Testowanie hipotez w wybranych wielowymiarowych rozkładach normalnych

Roman Zmyślony

Kościelisko 2016, 5-13 Września

Plan prezentacji

1. Wielowymiarowy model normalny ze strukturą BS macierzy kowariancji
2. Algebry Jordana - zastosowania w statystyce
3. Model mieszany jednowymiarowy i optymalne estymatory i testy dla komponentów wariancyjnych
4. Test F i LRT dotyczący struktury kowariancji
5. Porównanie mocy (symulacje)

1. Wprowadzenie

Blokowo-symetryczna (BS) struktura kowariancji dla podwójnie wielowymiarowych danych (m wymiarowe wektory obserwacji powstałe poprzez powtarzane pomiary ze względu na u czynników lub punktów czasowych), która jest wielowymiarowym uogólnieniem symetrycznej struktury kowariancji dla wielowymiarowych obserwacji, została wprowadzona przez Rao (1945, 1953) przy okazji klasyfikacji genetycznie różnych grup, a następnie Arnold (1979) badał tę strukturę kowariancji przy opracowywaniu uogólnionego modelu liniowego z wymienialnym wektorem reszt o łącznym rozkładzie normalnym.

1. Wprowadzenie

Rozważane będą podwójnie wielowymiarowe dane następującej postaci:

$$\underset{num\times 1}{\boldsymbol{y}} = vec(\underset{um\times n}{\boldsymbol{Y}'}) \sim N\Big((\mathbf{1}_n \otimes \boldsymbol{I}_{um})\boldsymbol{\mu}, \boldsymbol{I}_n \otimes \boldsymbol{\Gamma}_{um}\Big),$$

gdzie $\boldsymbol{\mu}$ i $\boldsymbol{\Gamma}$ są nieznane, $\boldsymbol{\mu}$ jest wektorem wymiaru $um \times 1$ oraz

$$\boldsymbol{\Gamma} = I_u \otimes \boldsymbol{\Gamma}_0 + (\boldsymbol{J}_u - \boldsymbol{I}_u) \otimes \boldsymbol{\Gamma}_1.$$

Poprzez $\boldsymbol{I}$ oznaczono macierz identycznościową wymiaru u, a $\boldsymbol{J} = \mathbf{11}'$ jest macierzą jedynek wymiaru u. $\boldsymbol{\Gamma}_0$ i $\boldsymbol{\Gamma}_1$ są nieznanymi parametrami wymiaru $m \times m$.

2. Najlepsze nieobciążone estymatory dla $\mathbf{\Gamma}_0$ oraz $\mathbf{\Gamma}_1$

Można pokazać, że najlepsze nieobciążone estymatory (BUE) dla $\mathbf{\Gamma}_0$ i $\mathbf{\Gamma}_1$ są postaci:

$$\widetilde{\mathbf{\Gamma}}_0 = \frac{1}{(n-1)u}\boldsymbol{C}_0,$$
$$\widetilde{\mathbf{\Gamma}}_1 = \frac{1}{(n-1)u(u-1)}\boldsymbol{C}_1,$$

gdzie

$$\boldsymbol{C}_0 = \sum_{s=1}^{u}\sum_{r=1}^{n}\left(\boldsymbol{y}_{r,s} - \overline{\boldsymbol{y}}_{\bullet s}\right)\left(\boldsymbol{y}_{r,s} - \overline{\boldsymbol{y}}_{\bullet s}\right)',$$
$$\boldsymbol{C}_1 = \sum_{s=1}^{u}\sum_{\substack{s^*=1 \\ s\neq s^*}}^{u}\sum_{r=1}^{n}\left(\boldsymbol{y}_{r,s} - \overline{\boldsymbol{y}}_{\bullet s}\right)\left(\boldsymbol{y}_{r,s^*} - \overline{\boldsymbol{y}}_{\bullet s^*}\right)'.$$

3. Testy dla komponentów wariancyjnych oparte na JNNE

$$H_0 : \sigma_i^2 = 0 \;\; vs. \;\; H_1 : \sigma_i^2 \neq 0$$

Niech $y'Ay$ będzie nieobciążonym estymatorem σ_i^2. Ponadto, niech A_+, A_- oznacza odpowiednio dodatnią i ujemną część macierzy A, tzn. $A = A_+ - A_-$.

