}, X\rangle + \langle \tilde{b}, Y\rangle)t + \frac{i(A+B)t}{\sqrt{||a||^2 + ||b||^2}}\right).$$

Z założenia wiemy, że ta wartość oczekiwana nie zależy od $||\tilde{a}||^2 + ||\tilde{b}||^2$. Wówczas dostajemy wartość oczekiwaną:

$$\mathbb{E}\exp\left(i(\langle \tilde{a}, X\rangle + \langle \tilde{b}, Y\rangle)t + \frac{i(A+B)t}{\sqrt{||a||^2 + ||b||^2}}\right),$$

która przy

$$||a||^2 + ||b||^2 \to +\infty$$

dąży do

$$\mathbb{E}\exp\left(i(\langle\tilde{a}, X\rangle + \langle\tilde{b}, Y\rangle)t\right))$$

i nie zależy od $||\tilde{a}||^2 + ||\tilde{b}||^2$. W szczególności otrzymujemy zależność rozkładu statystyki

$$\langle a, X\rangle + \langle b, Y\rangle = (\langle\tilde{a}, X\rangle + \langle\tilde{b}, Y\rangle)\sqrt{||a||^2 + ||b||^2}$$

tylko od $||a||^2 + ||b||^2$.
Niech funkcja h będzie postaci $h(||a||^2 + ||b||^2) = \mathbb{E}\exp\left(i(\langle a, X\rangle + \langle b, Y\rangle)\right)$. Korzystając z niezależności X i Y możemy to zapisać jako:

$$(1) \qquad h(||a||^2 + ||b||^2) = \mathbb{E}\exp\left(i(\langle a, X\rangle)\right)\mathbb{E}\exp\left(i(\langle b, Y\rangle)\right).$$

Przekształcając (1) najpierw dla $a = \mathbf{0} \in \mathbb{R}^m$ a następnie dla $b = \mathbf{0} \in \mathbb{R}^n$ mamy kolejno

$$h(||b||^2) = \mathbb{E}\exp\left(i\langle b, Y\rangle\right) \quad \text{oraz} \quad h(||a||^2) = \mathbb{E}\exp\left(i\langle a, X\rangle\right).$$

Podstawiając to do równości (1) dostajemy

$$h(||a||^2 + ||b||^2) = h(||a||^2)h(||b||^2).$$

Zauważamy, że funkcja h jest ciągła stąd korzystając z multiplikatywnego równania funkcjonalnego Cauchy'ego dostajemy

$$h(||a||^2 + ||b||^2) = \exp\left(c(||a||^2 + ||b||^2)\right).$$

Podstawiając $a = (a_1, 0, 0, ..., 0)$ oraz $b = \mathbf{0}$ w tym równaniu dostajemy postać $\mathbb{E}\exp\left(iX_1a_1\right) = \exp\left(ca_1^2\right)$ a to oznacza, że X_1 ma rozkład normalny z średnią równą zero. Powtarzając to rozumowanie dla innych zmiennych losowych widzimy, że X_i i Y_j mają taki sam rozkład normalny z zerową średnią. Niezależność zmiennych losowych $X_1, ..., X_m$ wynika z obserwacji, że dla wszystkich $a = (a_1, ..., a_m) \in \mathbb{R}^m$ mamy

$$\mathbb{E}\exp\left(i\langle a, X\rangle\right) = \exp\left(c\sum_{j=1}^{m} a_j^2\right) = \mathbb{E}\exp\left(iX_1a_1\right)...\mathbb{E}\exp\left(iX_ma_m\right).$$

□

Rozdział 3

Statystyka testowa

3.1 Propozycja dla statystyki

Konstrukcję nowego testu opieramy bezpośrednio na poniższej propozycji 3.1.1. Jest ona szczególnym przypadkiem twierdzenia 2.2.1 dla $A = B = 0$.

