PRZEGLĄD STATYSTYCZNY
R. LVI – ZESZYT 2 – 2009

ANNA WITASZCZYK

ESTYMACJA UOGÓLNIONEJ WARIANCJI WYBRANYCH ROZKŁADÓW WIELOWYMIAROWYCH

1. WPROWADZENIE

Potrzeba stosowania metod statystyki wielowymiarowej występuje w wielu dziedzinach nauki. Podstawowym założeniem większości metod wielowymiarowej analizy statystycznej jest to, że próba pochodzi z populacji o wielowymiarowym rozkładzie normalnym. Celem pracy jest przedstawienie problemu estymacji uogólnionej wariancji, czyli wyznacznika macierzy kowariancji, w przypadku obserwacji pochodzących z populacji o rozkładzie wielowymiarowym niekoniecznie normalnym. Przedstawiono tzw. G-estymator uogólnionej wariancji otrzymany przez Girko w oparciu o twierdzenia graniczne dla wyznaczników losowych przy bardzo ogólnych założeniach dotyczących rozkładu wektora losowego [np. 4, 5, 6]. Własności G-estymatora zostały zbadane za pomocą eksperymentów symulacyjnych.

2. UOGÓLNIONA WARIANCJA

Koncepcja wariancji, jako miary koncentracji rozkładów jednowymiarowych, może być rozszerzona na przypadek rozkładów wielowymiarowych na dwa sposoby: jako obiekt geometryczny zwany elipsoidą koncentracji oraz jako miara liczbowa, która nazywa się uogólnioną wariancją.

Niech $\vec{\xi}$ będzie m wymiarowym wektorem losowym o wektorze wartości oczekiwanych $\mathbf{a} = E\vec{\xi}$ i macierzy kowariancji $\mathbf{R}_m = E\left(\vec{\xi} - \mathbf{a}\right)\left(\vec{\xi} - \mathbf{a}\right)^T$. Wyznacznik macierzy kowariancji nazywa się uogólnioną wariancją wektora $\vec{\xi}$.

Niech $\mathbf{x}_1, \ldots, \mathbf{x}_n$ będą niezależnymi obserwacjami wektora losowego $\vec{\xi}$ i niech $\mathbf{A} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$, gdzie $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$. Macierz $\hat{\mathbf{R}}_m = \frac{1}{n-1}\mathbf{A}$ nazywa się próbkową macierzą kowariancji. Jeżeli próba pochodzi z rozkładu normalnego $N_m(\mathbf{a}, \mathbf{R}_m)$, to $\hat{\mathbf{R}}_m$ jest nieobciążonym estymatorem parametru $\mathbf{R}_m$. Próbkowa macierz kowariancji ma rozkład Wisharta z $n - 1$ stopniami swobody i macierzą kowariancji $\frac{1}{n-1}\mathbf{R}_m$ [np. 2, 8, 9].

Jeżeli wykładnik funkcji gęstości wielowymiarowego rozkładu normalnego $(\mathbf{x} - \mathbf{a})^T\mathbf{R}_m^{-1}(\mathbf{x} - \mathbf{a})$ przyrównamy do pewnej stałej c, to tak otrzymane równanie jest równaniem elipsoidy w m wymiarowej przestrzeni euklidesowej. Zmieniając c

otrzymuje się rodzinę elipsoid o wspólnym środku w punkcie **a**, które nazywa się elipsoidami koncentracji. Pierwszą osią główną każdej z elipsoid nazywa się prostą przechodzącą przez środek elipsoidy oraz przez punkt najbardziej od tego środka oddalony. Długość pierwszej osi głównej jest równa $2\sqrt{\lambda_1 c}$, gdzie λ_1 oznacza największą wartość własną macierzy $\mathbf{R}_m$, natomiast jej kierunek określony jest przez kosinusy kierunkowe unormowanego wektora własnego α_1 odpowiadającego wartości własnej λ_1. Druga oś określona jest przez wektor własny odpowiadający drugiej co do wielkości wartości własnej macierzy $\mathbf{R}_m$ itd.

W analizie wielowymiarowej wiele statystyk można interpretować za pomocą objętości, są one odpowiednikami odległości, z którymi mamy do czynienia w przypadku jednowymiarowym. Objętość dowolnej elipsoidy jest równa $\dfrac{\pi^{\frac{m}{2}} c^{\frac{m}{2}} |\mathbf{R}_m|^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}m + 1\right)}$ gdzie $c > 0$. Jeżeli więc porównuje się tylko objętości elipsoid koncentracji dwóch rozkładów, to jest to równoważne z porównaniem ich uogólnionych wariancji.

Analogicznie do uogólnionej wariancji, uogólnioną wariancją z próby nazywa się wyznacznik macierzy $\hat{\mathbf{R}}_m$.