Uwaga. *Dla $i < k$ estymator $y'Ay$ jest nieokreślony, tj. $A = A_+ - A_-$, gdzie $A_+, A_- \neq 0$. Zauważmy, że*

- *jeśli H_0 jest prawdziwa, to $E\left(y'A_+y\right) = E\left(y'A_-y\right)$,*
- *jeśli H_1 jest prawdziwa, to $E\left(y'A_+y\right) > E\left(y'A_-y\right)$.*

3. Testy dla komponentów wariancyjnych oparte na JNNE

Test powinien odrzucać hipotezę (intuicyjnie)

$$H_0 : \sigma_i^2 = 0$$

jeśli

$$F = \frac{y' A_+ y}{y' A_- y}$$

jest dostatecznie duże.

3. Testy dla komponentów wariancyjnych oparte na JNNE

Definicja. *Algebrę nazywamy Jordana jeśli jest zamknięta ze względu na działanie* $A \circ B = \frac{AB+BA}{2}$.

Pełną charakteryzację nieredukowalnych algebr Jordana podali autorzy Jordan, Neumann i Wigner (1934):

- $\mathbb{R}$ z działaniami mnożenia i dodawania;
- S^n - zbiór macierzy symetrycznych z działaniami $A \circ B$;
- kwaterniony;
- specjalna algebra.

3. Testy dla komponentów wariancyjnych oparte na JNNE

Model normalny jest postaci:

$$y \sim N\left(X\beta, \sum_{i=1}^{k} \sigma_i^2 V_i\right)$$

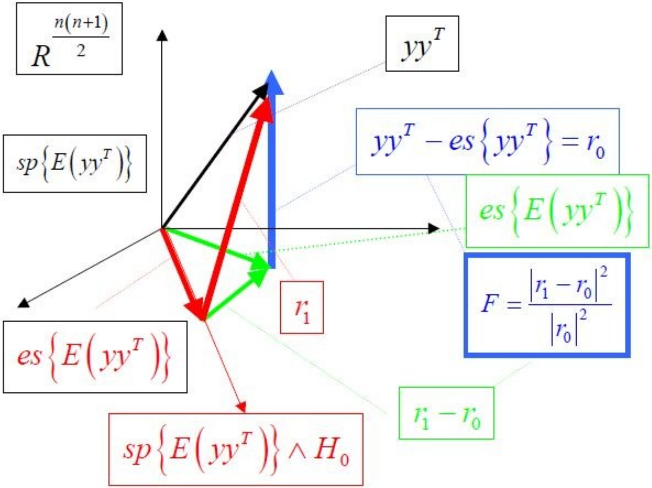
Rozważmy trzy warunki:

1. $sp\{MV_1M, \ldots, MV_kM\}$ jest algebrą Jordana,
2. $sp\{\{MV_1M, \ldots, MV_kM\} \setminus \{MV_iM\}\}$ jest przemienną algebrą Jordana,
3. $F = \frac{y'A_+y}{y'A_-y}$ ma rozkład F-Snedecora przy prawdziwości $H_0 : \sigma_i^2 = 0$.

Twierdzenie (1996): 1. $\wedge$ 2. $\Rightarrow$ 3.

Twierdzenie (2002): 1. $\wedge$ 3. $\Rightarrow$ 2.

Test F dla hipotezy $H_0 : \sigma_i = 0$



Długość wektora r_0 nie jest funkcją minimalnych statystyk dostatecznych.