Propozycja 3.1.1 *Niech X i Y będą niezależnymi wektorami losowymi, takimi że X_i i Y_j nie są zdegenerowane oraz $\mathbb{E}(X_i^2) = 1$, $\mathbb{E}(Y_j^2) = 1$ dla wszystkich $i \in [m],\ j \in [n]$. Wtedy następujące stwierdzenia są równoważne:*

(i) statystyka $\langle a, X\rangle + \langle b, Y\rangle$ ma rozkład, który nie zależy od $(a_1, ..., a_m, b_1, ..., b_n)$ dla każdego a i b spełniających warunek $||a||^2 + ||b||^2 = 1$;

(ii) zmienne losowe $X_1, ..., X_m, Y_1, ..., Y_n$ są niezależne i mają ten sam rozkład $N(0, 1)$.

DOWÓD
(i) $\Rightarrow$ (ii)
Udowodnimy najpierw, że z stwierdzenia *(i)* wynika stwierdzenie *(ii)*. Z pierwszego stwierdzenia wiemy, że rozkład

$$\langle a, X\rangle + \langle b, Y\rangle = \sqrt{||a||^2 + ||b||^2}\ \frac{\langle a, X\rangle + \langle b, Y\rangle}{\sqrt{||a||^2 + ||b||^2}}$$

zależy tylko od $||a||^2 + ||b||^2$. Z twierdzenia 2.2.1 wynika, że wówczas X_i i Y_j są niezależne i mają ten sam rozkład normalny $N(0, 1)$. Jest tak, ponieważ zakładamy, że $\mathbb{E}(X_i^2) = \mathbb{E}(Y_j^2) = 1$.
(ii) $\Rightarrow$ (i)
Obiczamy funkcję charakterystyczną

$$\mathbb{E} \exp\left(i\langle a, X\rangle + i\langle b, Y\rangle\right) = \exp \frac{-(||a||^2 + ||b||^2)}{2}.$$

Widzimy, że zależy ona tylko od $||a||^2 + ||b||^2$ zatem stwierdzenie *(i)* jest prawdziwe.

□

3.2 Postać nowej statystyki

Do konstrukcji omawianego testu będziemy używać odległości. Istnieje wiele typów odległości w teorii testowania hipotez statystycznych, które są definiowane pomiędzy obiektami statystycznymi. Jedną z najbardziej znanych i najczęściej używanych jest odległość L_2. Przyjmijmy, że F jest skumulowaną funkcją rozkładu zmiennych losowych a F_n jest funkcją empiryczną. Odległość L_2 między nimi możemy zapisać jako $\int_{-\infty}^{\infty}(F_n(x)-F(x))^2dx$, czyli całkę z kwadratu różnicy pomiędzy nimi. Do testowania rozkładu wielowymiarowego normalnego stosujemy odległość pomiędzy empiryczną a teoretyczną funkcją charakterystyczną. Załóżmy, że $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^n$ są losowymi wektorami o wartościach rzeczywistych i funkcjach charakterystycznych ϕ_X i ϕ_Y. Wtedy do sprawdzenia niezależności możemy użyć następującej odległości

$$\int\limits_{\mathbb{R}^{m+n}} |\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 w(t,s)dtds$$

gdzie $w(t,s)$ jest arbitralną dodatnią funkcją wag, dla której powyższa całka istnieje. Proponowany przez nas test bazuje na odległości pomiędzy funkcją charakterystyczną a pewną stałą i został zainspirowany artykułami [3, 5, 7, 8]. Stwierdzenie *(i)* z propozycji 3.1.1 wprost mówi nam, że dostajemy stwierdzenie *(ii)* gdy rozkład statystyki $\langle a, X\rangle + \langle b, Y\rangle$ jest stały na $n+m$ wymiarowej sferze o promieniu 1. Ten warunek możemy zapisać za pomocą funkcji charakterystycznych. Otrzymujemy stwierdzenie *(ii)* wtedy i tylko wtedy gdy funkcja $\mathbb{E}(\exp(i\langle a, X\rangle + i\langle b, Y\rangle)) = \phi_X(a)\phi_Y(b)$ jest stała na sferze jednostkowej dla $||a||^2 + ||b||^2 = 1$ gdzie $a \in \mathbb{R}^m$ i $b \in \mathbb{R}^n$. Z dowodu propozycji 3.1.1 wiemy również, że ta stała funkcja musi być równa $\exp(-\frac{1}{2})$. Otrzymujemy $\phi_X(a)\phi_Y(b) - \exp(-\frac{1}{2}) = 0$ dla wszystkich $||a||^2 + ||b||^2 = 1$. Równoważnie