Podamy geometryczną interpretację próbkowej uogólnionej wariancji. Niech $\mathbf{t}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, $i = 1, \ldots, n$ i

$$\mathbf{Y}^T = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_m^T \end{bmatrix} = [\mathbf{t}_1, \ldots, \mathbf{t}_n], \quad \mathbf{y}_i \in R^n.$$

Współrzędnymi wektora $\mathbf{y}_i$ są i-te składowe wektorów $\mathbf{t}_1, \ldots, \mathbf{t}_n$. Wtedy $\mathbf{A} = \mathbf{Y}^T\mathbf{Y}$ i $a_{ii} = \mathbf{y}_i^T \mathbf{y}_i$ jest kwadratem długości wektora $\mathbf{y}_i$, a $a_{ij} = \mathbf{y}_i^T \mathbf{y}_j$ jest iloczynem długości wektorów $\mathbf{y}_i$ i $\mathbf{y}_j$ przez kosinus kąta między nimi. Rozważmy równoległościan zbudowany na tych m wektorach. Dla $m = 2$ będzie to równoległobok zbudowany na wektorach $\mathbf{y}_1$ i $\mathbf{y}_2$, dla $m = 3$ otrzymuje się równoległościan zbudowany na wektorach $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$ itd. Dla dowolnego m jest to figura ograniczona parami równoległych $(m - 1)$ wymiarowych hiperpłaszczyzn, z których jedna zawiera $m - 1$ wektorów spośród $\mathbf{y}_1, \ldots, \mathbf{y}_m$, a druga, równoległa do niej, przechodzi przez koniec pozostałego wektora. Wyznacznik macierzy **A**, tj. $|\mathbf{A}| = \left|(n-1)\hat{\mathbf{R}}_m\right| = (n-1)^m \left|\hat{\mathbf{R}}_m\right|$ jest równy kwadratowi objętości tak utworzonego równoległościanu [1].

Przedstawimy teraz interpretację geometryczną $|\mathbf{A}|$ wykorzystując wektory $\mathbf{t}_1, \ldots, \mathbf{t}_n$, które są punktami przestrzeni m wymiarowej. Gdy $m = 1$, to $|\mathbf{A}| = \sum_{i=1}^{n} t_{1i}^2$, czyli wyznacznik macierzy A jest równy sumie kwadratów odległości tych punktów od początku układu współrzędnych. W ogólnym przypadku wyznacznik macierzy **A** jest równy sumie kwadratów objętości wszystkich równoległościanów zbudowanych na m wektorach wybranych spośród $\mathbf{t}_1, \ldots, \mathbf{t}_n$. Biorąc pod uwagę, że $\mathbf{t}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ otrzymujemy, że próbkowa uogólniona wariancja $\left|\hat{\mathbf{R}}_m\right|$ jest proporcjonalna do sumy kwadratów objętości wszystkich równoległościanów utworzonych w następujący sposób: wektorami tworzącymi każdy równoległościan jest m wektorów, których wspólnym początkiem jest punkt $\bar{\mathbf{x}}$, a końce znajdują się w m punktach spośród $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Współczynnik proporcjonalności wynosi $\frac{1}{(N-1)^p}$.

Uogólniona wariancja z próby pochodzącej z populacji o rozkładzie normalnym ma taki rozkład jak zmienna $\frac{|\mathbf{R}_m|}{(n-1)^m}$ pomnożona przez iloczyn m niezależnych zmiennych, z których i-ta ma rozkład χ^2 z $(n-i)$ stopniami swobody $(i = 1, \ldots, m)$ (por. [8]).

Dla $m = 1$ lub $m = 2$ można otrzymać dokładny rozkład $|\hat{\mathbf{R}}_m|$, jednak dla większych wartości m jest on coraz bardziej skomplikowany i trudniej powiedzieć coś o jego własnościach, dlatego wygodniej posługiwać się rozkładem asymptotycznym. Jeżeli $\hat{\mathbf{R}}_m$ jest próbkową macierzą kowariancji o wymiarach $m \times m$ i $(n-1)$ stopniach swobody otrzymaną z próby z rozkładu normalnego $N_m(\mathbf{a}, \mathbf{R}_m)$, to zmienna losowa $\sqrt{n-1}\left(\frac{|\hat{\mathbf{R}}_m|}{|\mathbf{R}_m|} - 1\right)$ ma rozkład asymptotycznie normalny z wartością oczekiwaną zero i wariancją $2m$ [1].

Nieobciążony estymator wyrażenia $\ln|\mathbf{R}_m|$ ma postać:
$\ln|\hat{\mathbf{R}}_m| - \sum_{i=1}^{m} \psi\left(\frac{n-i}{2}\right) + m \ln \frac{n-1}{2}$, gdzie $\psi(x) = \frac{d \ln \Gamma(x)}{dx}$ (por. [8]).

3. G-ESTYMATOR UOGÓLNIONEJ WARIANCJI

W większości przypadków dokładne rozkłady wyznaczników macierzy losowych są nieznane albo mają skomplikowaną postać, dlatego poszukuje się ich rozkładów granicznych. W tradycyjnym podejściu przyjmuje się, że wymiar wektora losowego jest stały (m = const) i bada się własności estymatorów, gdy n dąży do nieskończoności. Girko w swoich pracach [np. 3, 4, 5, 6] rozważa ogólniejszy przypadek, kiedy wraz ze wzrostem liczby obserwacji również liczba składowych wektora losowego dąży do nieskończoności, przy czym faktycznie m zależy od n, co zaznaczymy stosując oznaczenie m_n. Liczby m_n i n spełniają następujący tzw. G-warunek:

$$\lim_{n \to \infty} \frac{m_n}{n} = const < \infty.$$

Przy powyższym założeniu rozważany jest następujący problem: przy odpowiednim doborze stałych normujących a_n, b_n znaleźć warunki zbieżności, a także ogólną postać rozkładów granicznych dla ciągu zmiennych losowych

$$\frac{\ln|\Xi_n| - a_n}{b_n}, \quad n = 1, 2, \ldots, \tag{1}$$

gdzie $\Xi_n = \left[\xi_{ij}^{(n)}\right]$ jest kwadratową macierzą losową stopnia n.