3. Testy dla komponentów wariancyjnych oparte na JNNE

Twierdzenie. *Załóżmy, że podprzestrzeń*

$$sp\,\{MV_1M, \ldots, MV_kM\}$$

jest przemienną algebrą Jordana, natomiast

$$sp\,\{\{MV_1M, \ldots, MV_kM\} \setminus \{MV_iM\}\}$$

nią nie jest. Wówczas statystyka

$$F = \frac{y'A_+y}{y'A_-y}$$

ma uogólniony rozkład F-Snedecora przy prawdziwości $H_0 : \sigma_i^2 = 0$*, gdzie* $y'Ay$ *jest BIQUE parametru* σ_i^2*.*

4. Hipotezy dotyczące struktury kowariancji dla podwójnie wielowymiarowych danych

$$H_0 : \mathbf{\Gamma}_1 = 0 \;\; vs. \;\; H_1 : \mathbf{\Gamma}_1 \neq 0.$$

Wykorzystując ideę zaprezentowaną w [A. Michalski, R. Zmyślony, *Testing hypotheses for variance components in mixed linear models*, Statistics 27(1996), 297-310] do testowania hipotez przy założeniu, że wszystkie elementy macierzy $\mathbf{\Gamma}_1$ są nieujemne bądź niedodatnie.

Lemat. *Jeśli* $\boldsymbol{W}_1 \sim W_m(\mathbf{\Sigma}, n_1)$ *i* $\boldsymbol{W}_2 \sim W_m(\mathbf{\Sigma}, n_2)$ *są niezależne to dla każdego ustalonego wektora* $\boldsymbol{x} \neq 0 \in \mathbb{R}^m$ *zachodzi:*

$$T = \frac{\boldsymbol{x}'\boldsymbol{W}_1\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{W}_2\boldsymbol{x}} \sim F_{n_1,n_2}.$$

Dowód. Zgodnie z twierdzeniem zaprezentowanym w [Mirosław Krzyśko, *Podstawy wielowymiarowego wnioskowania statystycznego*, UAM (2009), 58] jeśli $\boldsymbol{W} \sim W_m(\boldsymbol{\Sigma}, n)$ to dla każdego $\boldsymbol{x} \neq 0 \in \mathbb{R}^m$ zachodzi:

$$\frac{\boldsymbol{x}'\boldsymbol{W}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}} \sim \chi^2_n.$$

Jeśli teraz obliczy się iloraz $\frac{\boldsymbol{x}'\boldsymbol{W}_1\boldsymbol{x}}{n_1}$ i $\frac{\boldsymbol{x}'\boldsymbol{W}_2\boldsymbol{x}}{n_2}$ to dostajemy:

$$\frac{\frac{\boldsymbol{x}'\boldsymbol{W}_1\boldsymbol{x}}{n_1}}{\frac{\boldsymbol{x}'\boldsymbol{W}_2\boldsymbol{x}}{n_2}} = \frac{\frac{\boldsymbol{x}'\boldsymbol{W}_1\boldsymbol{x}}{n_1\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}}}{\frac{\boldsymbol{x}'\boldsymbol{W}_2\boldsymbol{x}}{n_2\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}}} \sim \frac{\frac{\chi^2_{n_1}}{n_1}}{\frac{\chi^2_{n_2}}{n_2}} \sim F_{n_1,n_2}.$$

$\square$

5. Dodatnia i ujemna część estymatora $\widetilde{\mathbf{\Gamma}}_1$

Wykorzystując wyniki otrzymane w [A. Roy, R. Leiva, I. Žežula, D. Klein, *Testing the equality of mean vectors for paired doubly multivariate observations in blocked compound symmetric covariance matrix setup*, Journal of Multivariate Analysis, 137, 50-60] otrzymujemy, że macierze:

$$(n-1)(u-1)\widetilde{\mathbf{\Delta}}_1 = (n-1)(u-1)\left(\widetilde{\mathbf{\Gamma}}_0 - \widetilde{\mathbf{\Gamma}}_1\right)$$

oraz

$$(n-1)\widetilde{\mathbf{\Delta}}_2 = (n-1)\left(\widetilde{\mathbf{\Gamma}}_0 + (u-1)\widetilde{\mathbf{\Gamma}}_1\right)$$

mają rozkład Wisharta i są niezależne.