$$(2) \qquad \int\limits_{S_{n+m}} |\phi_X(a)\phi_Y(b) - \exp(-\frac{1}{2})|^2 dS_{n+m} = 0,$$

gdzie całka w (2) jest całką powierzchniową na $S_{n+m} = \{t \,\epsilon\, \mathbb{R}^{n+m} : ||t|| = 1\}$. Skończoność tej całki wynika bezpośrednio z tego, że $|\phi_X(a)\phi_Y(b)| \leq 1$ oraz $\exp(-\frac{1}{2}) < 1$. Widzimy zatem, że $\int_{S_{n+m}} |\phi_X(a)\phi_Y(b) - \exp(-\frac{1}{2})|^2 dS_{n+m} \leq$

$(1-\exp(-\frac{1}{2}))^2|S_{n+m}|$.
Próba z rozkładu X z $\mathbb{R}^m$ (w przypadku Y z $\mathbb{R}^n$) jest oznaczana przez $N \times m$ (lub analogicznie $N \times n$) wymiarową macierz $\mathbf{X}$ ($\mathbf{Y}$), gdzie wektory losowe są kolumnami. Dane możemy zapisać jak poniżej:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,m} \end{bmatrix} \quad \text{oraz} \quad \mathbf{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,n} \\ y_{2,1} & y_{2,2} & \dots & y_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N,1} & y_{N,2} & \dots & y_{N,n} \end{bmatrix}.$$

Zakładamy również, że $\mathbf{X}$ oraz $\mathbf{Y}$ są niezależne. Chcemy testować hipotezę:

$$H_0^{(m,n)} : \text{wszystkie kolumny } \mathbf{X} \text{ i } \mathbf{Y} \text{ są niezależne i mają rozkład normalny}$$
$$\text{vs}$$
$$H_1^{(m,n)} : \ H_0^{(m,n)} \text{ nie jest prawdziwa.}$$

Przez $\widetilde{\mathbf{X}}$ oraz $\widetilde{\mathbf{Y}}$ oznaczamy macierze otrzymane poprzez normalizację kolumn macierzy $\mathbf{X}$ oraz $\mathbf{Y}$. Musimy tego dokonać, ponieważ test powinien być niezależny ze względu na parametry przesunięcia i skali. Niech $\Phi_{\widetilde{\mathbf{X}}}(a)$, $\Phi_{\widetilde{\mathbf{Y}}}(b)$ będą empirycznymi funkcjami charakterystycznymi macierzy $\widetilde{\mathbf{X}}$ oraz $\widetilde{\mathbf{Y}}$. Definiujemy je w następujący sposób:

$$\Phi_{\widetilde{\mathbf{X}}}(a) = \frac{1}{N}\sum_{k=1}^{N} \exp(i\langle a, \widetilde{\mathbf{X}}_k\rangle) \quad \text{i} \quad \Phi_{\widetilde{\mathbf{Y}}}(b) = \frac{1}{N}\sum_{k=1}^{N} \exp(i\langle b, \widetilde{\mathbf{Y}}_k\rangle),$$

gdzie $\widetilde{\mathbf{X}}_k$ i $\widetilde{\mathbf{Y}}_k$ są k-tym wierszem macierzy $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{Y}}$. Obserwujemy, że po normalizacji kolumn macierze $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{Y}}$ spełniają założenia propozycji 3.1.1 i w końcu jesteśmy gotowi do zdefiniowania naszej statystyki. Proponujemy statystykę, która bazuje na całce oznaczonej przez (2), do której podstawiamy odpowiednie estymatory funkcji charakterystycznych i mnożymy przez liczebność próby. Otrzymujemy:

$$(3) \quad T_{m,n} := N \int\limits_{S_{n+m}} |\Phi_{\widetilde{\mathbf{X}}}(a)\Phi_{\widetilde{\mathbf{Y}}}(b) - \exp(-\frac{1}{2})|^2 dS_{n+m}.$$

Interesuje nas test prawostronny. Dla podanej wyżej konstrukcji odrzucamy hipotezę zerową $H_0^{(m,n)}$ dla dużych wartości statystyki $T_{m,n}$.