W GSA (General Statistical Analysis) zakłada się, że elementy macierzy Ξ_n są niezależne i mają niektóre jednakowe własności. W dowodach twierdzeń granicznych wykorzystuje się między innymi metodę przekształceń ortogonalnych. Polega ona na tym, że wyznacznik macierzy Ξ_n przedstawia się w postaci iloczynu zmiennych losowych, co pozwala na uogólnienie teorii sumowania zmiennych losowych dla logarytmu wyznacznika macierzy losowej.

Poniżej przedstawione są dwa twierdzenia, które są podstawą konstrukcji estymatora uogólnionej wariancji [5].

Przy pewnych założeniach dotyczących elementów macierzy losowej Ξ_n zachodzi następujące centralne twierdzenie graniczne dla wyznaczników losowych:

Twierdzenie 1. Niech dla każdej wartości n elementy losowe $\xi_{ij}^{(n)}$, $i, j = 1, \dots n$ macierzy Ξ_n będą niezależne, $E\xi_{ij}^{(n)} = 0$, $D\xi_{ij}^{(n)} = 1$, $E\left(\xi_{ij}^{(n)}\right)^4 = 3$, dla pewnego $\delta > 0$

$$\sup_n \max_{i,j=1,\dots,n} E\left|\xi_{ij}^{(n)}\right|^{4+\delta} < \infty.$$

Wtedy

$$\lim_{n \to \infty} P\left\{\frac{\ln\left|\Xi_n^2\right| - \ln(n-1)!}{\sqrt{2\ln n}} < x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy.$$

Dla zmiennej $\frac{1}{c_n}\ln\left|\hat{\mathbf{R}}_m\right|$, gdzie $\{c_n\}$ jest pewnym ciągiem zachodzi słabe prawo wielkich liczb i centralne twierdzenie graniczne.

Twierdzenie 2. Niech m_n wymiarowe wektory $\mathbf{x}_1, \dots, \mathbf{x}_n$ dla każdego $n > m_n$ będą niezależne, mają jednakowe rozkłady o wektorze wartości oczekiwanych $\mathbf{a}$ i nieosobliwej macierzy kowariancji $\mathbf{R}_{m_n}$, dla pewnego $\delta > 0$

$$\sup_n \sup_{\substack{i=1,\dots,n \\ j=1,\dots,m_n}} E\left|\tilde{x}_{ij}\right|^{4+\delta} < \infty \tag{2}$$

gdzie $\tilde{x}_{ij}$ są składowymi wektora $\tilde{\mathbf{x}}_i = \mathbf{R}_{m_n}^{-\frac{1}{2}}(\mathbf{x}_i - \mathbf{a})$,

$$\lim_{n \to \infty}(n - m_n) = \infty, \quad \lim_{n \to \infty}\frac{m_n}{n} = 1 \tag{3}$$

1i dla każdej wartości $n > m_n$ wektory losowe $\tilde{\mathbf{x}}_i$ będą niezależne.

Wtedy

$$p\lim_{n \to \infty}\frac{1}{c_n}\left\{\ln\left|\hat{\mathbf{R}}_{m_n}\right| + \ln\frac{(n-1)^{m_n} n}{A_{n-1}^{m_n}(n-m_n)} - \ln\left|\mathbf{R}_{m_n}\right|\right\} = 0, \tag{4}$$

gdzie $A_n^m = n(n-1)\ldots(n-m+1)$ a $\{c_n\}$ jest pewnym ciągiem takim, że $\lim\limits_{n\to\infty}\frac{1}{c_n^2}\ln\frac{n}{n-m_n} = 0.$

Jeżeli ponadto

$$E\left(\tilde{x}_{ij}\right)^4 = 3 \quad \text{dla } i = 1,\ldots,n,\ j = 1,\ldots,m_n, \tag{5}$$

to

$$\lim_{n\to\infty} P\left\{\frac{\ln\left|\hat{\mathbf{R}}_{m_n}\right| + \ln\frac{(n-1)^{m_n}n}{A_{n-1}^{m_n}(n-m_n)} - \ln\left|\mathbf{R}_{m_n}\right|}{\sqrt{-2\ln\left(1-\frac{m_n}{n}\right)}} < x\right\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}dy. \tag{6}$$

Z twierdzenia 2 wynika, że jeśli spełnione są warunki (2) i (3), to za estymator wielkości $\ln|\mathbf{R}_{m_n}|$ można przyjąć

$$G\left(\hat{\mathbf{R}}_{m_n}\right) = \ln\left(\left|\hat{\mathbf{R}}_{m_n}\right|\frac{(n-1)^{m_n}n}{A_{n-1}^{m_n}(n-m_n)}\right), \tag{7}$$

Przy czym, jeżeli spełniony jest warunek (5), to zachodzi centralne twierdzenie graniczne (6). $G\left(\hat{\mathbf{R}}_{m_n}\right)$ nazywa się G-estymatorem uogólnionej wariancji.

Twierdzenie 2 można uogólnić na przypadek, gdy $n > m_n$, $\lim\limits_{n\to\infty}(n-m_n) < \infty$, $\overline{\lim\limits_{n\to\infty}}\frac{m_n}{n} \le 1$.