5. Dodatnia i ujemna część estymatora $\widetilde{\mathbf{\Gamma}}_1$

Warto zwrócić uwagę na to, że:

$$(n-1)(u-1)\widetilde{\mathbf{\Delta}}_1 = (n-1)(u-1)(\widetilde{\mathbf{\Gamma}}_0 - \widetilde{\mathbf{\Gamma}}_1) \sim W_m(\mathbf{\Gamma}_0 - \mathbf{\Gamma}_1, (n-1)(u-1)),$$

$$(n-1)\widetilde{\mathbf{\Delta}}_2 = (n-1)(\widetilde{\mathbf{\Gamma}}_0 + (u-1)\widetilde{\mathbf{\Gamma}}_1) \sim W_m(\mathbf{\Gamma}_0 + (u-1)\mathbf{\Gamma}_1, n-1).$$

Odejmując lewe strony drugiej z powyższych równości pomnożonej przez $u-1$ od pierwszej, oraz dokonując kilku prostych przekształceń dostajemy, że:

$(n-1)(u-1)\widetilde{\mathbf{\Delta}}_2-(n-1)(u-1)\widetilde{\mathbf{\Delta}}_1 = (n-1)(u-1)\widetilde{\mathbf{\Gamma}}_0+(n-1)(u-1)^2\widetilde{\mathbf{\Gamma}}_1-(n-1)(u-1)\widetilde{\mathbf{\Gamma}}_0+(n-1)(u-1)\widetilde{\mathbf{\Gamma}}_1 = (n-1)(u-1)[u-1+1]\widetilde{\mathbf{\Gamma}}_1 = (n-1)u(u-1)\widetilde{\mathbf{\Gamma}}_1.$

Zatem $\widetilde{\mathbf{\Gamma}}_1 = \frac{(n-1)(u-1)\widetilde{\mathbf{\Delta}}_2-(n-1)(u-1)\widetilde{\mathbf{\Delta}}_1}{(n-1)u(u-1)} = \frac{\widetilde{\mathbf{\Delta}}_2-\widetilde{\mathbf{\Delta}}_1}{u}$.

5. Dodatnia i ujemna część estymatora $\widetilde{\mathbf{\Gamma}}_1$

Wykorzystując podejście Michalskiego i Zmyślonego, łatwo zauważyć, że estymator $\widetilde{\mathbf{\Gamma}}_1$ można wyrazić jako różnicę jego części dodatniej i ujemnej, a więc:

$$\widetilde{\mathbf{\Gamma}}_1 = \frac{\widetilde{\mathbf{\Delta}}_2 - \widetilde{\mathbf{\Delta}}_1}{u} = \widetilde{\mathbf{\Gamma}}_{1+} - \widetilde{\mathbf{\Gamma}}_{1-},$$

gdzie $\widetilde{\mathbf{\Gamma}}_{1+} = \frac{\widetilde{\mathbf{\Delta}}_2}{u}$ to dodatnia część estymatora, natomiast $\widetilde{\mathbf{\Gamma}}_{1-} = \frac{\widetilde{\mathbf{\Delta}}_1}{u}$ to jego ujemna część.