3.3 Postać statystyki $T_{1,1}$

Pokażemy teraz, że dla $n = m = 1$ istnieje dość prosta postać statystyki $T_{m,n}$, którą możemy obliczyć. Reprezentację statystyki w tym konkretnym wypadku otrzymujemy wykonując całkowanie z (3). Zakładamy, że znormalizowane próby są dane przez dwa niezależne wektory:

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \\ \vdots \\ \widetilde{x}_N \end{bmatrix} \quad i \quad \widetilde{\mathbf{Y}} = \begin{bmatrix} \widetilde{y}_1 \\ \widetilde{y}_2 \\ \vdots \\ \widetilde{y}_N \end{bmatrix}.$$

W tym wypadku hipoteza przyjmuje formę:

$$H_0^{(1,1)} : \text{Elementy } \widetilde{\mathbf{X}} \text{ i } \widetilde{\mathbf{Y}} \text{ mają rozkład normalny}$$
$$\text{vs.}$$
$$H_1^{(1,1)} : \ H_0^{(1,1)} \text{ nie jest prawdziwa.}$$

Uwaga 3.3.1 *Podkreślamy, że nie testujemy tutaj dwuwymiarowego rozkładu normalnego wektorów losowych (X, Y), ale testujemy normalność X i Y, ponieważ zakładamy niezależność ich rozkładów brzegowych.*

Dostajemy następującą reprezentację statystyki $T_{1,1}$ w zakresie funkcji specjalnych.

Propozycja 3.3.2 (Statystyka testowa dla $n = m = 1$)

$$T_{1,1} = 2\pi N\Big[\frac{1}{N^4}\sum_{n,j,k,l=1}^{N} J\Big(\sqrt{(\widetilde{x}_n - \widetilde{x}_k)^2 + (\widetilde{y}_j - \widetilde{y}_l)^2}\,\Big) - e^{-\frac{1}{2}}\frac{2}{N^2}\sum_{n,j=1}^{N} J\Big(\sqrt{\widetilde{x}_n^2 + \widetilde{y}_j^2}\,\Big) + e^{-1}\Big],$$

gdzie J to funkcja Bessela pierwszego rodzaju opisana w 2.1.

DOWÓD
Dla $a, b \in \mathbb{R}$ obliczamy

$$W(a,b) = |\Phi_{\widetilde{\mathbf{X}}}(a)\Phi_{\widetilde{\mathbf{Y}}}(b) \ - \ \exp(-\tfrac{1}{2})|^2$$
$$= \Phi_{\widetilde{\mathbf{X}}}(a)\Phi_{\widetilde{\mathbf{Y}}}(b)\overline{\Phi_{\widetilde{\mathbf{X}}}(a)\Phi_{\widetilde{\mathbf{Y}}}(b)} \ - \ \Phi_{\widetilde{\mathbf{X}}}(a)\Phi_{\widetilde{\mathbf{Y}}}(b)\exp(-\tfrac{1}{2})$$
$$- \ \overline{\Phi_{\widetilde{\mathbf{X}}}(a)\Phi_{\widetilde{\mathbf{Y}}}(b)}\exp(-\tfrac{1}{2}) \ + \ \exp(-1)$$
$$= \tfrac{1}{N^4}\sum_{n,j,k,l=1}^{N} \exp\left(ia(\widetilde{x}_n - \widetilde{x}_k) + ib(\widetilde{y}_j - \widetilde{y}_l)\right) - \tfrac{1}{N^2}\sum_{n,j}^{N}\exp\left(ia\widetilde{x}_n + ib\widetilde{y}_j\right)\exp(-\tfrac{1}{2})$$
$$-\tfrac{1}{N^2}\sum_{n,j=1}^{N}\exp\left(-ia\widetilde{x}_n - ib\widetilde{y}_j\right)\exp(-\tfrac{1}{2}) + \exp(-1).$$