4. PRZYKŁADY SYMULACYJNE

W celu zbadania własności G-estymatora przeprowadzone zostały eksperymenty symulacyjne.

Niech wektor losowy $(X_1, \ldots, X_p)$ będzie p wymiarową zmienną losową o rozkładzie normalnym $N_p(\mathbf{a}, \mathbf{R})$. Według tego rozkładu wygenerowano n wektorów i na ich podstawie wyznaczono wartość estymatora $G(\hat{\mathbf{R}})$ oraz klasycznego estymatora największej wiarogodności $\ln|\hat{\mathbf{R}}|$. Eksperymenty realizowano dla $k = 10000$ powtórzeń. Obliczone zostały średnie wartości estymatorów oraz błędy średniokwadratowe. Obliczenia wykonano dla różnych wartości p, n, $\ln|\mathbf{R}|$, $\mathbf{a} = \mathbf{0}$.

Następne eksperymenty, analogiczne, zostały przeprowadzone dla innych niż normalny rozkładów wielowymiarowych: rozkładu t-Studenta, rozkładów Pearsona, wykładniczego, Laplace'a, Weibulla, Beta.

Wyniki przedstawione są w tablicach 1-7. Na ich podstawie widać, że estymator $\ln|\hat{\mathbf{R}}|$ ze wzrostem liczebności próby dąży do wartości szacowanego parametru monotonicznie, ale o wiele wolniej niż G-estymator. $G(\hat{\mathbf{R}})$ stosunkowo szybko osiąga wartości bliskie $\ln|\mathbf{R}|$ i, w przypadku rozkładu normalnego, oscyluje w otoczeniu rzeczywistej wartości parametru przyjmując wartość większą bądź mniejszą niż $\ln|\mathbf{R}|$. Dla ustalonego p wartości błędów średniokwadratowych obu estymatorów ze wzrostem n przyjmują zbliżone wartości. Dla małych n błąd średniokwadratowy G-estymatora jest zdecydowanie mniejszy.

Dla rozkładów innych niż wielowymiarowy rozkład normalny estymator $G(\hat{\mathbf{R}})$ ma podobne własności jak w przypadku N_p.

Eksperymenty pokazały, że estymator uogólnionej wariancji, otrzymany za pomocą metodologii GSA daje bardzo dobre oszacowanie i można go stosować nawet wtedy, gdy próba losowa nie jest zbyt liczna.

Ze względu na swoje własności G-estymatory znajdują z powodzeniem zastosowanie w zagadnieniach praktycznych [7].

Tabela 1

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu normalnego

Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
1,609	2	5	1,476	0,678	1,594	2,445
		10	1,583	1,242	0,534	0,668
		20	1,605	1,445	0,227	0,253
		30	1,613	1,509	0,147	0,157
		40	1,606	1,529	0,110	0,116
		50	1,610	1,548	0,085	0,089
		70	1,608	1,564	0,059	0,061
		100	1,611	1,581	0,041	0,042
1,946	3	5	1,643	–0,255	3,342	8,093
		10	1,902	1,176	0,871	1,462
		20	1,937	1,609	0,356	0,469
		30	1,948	1,736	0,223	0,267
		40	1,944	1,787	0,162	0,188
		50	1,944	1,820	0,129	0,145
		70	1,948	1,860	0,090	0,097
		100	1,938	1,877	0,062	0,067
2,398	5	8	2,182	–0,697	2,810	12,339
		10	2,295	0,239	1,750	6,398
		20	2,392	1,531	0,628	1,380
		30	2,391	1,845	0,398	0,704
		40	2,385	1,984	0,271	0,442
		50	2,386	2,070	0,220	0,327
		70	2,402	2,180	0,152	0,199
		100	2,394	2,240	0,104	0,129

cd. tabeli 1

Wartość parametru ln $\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
3,045	10	12	2,621	–5,647	5,499	80,867
		20	3,013	–0,587	1,556	14,740
		30	3,020	0,858	0,856	5,635
		40	3,054	1,506	0,591	2,959
		50	3,039	1,831	0,466	1,938
		70	3,034	2,196	0,316	1,036
		100	3,045	2,460	0,216	0,546
3,434	15	18	3,165	–8,313	4,655	142,579
		20	3,289	–6,102	3,335	94,257
		30	3,393	–1,744	1,462	28,269
		40	3,419	–0,157	0,992	13,891
		50	3,418	0,669	0,733	8,375
		70	3,427	1,545	0,505	4,073
		100	3,437	2,158	0,336	1,965
3,714	20	23	3,418	–12,662	5,113	273,175
		30	3,658	–6,331	2,420	103,311
		40	3,704	–2,969	1,484	46,137
		50	3,708	–1,331	1,058	26,505
		70	3,704	0,310	0,699	12,282
		100	3,703	1,420	0,447	5,706
3,932	25	28	3,611	–17,154	5,690	450,197
		30	3,761	–14,134	4,262	330,611
		40	3,879	–7,250	2,070	127,103
		50	3,892	–4,316	1,449	69,476
		70	3,925	–1,496	0,918	30,380
		100	3,924	0,320	0,586	13,635

Źródło: obliczenia własne.