6. Postać testu F

Twierdzenie. *Z faktu, że estymator parametru* $\mathbf{\Gamma}_1$ *jest dany poprzez:*

$$\widetilde{\mathbf{\Gamma}}_1 = \widetilde{\mathbf{\Gamma}}_{1+} - \widetilde{\mathbf{\Gamma}}_{1-} = \frac{\widetilde{\mathbf{\Delta}}_2 - \widetilde{\mathbf{\Delta}}_1}{u},$$

to statystyka testowa testu F ma postać:

$$T = \frac{\mathbf{1}'\widetilde{\mathbf{\Gamma}}_{1+}\mathbf{1}}{\mathbf{1}'\widetilde{\mathbf{\Gamma}}_{1-}\mathbf{1}}$$

i przy prawdziwości hipotezy zerowej $H_0 : \mathbf{\Gamma}_1 = 0$ *ma rozkład F-Snedecora z* $n-1$ *i* $(n-1)(u-1)$ *stopniami swobody.*

6. Postać testu F

Dowód. Przy prawdziwości hipotezy zerowej $\mathbf{\Gamma}_1 = 0$ zachodzi, że:

$$(n-1)u\widetilde{\mathbf{\Gamma}}_{1+} = (n-1)\widetilde{\mathbf{\Delta}}_2 \sim W_m(\mathbf{\Gamma}_0, n-1),$$

$$(n-1)u(u-1)\widetilde{\mathbf{\Gamma}}_{1-} = (n-1)(u-1)\widetilde{\mathbf{\Delta}}_1 \sim W_m(\mathbf{\Gamma}_0, (n-1)(u-1)).$$

Wykorzystując powyższe równości oraz wcześniejszy lemat, można pokazać, że:

$$T = \frac{\frac{(n-1)u\mathbf{1}'\widetilde{\mathbf{\Gamma}}_{1+}\mathbf{1}}{(n-1)}}{\frac{(n-1)u(u-1)\mathbf{1}'\widetilde{\mathbf{\Gamma}}_{1-}\mathbf{1}}{(u-1)(n-1)}} = \frac{\mathbf{1}'\widetilde{\mathbf{\Gamma}}_{1+}\mathbf{1}}{\mathbf{1}'\widetilde{\mathbf{\Gamma}}_{1-}\mathbf{1}}.$$

$\square$

7. Postać testu ilorazu wiarygodności

W przypadku testu ilorazu wiarygodności (LRT) do weryfikacji hipotez postaci:

$$H_0 : \mathbf{\Gamma}_1 = 0 \quad vs. \quad H_1 : \mathbf{\Gamma}_1 \neq 0$$

nie wymaga się dodatkowo założenia odnośnie jednakowego znaku wszystkich elementów macierzy $\mathbf{\Gamma}_1$, w odróżnieniu od testu F, który de facto bada czy suma wszystkich elementów tej macierzy jest różna od zera.

Postać statystyki testowej testu LRT jest następująca:

$$L = \frac{\left| \frac{(n-1)(u-1)\widetilde{\mathbf{\Delta}}_1 + (n-1)\widetilde{\mathbf{\Delta}}_2}{nu} \right|^{-\frac{nu}{2}}}{\left| \frac{n-1}{n} \widetilde{\mathbf{\Delta}}_1 \right|^{-\frac{n(u-1)}{2}} \left| \frac{n-1}{n} \widetilde{\mathbf{\Delta}}_2 \right|^{-\frac{n}{2}}}.$$

7. Postać testu ilorazu wiarygodności

Po kilku prostych przekształceniach można pokazać, że:

$$L = \frac{\left|\frac{(u-1)\widetilde{\boldsymbol{\Delta}}_1+\widetilde{\boldsymbol{\Delta}}_2}{u}\right|^{-\frac{nu}{2}}}{\left|\widetilde{\boldsymbol{\Delta}}_1\right|^{-\frac{n(u-1)}{2}}\left|\widetilde{\boldsymbol{\Delta}}_2\right|^{-\frac{n}{2}}}.$$

Przy prawdziwości hipotezy zerowej $H_0 : \boldsymbol{\Gamma}_1 = 0$ statystyka $-2\ln(L)$ ma graniczny rozkład chi-kwadrat z $\frac{m(m+1)}{2}$ stopniami swobody.