Jeśli użyjemy zmiennych biegunowych podstawiając $a = \cos(\alpha)$ i $b = \sin(\alpha)$ otrzymujemy

$$= \frac{1}{N^4} \sum_{n,j,k,l=1}^{N} \cos\Big(\cos(\alpha)(\widetilde{x}_n - \widetilde{x}_k) + \sin(\alpha)(\widetilde{y}_j - \widetilde{y}_l\Big)$$
$$-\frac{2}{N^2} \sum_{n,j=1}^{N} \cos\Big(\cos(\alpha)\widetilde{x}_n + \sin(\alpha)\widetilde{y}_j\Big) \exp(-\tfrac{1}{2}) + \exp(-1).$$

Biorąc pod uwagę, że całkowanie po S_2 jest liniowe wzdłuż okręgu o środku w punkcie $(0,0)$ i promieniu równym 1, proponowana przez nas statystyka testowa wygląda tak:

$$T_{1,1} = N \int_{S_2} |\Phi_{\widetilde{\mathbf{X}}}(a)\Phi_{\widetilde{\mathbf{Y}}}(b) - \exp(-\frac{1}{2})|^2 dS_2 = N \int_0^{2\pi} W(\cos(\alpha), \sin(\alpha))d\alpha.$$

Oznacza to, że skupiamy się na obliczeniu całki

$$\int_0^{2\pi} \cos(\cos(\alpha)x + \sin(\alpha)y)d\alpha, \quad \text{dla } x, y \in \mathbb{R}.$$

Używając liniowych tożsamości trygonometrycznych i dodając harmoniczne sinusoidy i cosinusoidy dostajemy pojedynczą cosinusoidę z przesunięciem fazowym i przeskalowaną amplitudą. Wygląda to następująco:

$$\cos(\alpha)x + \sin(\alpha)y = \sqrt{x^2 + y^2}\cos(\alpha - atan2(y, x)),$$

gdzie $atan2(y, x)$ jest uogólnieniem $\arctan(y/x)$, który obejmuje cały zakres kołowy. Formalna definicja $atan2$ nie będzie nam potrzebna. Dzięki temu dla $x, y \in \mathbb{R}$ oraz $xy \neq 0$ otrzymujemy

$$\int_0^{2\pi} \cos(\cos(\alpha)x + \sin(\alpha)y)s\alpha = \int_0^{2\pi} \cos\Big(\sqrt{x^2 + y^2}\cos(\alpha - atan2(y, x))\Big)d\alpha$$

$$= \int_{-atan2(x,y)}^{2\pi - atan2(y,x)} cos\Big(\sqrt{x^2 + y^2}cos(t)\Big)dt = 2\pi J(\sqrt{x^2 + y^2}).$$

Używamy tutaj następującej zależności (szerzej opisanej w [2] str.360)

$$(4) \quad 2\pi J(z) = \int_0^{2\pi} \exp(iz\cos(\alpha))d\alpha = \int_0^{2\pi} \cos(z\cos(\alpha))d\alpha = \int_0^{2\pi} \cos(z\sin(\alpha))d\alpha.$$

Nawet dla $x = 0$ lub $y = 0$ powyższa formuła jest prawdziwa, ponieważ wówczas możemy wprost użyć równości (4). Ostatecznie, używając całkowania jak wyżej dla każdego składnika wyrażenia $W(\cos(\alpha), \sin(\alpha))$, dostajemy statystykę $T_{1,1}$.

□

W następnym rozdziale będziemy wykonywać symulacje dla tak określonej statystyki testowej w przypadku dwóch wektorów jednowymiarowych.