Tabela 2

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Studenta

Liczba stopni swobody v	Wartość parametru $\ln|\mathbf{R}|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln|\hat{\mathbf{R}}|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln|\hat{\mathbf{R}}|)$
6	0,811	2	5	0,376	–0,422	2,283	3,315
			10	0,585	0,244	0,904	1,175
			20	0,691	0,531	0,453	0,517
			30	0,718	0,613	0,320	0,350
			40	0,738	0,660	0,242	0,259
			50	0,758	0,697	0,190	0,201
			70	0,771	0,727	0,139	0,144
			100	0,772	0,742	0,102	0,105
6	2,027	5	8	0,955	–1,924	5,535	19,995
			10	1,201	–0,855	3,739	11,362
			20	1,540	0,679	1,713	3,294
			30	1,674	1,127	1,153	1,838
			40	1,743	1,343	0,873	1,261
			50	1,795	1,479	0,717	0,964
			70	1,848	1,625	0,519	0,649
			100	1,888	1,734	0,379	0,446
6	4,055	10	12	1,592	–6,676	15,235	124,320
			20	2,586	–1,013	6,237	29,766
			30	2,987	0,826	3,974	13,258
			40	3,170	1,621	2,973	8,111
			50	3,327	2,120	2,324	5,538
			70	3,491	2,652	1,661	3,309
			100	3,643	3,067	1,198	2,003
15	0,286	2	5	0,082	–0,717	1,822	2,787
			10	0,209	–0,132	0,653	0,822
			20	0,247	0,087	0,291	0,329
			30	0,259	0,154	0,190	0,206
			40	0,273	0,196	0,137	0,145
			50	0,270	0,209	0,113	0,119
			70	0,278	0,234	0,078	0,081
			100	0,280	0,250	0,055	0,056

cd. tabeli 2

Liczba stopni swobody v	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
15	0,716	5	8	0,270	–2,608	3,543	14,392
			10	0,400	–1,655	2,271	7,793
			20	0,569	–0,293	0,930	1,925
			30	0,611	0,064	0,567	0,980
			40	0,638	0,238	0,426	0,648
			50	0,664	0,348	0,341	0,473
			70	0,676	0,454	0,237	0,303
			100	0,689	0,535	0,164	0,196
15	1,431	10	12	0,365	–7,903	7,838	93,831
			20	0,965	–2,634	2,438	16,818
			30	1,118	–1,044	1,567	7,592
			40	1,180	–0,368	1,125	4,300
			50	1,221	0,014	0,861	2,825
			70	1,275	0,436	0,627	1,592
			100	1,320	0,745	0,419	0,877
30	0,138	2	5	–0,022	–0,820	1,718	2,611
			10	0,088	–0,253	0,583	0,733
			20	0,120	–0,039	0,253	0,284
			30	0,122	0,018	0,165	0,180
			40	0,128	0,051	0,122	0,130
			50	0,130	0,069	0,095	0,100
			70	0,132	0,088	0,067	0,070
			100	0,133	0,103	0,047	0,048
30	0,345	5	8	–0,008	–2,886	3,133	13,448
			10	0,146	–1,909	1,985	7,028
			20	0,271	–0,590	0,745	1,613
			30	0,313	–0,234	0,475	0,809
			40	0,310	–0,090	0,350	0,538
			50	0,333	0,017	0,263	0,371
			70	0,321	0,099	0,186	0,247
			100	0,335	0,181	0,133	0,160

cd. tabeli 2

Liczba stopni swobody v	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
30	0,690	10	12	0,010	–8,258	6,474	86,086
			20	0,438	–3,161	1,993	16,762
			30	0,531	–1,630	1,145	6,502
			40	0,567	–0,981	0,822	3,600
			50	0,599	–0,609	0,628	2,306
			70	0,627	–0,211	0,442	1,251
			100	0,641	0,066	0,304	0,691

Źródło: obliczenia własne.

Tabela 3

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowych rozkładów Pearsona

Wartość parametru β_1	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
0,2	0	2	5	–0,104	–0,902	1,582	2,385
			10	–0,023	–0,364	0,530	0,662
			20	–0,006	–0,165	0,232	0,259
			30	–0,007	–0,112	0,148	0,160
			40	–0,001	–0,078	0,108	0,114
			50	0,001	–0,060	0,085	0,088
			70	–0,002	–0,046	0,059	0,061
			100	–0,005	–0,035	0,041	0,042
0,2	0	3	5	–0,290	–2,188	3,378	8,079
			10	–0,039	–0,764	0,874	1,456
			20	–0,003	–0,324	0,352	0,457
			30	–0,004	–0,216	0,219	0,266
			40	–0,003	–0,159	0,159	0,184
			50	0,000	–0,124	0,128	0,143
			70	0,001	–0,086	0,092	0,099
			100	–0,001	–0,062	0,062	0,066