8. Porównanie mocy testów: F i LRT

W celu porównania mocy obu testów przyjęto, że macierze $\mathbf{\Gamma}_0$ oraz $\mathbf{\Gamma}_1$ mają następującą postać:

$$\mathbf{\Gamma}_0 = \begin{bmatrix} 0.01221 & 0.02172 & 0.00901 \\ 0.02172 & 0.07492 & 0.01682 \\ 0.00901 & 0.01682 & 0.01108 \end{bmatrix},$$

$$\mathbf{\Gamma}_1 = \begin{bmatrix} 0.01038 & 0.01931 & 0.00824 \\ 0.01931 & 0.06678 & 0.01529 \\ 0.00824 & 0.01529 & 0.00807 \end{bmatrix}.$$

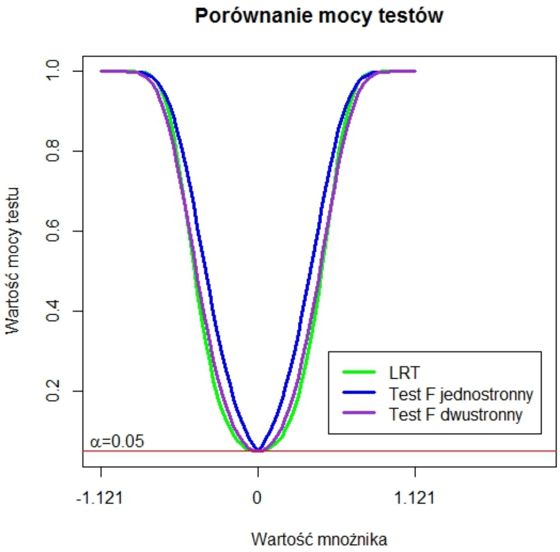
Dodatkowo założono, że $u = 2$ oraz $n = 25$.

8. Porównanie mocy testów: F i LRT

Dla ustalonej postaci macierzy $\mathbf{\Gamma}_0$ i $\mathbf{\Gamma}_1$ oraz wartości u, ustalono przedział dla wartości mnożnika λ, taki, że dla każdej wartości z tego przedziału poniższe dwa warunki są spełnione:

1. $\mathbf{\Gamma}_0 + (u-1)\lambda\mathbf{\Gamma}_1$ jest dodatnio określoną macierzą,
2. $\mathbf{\Gamma}_0 - \lambda\mathbf{\Gamma}_1$ jest dodatnio określoną macierzą.

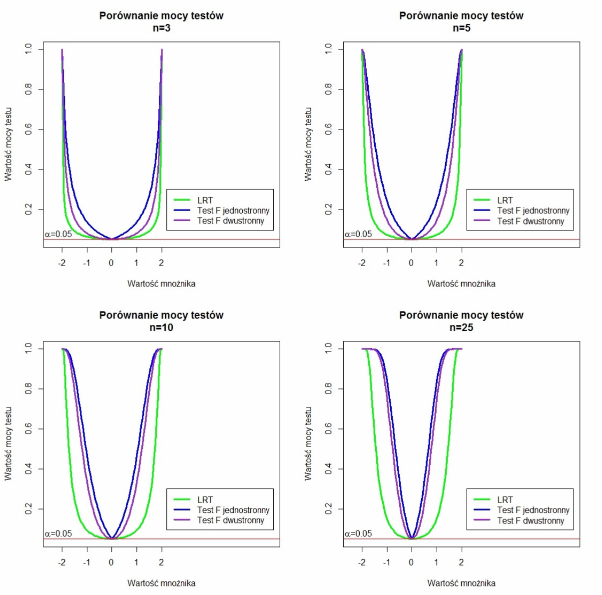
Ich spełnienie gwarantuje dodatnią określoność macierzy $\mathbf{\Gamma}$.
Przyjęty poziom istotności $\alpha = 5\%$.