Rozdział 4

Symulacje

4.1 Wartości krytyczne

Niestety nie jesteśmy w stanie określić asymptotycznego rozkładu statystyki $T_{1,1}$. W tej części przedstawimy wyniki wykonanych przeze mnie symulacji pozwalających empirycznie przybliżyć wartości krytyczne tej statystyki. Symulacje te zostały wykonane w środowisku R. Dla różnych długości wektorów X i Y wykonujemy eksperyment Monte Carlo. Obliczamy wartość statystyki $T_{1,1}$ 1000 razy przy prawdziwości hipotezy zerowej i na tej podstawie wyznaczamy kwantyl odpowiedniego rzędu. Jest on szukaną przez nas empirycznie wyznaczoną wartością krytyczną. Ograniczamy się do wykonania tych symulacji tylko dla $N = 25$ oraz $N = 50$. Wyniki opisanych symulacji zostały przedstawione w tabeli 4.1.

Tablica 4.1: Empiryczne wartości krytyczne statystyki $T_{1,1}$

	Wielkość wektorów X i Y	
Poziom istotności	$N = 25$	$N = 50$
$\alpha = 0.1$	0.519	0.597
$\alpha = 0.05$	0.714	0.790

Patrząc na te wyniki możemy zauważyć, że wartości krytyczne dla tych samych poziomów istotności są sobie bliskie nawet dla różnego N. Możemy na tej podstawie wywnioskować, że statystyka testowa dość szybko osiąga swój asymptotyczny rozkład przy hipotezie zerowej. Zwracamy również uwagę na to, że mamy stabilną wartość krytyczną dzięki czemu możemy kontrolować błąd I-go rodzaju.

4.2 Moc testu

Sprawdzimy teraz jaką moc ma test oparty na omawianej przez nas statystyce przy różnych hipotezach alternatywnych. Te symulacje również zostały wykonane przeze mnie w środowisku R. Moce te zostały obliczone dla długości wektorów X i Y $N = 25$ oraz $N = 50$ na dwóch poziomach istotności $\alpha = 0.05$ oraz $\alpha = 0.1$. Do ich wyznaczenia wykorzystujemy oczywiście odpowiednie wartości krytyczne z tabeli 4.1. Wybieramy kilka podstawowych rozkładów prawdopodobieństwa, dla których sprawdzimy jak się zachowuje moc testu przy różnych alternatywach. W tabeli zostają one oznaczone następująco:

- $N(\mu, \sigma)$ - rozkład normalny o średniej μ i wariancji σ^2,
- $Exp(\lambda)$ - rozkład wykładniczy o parametrze λ,
- $Poi(\lambda)$ - rozkład wykładniczy ze średnią λ,
- $Unif(a, b)$ - ciągły rozkład jednostajny na odcinku $[a, b]$,
- χ^2_n - rozkład χ^2 z n stopniami swobody,
- $Gamma(k, \theta)$ - rozkład gamma z parametrem kształtu k oraz skali θ,
- $Beta(\alpha, \beta)$ - rozkład beta z parametrami kształtu α i β.

Wykonujemy symulacje dla różnych par tych rozkładów z wybranymi parametrami. Wyniki zostały przedstawione w tabeli 4.2.
Pierwszy rzut oka na tabelę 4.2 pozwala nam stwierdzić, że uzyskane przez nas wyniki zachowują się poprawnie. Dla większej liczby elementów wektorów X i Y moc się zwiększa. Podobnie jest w przypadku większej wartości poziomu istotności. Tak więc przy każdej alternatywie największą moc otrzymujemy dla przypadku $N = 50$ oraz $\alpha = 0.1$.
Wykonane symulacje pozwalają nam stwierdzić, że proponowany przez nas test sprawdza się dobrze tylko w niektórych przypadkach hipotezy alternatywnej. Są takie pary rozkładów w alternatywie, że moc testu jest bardzo słaba. Jako przykład możemy podać parę rozkładu normalnego z jednostajnym: $N(0, 1)$, $Unif(0, 1)$, normalnego z beta: $N(0, 1)$, $Beta(3, 2)$, dwóch jednostajnych: $Unif(0, 1)$, $Unif(0, 1)$, jednostajnego z beta: $Unif(0, 1)$, $Beta(3, 2)$ oraz beta z beta: $Beta(3, 2)$, $Beta(3, 2)$. Wówczas moc w najlepszym wypadku wynosi nie dużo ponad wartość 0.1. Możemy na tej podstawie wysnuć wniosek, że symulowane przez nas moce są słabsze dla alternatyw związanych z rozkładem gamma oraz jednostajnym. Widzimy również, że są pary rozkładów na których ten test działa bardzo dobrze i moc w tych