cd. tabeli 3

Wartość parametru β_1	Wartość parametru $\ln\lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\lvert\hat{\mathbf{R}}\rvert)$
0,2	0	5	10	–0,080	–2,136	1,735	6,290
			20	–0,006	–0,868	0,637	1,390
			30	–0,002	–0,549	0,389	0,690
			40	–0,008	–0,408	0,275	0,441
			50	–0,001	–0,317	0,215	0,316
			70	–0,002	–0,224	0,150	0,200
			100	–0,001	–0,155	0,104	0,128
0,4	0	2	5	–0,120	–0,918	1,629	2,457
			10	–0,028	–0,369	0,543	0,679
			20	–0,010	–0,169	0,232	0,261
			30	–0,002	–0,106	0,147	0,156
			40	–0,005	–0,083	0,107	0,114
			50	0,003	–0,059	0,085	0,089
			70	–0,002	–0,046	0,060	0,062
			100	–0,004	–0,035	0,040	0,042
0,4	0	3	5	–0,328	–2,226	3,409	8,255
			10	–0,052	–0,778	0,877	1,481
			20	–0,011	–0,339	0,358	0,472
			30	–0,002	–0,214	0,226	0,272
			40	0,000	–0,157	0,163	0,187
			50	0,002	–0,122	0,128	0,143
			70	0,006	–0,082	0,089	0,096
			100	0,000	–0,061	0,062	0,066
0,4	0	5	10	–0,101	–2,156	1,763	6,402
			20	–0,023	–0,884	0,648	1,429
			30	–0,003	–0,550	0,395	0,697
			40	0,003	–0,397	0,280	0,438
			50	–0,009	–0,325	0,219	0,325
			70	–0,002	–0,224	0,155	0,205
			100	–0,007	–0,161	0,104	0,130

cd. tabeli 3

Wartość parametru β_1	Wartość parametru $\ln\lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\lvert\hat{\mathbf{R}}\rvert)$
1	0	2	5	–0,315	–1,113	2,514	3,654
			10	–0,067	–0,408	0,730	0,892
			20	–0,011	–0,170	0,268	0,297
			30	0,003	–0,101	0,163	0,173
			40	0,009	–0,069	0,114	0,118
			50	0,007	–0,055	0,091	0,094
			70	0,012	–0,032	0,063	0,064
			100	0,011	–0,020	0,043	0,043
1	0	3	5	–0,601	–2,498	4,901	10,781
			10	–0,108	–0,834	1,163	1,846
			20	0,001	–0,327	0,407	0,514
			30	0,003	–0,209	0,245	0,289
			40	0,006	–0,151	0,178	0,200
			50	0,012	–0,112	0,134	0,146
			70	0,016	–0,072	0,095	0,099
			100	0,012	–0,049	0,064	0,067
1	0	5	10	–0,198	–2,254	2,273	7,312
			20	–0,011	–0,872	0,729	1,490
			30	–0,002	–0,549	0,434	0,736
			40	0,016	–0,384	0,304	0,451
			50	0,016	–0,300	0,234	0,324
			70	0,023	–0,200	0,163	0,203
			100	0,021	–0,133	0,107	0,124

Źródło: obliczenia własne.

Tabela 4

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu wykładniczego z parametrem $\lambda = 1$

Wartość parametru $\ln\lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln\lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\lvert\hat{\mathbf{R}}\rvert)$
0	2	5	–0,795	–1,594	3,618	5,525
		10	–0,437	–0,778	1,604	2,018
		20	–0,241	–0,400	0,761	0,863
		30	–0,157	–0,261	0,497	0,540
		40	–0,123	–0,200	0,381	0,406
		50	–0,105	–0,166	0,325	0,341
		70	–0,075	–0,119	0,221	0,230
		100	–0,052	–0,083	0,153	0,157
0	3	5	–1,335	–3,232	7,213	15,878
		10	–0,648	–1,374	2,559	4,026
		20	–0,354	–0,682	1,198	1,538
		30	–0,263	–0,475	0,794	0,951
		40	–0,197	–0,353	0,592	0,678
		50	–0,169	–0,293	0,471	0,529
		70	–0,114	–0,202	0,342	0,370
		100	–0,085	–0,146	0,236	0,250
0	5	8	–1,488	–4,366	7,475	24,327
		10	–1,137	–3,192	5,208	14,106
		20	–0,591	–1,452	2,193	3,952
		30	–0,423	–0,970	1,434	2,195
		40	–0,334	–0,734	1,049	1,477
		50	–0,261	–0,577	0,832	1,097
		70	–0,195	–0,417	0,580	0,716
		100	–0,140	–0,294	0,414	0,481

Źródło: obliczenia własne

Tabela 5

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Laplace'a z parametrem $\lambda = 1$

Wartość parametru $\ln\lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
			$G(\hat{\mathbf{R}})$	$\ln\lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\lvert\hat{\mathbf{R}}\rvert)$
1,386	2	5	0,845	0,046	2,750	4,252
		10	1,127	0,786	1,082	1,375
		20	1,250	1,091	0,512	0,580
		30	1,300	1,204	0,339	0,368
		40	1,321	1,244	0,245	0,261
		50	1,328	1,266	0,194	0,205
		70	1,351	1,307	0,142	0,147
		100	1,355	1,325	0,098	0,101
2,079	3	5	1,160	–0,737	5,399	12,485
		10	1,701	0,975	1,751	2,827
		20	1,887	1,559	0,760	0,994
		30	1,944	1,732	0,518	0,620
		40	1,982	1,826	0,378	0,433
		50	1,996	1,872	0,304	0,340
		70	2,011	1,923	0,218	0,238
		100	2,039	1,978	0,152	0,160
3,466	5	8	2,574	–0,304	5,095	18,513
		10	2,784	0,728	3,460	10,488
		20	3,141	2,280	1,399	2,699
		30	3,215	2,669	0,924	1,496
		40	3,279	2,879	0,657	0,966
		50	3,317	3,001	0,523	0,717
		70	3,362	3,140	0,364	0,460
		100	3,396	3,242	0,251	0,296

Źródło: obliczenia własne.