Porównanie mocy testów
Wartość mocy testu
Wartość mnożnika
0.2
0.4
0.6
0.8
1.0
-1.121
0
1.121
α=0.05
LRT
Test F jednostronny
Test F dwustronny

8. Porównanie mocy testów: F i LRT

Rozważany będzie jeszcze jeden przypadek. Tym razem bardzo szczególny, gdyż mianowicie Γ_0 i Γ_1 będą skalarami, a więc $m = 1$. Niech $\Gamma_0 = 2$ oraz $\Gamma_1 = 1$. Dodatkowo przyjęto, że $u = 2$, natomiast parametr n będzie przyjmować jedną z wartości ze zbioru $\{3, 5, 10, 25\}$.

Z warunków na dodatnią określoność macierzy $\mathbf{\Gamma}$ nietrudno wywnioskować, że odpowiednie wartości mnożnika λ powinny zawierać się w przedziale $[-2, 2]$.



Porównanie mocy testów
n=3
Porównanie mocy testów
n=5
Porównanie mocy testów
n=10
Porównanie mocy testów
n=25
Wartość mocy testu
Wartość mnożnika
LRT
Test F jednostronny
Test F dwustronny
α=0.05

References

[1] Arnold, S.F. 1979. Linear models with exchangeably distributed errors. Journal of the American Statistical Association. 74, 194-199.

[2] Drygas, H., 1970. The Coordinate-Free Approach to Gauss-Markov Estimation, Berlin, Heidelberg: Springer.

[3] Gąsiorek E., Michalski A., Zmyślony R. 2000, Tests of independence of normal random variables with known and unknown variance ratio, Discussiones Mathematicae Probability and Statistics 20 2, 233-247.

[4] Gnot, S., Klonecki, W. and Zmyślony, R. 1976. Uniformly minimum variance unbiased estimation in Euclidean vector spaces, Bull. de l.Academie Polon. des Sciences XXIV 4, 281-286.

[5] Gnot, S., Klonecki, W. and Zmyślony, R. 1977a. Uniformly minimum variance unbiased estimation in various classes of estimators. Statistics 8(2), 199-210.

[6] Gnot, S., Klonecki, W. and Zmyślony, R. 1977b. Uniformly minimum variance unbiased estimation in various classes of estimators, Math. Operationsforsch. Statist.,Ser. Statistics 8 2, 199-210.

[7] Gnot, S., Klonecki, W. and Zmyślony, R. 1980. Best unbiased estimation: a coordinate free-approach. Probability and Statistics, 1(1), 1-13.

[8] Jordan, P., Neumann, von, J. and Wigner, E., 1934. On an algebraic generalization of the quantum mechanical formalism. The Annals of Mathematics, 35(1), 29-64.

[9] Kruskal, W., 1968. When are Gauss-Markov and Least Squares Estimators Identical? A Coordinate-Free Approach. The Annals of Mathematical Statistics, 39(1), pp.70-75.

[10] Lehmann, E.L. and Casella, G., 1998. Theory of Point Estimation Second Edition, Springer.

[11] Roy, A. and Leiva, R. 2008. Likelihood ratio tests for triply multivariate data with structured correlation on spatial repeated measurements. Statistics & Probability Letters, 78(13), 1971-1980.

[12] Seely, J.F., 1971. Quadratic subspaces and completeness. The Annals of Mathematical Statistics, 42(2), 710-721.

[13] Seely, J.F., 1972. Completeness for a family of multivariate normal distributions. The Annals of Mathematical Statistics, 43, 1644-1647.

[14] Seely, J.F., 1977. Minimal sufficient statistics and completeness for multivariate normal families. Sankhya (Statistics). The Indian Journal of Statistics. Series A, 39(2), 170-185.

[15] Zmyślony, R. 1976. On estimation of parameters in linear models, Applicationes Mathematicae XV 3(1976), 271-276.

[16] Zmyślony, R. 1978. A characterization of best linear unbiased estimators in the general linear model, Lecture Notes in Statistics, 2, 365-373.

[17] Zmyślony, R. 1980. Completeness for a family of normal distributions, Mathematical Statistics, Banach Center Publications 6, 355-357.