wypadkach jest bardzo bliska wartości 1. Są to na przykład wszystkie pary z rozkładem $Exp(2)$ jak również te z rozkładem $Gamma(3, 2)$.
Obserwujemy zatem, że omawiany przez nas test działa bardzo dobrze dla alternatyw związanych z rozkładami gamma oraz wykładniczym. Dla takich hipotez badawczych możemy spokojnie używać omawianego w tej pracy testu. Natomiast najsłabsze mocy uzyskujemy w wypadku alternatyw związanych z rozkładami jednostajnym oraz beta. W tych wypadkach stosowanie testu o statystyce testowej $T_{1,1}$ nie jest zalecane. Dla pozostałych par rozkładów wyniki mocy są średnie. Oczywiście dla małych wartości N są one gorsze. Możemy przypuszczać, że gdybyśmy rozszerzyli symulacje na przykład dla $N = 100$ lub $N = 200$ to uzyskane przez nas moce byłyby jeszcze większe. Wówczas moglibyśmy stwierdzić, że możemy używać tego testu dla takich alternatyw, ale tylko dla większych wartości N.

Tablica 4.2: Symulacje mocy testu opartego na statystyce $T_{1,1}$ przy różnych alternatywach

Poziom istotności:		$\alpha = 0.05$		$\alpha = 0.1$	
Alternatywa					
rozkład X	rozkład Y	$N = 25$	$N = 50$	$N = 25$	$N = 50$
$N(0,1)$	$Exp(2)$	0.729	0.986	0.859	0.992
$N(0,1)$	$Poi(3)$	0.156	0.262	0.225	0.355
$N(0,1)$	$Unif(0,1)$	0.026	0.017	0.064	0.061
$N(0,1)$	χ^2_{10}	0.259	0.509	0.400	0.628
$N(0,1)$	$Gamma(2,1)$	0.496	0.839	0.675	0.934
$N(0,1)$	$Beta(3,2)$	0.056	0.064	0.110	0.110
$Exp(2)$	$Exp(2)$	0.962	1.000	0.990	1.000
$Exp(2)$	$Poi(3)$	0.801	0.988	0.885	0.999
$Exp(2)$	$Unif(0,1)$	0.739	0.978	0.851	0.997
$Exp(2)$	χ^2_{10}	0.851	0.994	0.952	1.000
$Exp(2)$	$Gamma(2,1)$	0.931	1.000	0.958	1.000
$Exp(2)$	$Beta(3,2)$	0.736	0.985	0.850	0.994
$Poi(3)$	$Poi(3)$	0.201	0.460	0.346	0.584
$Poi(3)$	$Unif(0,1)$	0.077	0.218	0.181	0.332
$Poi(3)$	χ^2_{10}	0.315	0.634	0.471	0.766
$Poi(3)$	$Gamma(2,1)$	0.567	0.897	0.718	0.958
$Poi(3)$	$Beta(3,2)$	0.124	0.252	0.249	0.410
$Unif(0,1)$	$Unif(0,1)$	0.007	0.003	0.021	0.025
$Unif(0,1)$	χ^2_{10}	0.242	0.485	0.319	0.637
$Unif(0,1)$	$Gamma(2,1)$	0.460	0.843	0.607	0.925
$Unif(0,1)$	$Beta(3,2)$	0.018	0.032	0.041	0.089
χ^2_{10}	χ^2_{10}	0.434	0.801	0.588	0.889
χ^2_{10}	$Gamma(2,1)$	0.648	0.963	0.799	0.983
χ^2_{10}	$Beta(3,2)$	0.260	0.505	0.354	0.663
$Gamma(2,1)$	$Gamma(2,1)$	0.814	0.993	0.891	0.997
$Gamma(2,1)$	$Beta(3,2)$	0.476	0.854	0.640	0.927
$Beta(3,2)$	$Beta(3,2)$	0.035	0.075	0.072	0.168