Tabela 6

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Weibulla

Wartość parametru λ	Wartość parametru $\ln\lvert\mathbf{R}\rvert$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\lvert\hat{\mathbf{R}}\rvert$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\lvert\hat{\mathbf{R}}\rvert)$
0,5	5,991	2	5	2,432	1,633	22,208	28,530
			10	3,697	3,356	10,587	12,267
			20	4,530	4,371	5,167	5,658
			30	4,900	4,796	3,384	3,622
			40	5,091	5,013	2,633	2,778
			50	5,219	5,158	2,221	2,319
			70	5,376	5,332	1,560	1,615
			100	5,529	5,499	1,122	1,151
0,5	8,987	3	5	3,381	1,483	46,849	71,721
			10	5,517	4,791	19,962	25,526
			20	6,795	6,469	9,400	10,945
			30	7,287	7,075	6,273	7,038
			40	7,605	7,448	4,639	5,097
			50	7,792	7,668	3,694	4,007
			70	8,063	7,975	2,665	2,835
			100	8,274	8,213	1,890	1,981
0,5	14,979	5	8	7,898	5,020	66,758	115,805
			10	9,021	6,965	49,150	77,869
			20	11,298	10,437	21,323	28,404
			30	12,165	11,618	13,719	17,093
			40	12,660	12,260	9,951	11,968
			50	13,020	12,704	7,767	9,105
			70	13,434	13,211	5,366	6,102
			100	13,790	13,636	3,745	4,135
2	–3,078	2	5	–3,221	–4,019	1,603	2,469
			10	–3,118	03,459	0,572	0,715
			20	–3,092	–3,251	0,246	0,276
			30	–3,078	–3,182	0,157	0,168
			40	–3,079	–3,157	0,116	0,122
			50	–3,082	–3,144	0,098	0,103
			70	–3,078	–3,122	0,067	0,068
			100	–3,079	–3110	0,045	0,046

cd. tabeli 6

Wartość parametru λ	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
2	–4,617	3	5	–4.945	–6,842	3,364	8,207
			10	–4.658	–5,384	0,904	1,491
			20	–4,631	–4,959	0,380	0,497
			30	–4,637	–4,849	0,245	0,298
			40	–4,626	–4,783	0,180	0,208
			50	–4,628	–4,753	0,142	0,160
			70	–4,618	–4,706	0,101	0,109
			100	–4,622	–4,683	0,068	0,072
2	–7,695	5	8	–7,947	–10,825	2,888	12,624
			10	–7,809	–9,865	1,857	6,552
			20	–7,714	–8,575	0,684	1,458
			30	–7,713	–8,259	0,423	0,742
			40	–7,712	–8,112	0,311	0,485
			50	–7,702	–8,018	0,242	0,347
			70	–7,704	–7,926	0,170	0,223
			100	–7,702	–7,855	0,119	0,145

Źródło: obliczenia własne.

Tabela 7

Własności estymatorów $G(\hat{\mathbf{R}})$ i $\ln|\hat{\mathbf{R}}|$ dla p wymiarowego rozkładu Beta

Wartości parametrów (α, β)	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
(0.5, 0.5)	–4,159	2	5	–4,063	–4,862	1,217	1,702
			10	–4,040	–4,381	0,239	0,274
			20	-4,090	–4,250	0,078	0,081
			30	–4,108	–4,212	0,046	0,046
			40	–4,123	–4,196	0,031	0,030
			50	–4,125	–4,187	0,024	0,024
			70	–4,135	–4,178	0,017	0,016
			100	–4,143	–4,173	0,011	0,011

cd. tabeli 7

Wartości para-metrów (α, β)	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
(0.5, 0.5)	–6,238	3	5	–6,221	–8,118	2,704	6,238
			10	–6,057	–6,783	0,441	0,705
			20	–6,125	–6,453	0,134	0,167
			30	–6,160	–6,372	0,074	0,086
			40	–6,183	–6,340	0,051	0,059
			50	–6,195	–6,320	0,039	0,044
			70	–6,202	–6,290	0,026	0,027
			100	–6,214	–6,275	0,017	0,018
(0.5, 0.5)	–10,397	5	8	–10,178	–13,056	1,977	8,997
			10	–10,123	–12,178	1,064	4,161
			20	–10,221	–11,082	0,273	0,711
			30	–10,273	–10,820	0,145	0,308
			40	–10,301	–10,701	0,096	0,179
			50	–10,316	–10,631	0,070	0,119
			70	–10,339	–10,562	0,046	0,069
			100	–10,354	–10,508	0,031	0,041
(2, 0.5)	–6,171	2	5	–6,698	–7,496	3,216	4,696
			10	–6,351	–6,692	1,006	1,245
			20	–6,236	–6,395	0,376	0,422
			30	–6,217	–6,321	0,233	0,254
			40	–6,190	–6,267	0,163	0,172
			50	–6,193	–6,255	0,128	0,135
			70	–6,179	–6,222	0,087	0,090
			100	–6,176	–6,206	0,060	0,061
(2, 0.5)	–9,256	3	5	–10,204	–12,101	6,350	13,545
			10	–9,503	–10,229	1,543	2,429
			20	–9,348	–9,676	0,579	0,747
			30	–9,308	–9,520	0,344	0,411
			40	–9,281	–9,438	0,247	0,280
			50	–9,286	–9,410	0,190	0,213
			70	–9,271	–9,359	0,130	0,140
			100	–9,267	–9,328	0,087	0,092