Rozdział 5

Podsumowanie

Niestety statystyka $T_{1,1}$ którą omawiamy w 3.3 jest jedynym przypadkiem, dla którego możemy przedstawić dokładną formę statystyki $T_{m,n}$. Jest to niemożliwe nawet dla bardzo prostych przypadków jak $T_{2,1}$. Próba rozszerzenia rozumowania na $n > 1$ lub $m > 1$ prowadzi do nieznanych całek. Wyższe formy omawianej statystyki możemy jednak obliczać numerycznie.
Omawiany przez nas w tej pracy test można również stosować do testowania standardowej normalności. W wypadku jednowymiarowym używamy statystyki $T_{1,1}$. Musimy wówczas założyć, że rozkład Y jest znany, normalny i niezależny z X. Jeśli tak nie jest to możemy łatwo skonstruować Y spełniającego te warunki. Wówczas hipotezy testowe mają postać:

$$H_0^{(1,0)} : \text{zmienna losowa } X \text{ ma rozkład normalny}$$
$$\text{przeciwko}$$
$$H_1^{(1,0)} : \text{hipoteza } H_0^{(1,0)} \text{ nie jest prawdziwa}$$

W teście wielowymiarowej normalności i niezależności korzystamy z statystyki $T_{m,n}$. Podobnie jak w wypadku jednowymiarowym musimy założyć, że wektor losowy Y konstruujemy jako niezależny z X z rozkładu normalnego. Wymiar tego wektora możemy określać arbitralnie. Hipotezy wyglądają wówczas następująco:

$$H_0^{(m,0)} : \text{wszystkie elementy wektora } X \text{ są niezależne i mają rozkład normalny}$$
$$\text{przeciwko}$$
$$H_1^{(m,0)} : \text{hipoteza } H_0^{(m,0)} \text{ nie jest prawdziwa}$$

Hipoteza zerowa tego typu nie jest niczym nowym i została omówiona w [6]. W powyższej pracy przedstawiliśmy propozycję nowego testu niezależności i normalności. Dowiedliśmy jego postaci oraz wyprowadziliśmy statystykę

testową. Szczegółowo rozpatrzyliśmy przypadek jednowymiarowy, dla którego przeprowadziliśmy również symulacje. Wynika z nich, że test ten ma stabilną wartość krytyczną a symulacje mocy potwierdzają skuteczność tego testu przy wybranych alternatywach.

Bibliografia

[1] Ejsmont, W. (2019). A TEST FOR NORMALITY AND INDEPENDENCE BASED ON THE BESSEL FUNCTION. - praca wysłana

[2] Abramowitz, M. and Stegun, I. A. (1972). Bessel Functions J and Y. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover, pp. 358-364.

[3] Baringhaus, L. and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. Metrika 35(6), 339–348.

[4] Ejsmont, W. (2016). A characterization of the normal distribution by the independence of a pair of random vectors. Statistics and Probability Letters 114, 1–5.

[5] Epps, T. W. and Lawrence, B. (1983). A test for normality based on the empirical characteristic function. Biometrika 70(3), 723–726.

[6] Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 19, 279–281.

[7] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics 35(2), 2769–2794.

[8] Székely, Gábor J. and Rizzo, Maria L. (2009). Brownian distance covariance. Ann. Appl. Stat. 3(4), 1236–1265.

[9] Cook, L., 1971. A characterization of the normal distribution by the independence of a pair of random vectors and a property of the noncentral chi-square statistic. J. Multivariate Anal. 1, 457–460

[10] Kagan, A., Shalaevski, O., 1967. Characterization of normal law by a property of the non-central χ^2-distribution. Lith. Math. J. 7, 57–58.

[11] https://www.britannica.com/science/Bessel-function.

[12] http://www.mhtlab.uwaterloo.ca/courses/me755/web_chap4.pdf.