cd. tabeli 7

Wartości para-metrów (α, β)	Wartość parametru $\ln\|\mathbf{R}\|$	Wymiar wektora p	Liczebność próby n	Średnia wartość estymatora		Błąd średniokwadratowy	
				$G(\hat{\mathbf{R}})$	$\ln\|\hat{\mathbf{R}}\|$	$BSK(G(\hat{\mathbf{R}}))$	$BSK(\ln\|\hat{\mathbf{R}}\|)$
(2, 0.5)	–15,427	5	8	–16,156	–19,034	5,067	17,549
			10	–15,891	–17,947	2,989	9,124
			20	–15,613	–16,474	1,049	2,111
			30	–15,523	–16,070	0,587	0,991
			40	–15,493	–15,894	0,435	0,648
			50	–15,480	–15,796	0,324	0,457
			70	–15,454	–15,676	0,218	0,279
			100	–15,444	–15,598	0,150	0,180
(1, 2)	–5.781	2	5	–5,818	–6,617	1,494	2,191
			10	–5,749	–6,090	0,456	0,551
			20	–5,757	–5,917	0,177	0,195
			30	–5,760	–5,864	0,107	0,113
			40	–5,773	–5,850	0,081	0,085
			50	–5,768	–5,829	0,063	0,065
			70	–5,769	–5,812	0,043	0,044
			100	–5,776	–5,806	0,030	0,031
(1, 2)	–8,671	3	5	–8,833	–10,730	3,189	7,402
			10	–8,642	–9,368	0,732	1,216
			20	–8,635	–8,963	0,282	0,366
			30	–8,645	–8,857	0,168	0,201
			40	–8,647	–8,804	0,119	0,135
			50	–8,650	–8,775	0,094	0,105
			70	–8,660	–8,748	0,065	0,070
			100	–8,663	–8,724	0,044	0,047
(1, 2)	–14,452	5	8	–14,525	–17,403	2,589	11,294
			10	–14,419	–16,475	1,551	5,642
			20	–14,398	–15,259	0,501	1,149
			30	–14,410	–14,957	0,290	0,543
			40	–14,424	–14,825	0,206	0,344
			50	–14,425	–14,742	0,157	0,240
			70	–14,434	–14,656	0,112	0,153
			100	–14,435	–14,588	0,075	0,093

Źródło: obliczenia własne.

Uniwersytet Łódzki, Katedra Metod Statystycznych

LITERATURA

[1] Anderson T.W., [1958], *An introduction to multivariate statistical analysis*, John Wiley and Sons, NY, London.
[2] Barra J.R., [1982], *Matematyczne podstawy statystyki*, PWN, Warszawa.
[3] Girko V.L., [1975], *Random Matrices* (in Russian), Kiev University, Kiev.
[4] Girko V.L., [1988], *Spectral Theory of Random Matrices* (in Russian), Nauka, Moscow.
[5] Girko V.L., [1995], *Statistical Analysis of Observations of Increasing Dimension*, Kluwer Academic Publishers.
[6] Girko V.L., [1990], *Theory of Random Determinants*, Kluwer Academic Publisher.
[7] Johnson B.A., Abramovich Y.I., Mestre X., [Aug. 2008], *MUSIC, G-MUSIC,and Maximum- Likelihood Performance Breakdown*, IEEE Transactions on Signal Processing, Vol. 56, No. 2.
[8] Krzyśko M., [2000], *Wielowymiarowa analiza statystyczna*, Rektor UAM, Poznań.
[9] Rao C.R., [1982], *Modele liniowe statystyki matematycznej*, PWN, Warszawa.

Praca wpłynęła do redakcji w maju 2009 r.

ESTYMACJA UOGÓLNIONEJ WARIANCJI WYBRANYCH ROZKŁADÓW WIELOWYMIAROWYCH

Streszczenie

Uogólniona wariancja, czyli wyznacznik macierzy kowariancji jest skalarną miarą rozrzutu rozkładów wielowymiarowych. Dokładny rozkład uogólnionej wariancji znany jest tylko dla wektorów losowych o wielowymiarowym rozkładzie normalnym. Dla wektorów losowych o dużych wymiarach przyjmuje on skomplikowaną postać, co stanowi utrudnienie w zastosowaniach praktycznych.

W pracy przedstawiono tzw. *G*-estymator logarytmu uogólnionej wariancji otrzymany na podstawie twierdzeń granicznych dla wyznaczników losowych i jego własności na przykładach symulacyjnych dla kilku wybranych rozkładów wielowymiarowych.

Słowa kluczowe: uogólniona wariancja, wektor losowy, macierz kowariancji, wyznacznik losowy.

ESTIMATION OF GENERALIZED VARIANCE OF CHOSEN MULTIVARIATE DISTRIBUTION

Summary

Generalized variance i.e. the determinant of the covariance matrix is a scalar measure of multivariate distribution dispersion. The exact distribution of the generalized variance is known only for multivariate normal vectors. For random vectors in high dimensional spaces it has a complicated formula very troublesome to apply.

An estimator of the logarithm of generalized variance derived with the help of limit theorems for random determinants was presented as well as its properties in examples of chosen simulation multivariate distributions.

Key words: generalized variance, random vector, covariance matrix, random determinant.