ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN—POLONIA
VOL. XV/XVI, 13 SECTIO H 1981/82

Zakład Statystyki, Ekonometrii i Informatyki
Wydziału Ekonomicznego UMCS

Kazimierz KACPRZAK

Idea zastosowania korelacji kanonicznej do wyboru optymalnego zbioru zmiennych objaśniających do modelu ekonometrycznego

Идея применения канонической корреляции для выбора оптимального множества объяснимых переменных в эконометрической модели

The Application of Canonical Correlation to the Selection of an Optimal Set of Explanatory Variables for an Econometric Model

UWAGI WSTĘPNE

Analiza kanoniczna stanowi jeden z elementów wielowymiarowej analizy statystycznej. Ogólnie można stwierdzić, że polega ona na badaniu związku pomiędzy dwoma układami (wektorami) zmiennych, przy czym jeden układ tworzą zmienne objaśniane, a drugi — zmienne objaśniające. Wynika z tego, że analizę kanoniczną można traktować jako uogólnienie regresji wielokrotnej, w której zmienność zmiennej objaśnianej można wyjaśnić zmiennością zespołu zmiennych objaśniających.

Pojęcie zmiennych kanonicznych i korelacji kanonicznej wprowadził do literatury statystyczno-ekonometrycznej H. Hotelling w r. 1936, a szerokie podstawy teorii korelacji kanonicznej dał T. W. Anderson w r. 1958.[1] Problem ten omówiony jest również przez wielu autorów (przede wszystkim zachodnich) w pracach dotyczących wielowymiarowej analizy statystycznej. Autorzy, do których między innymi należą: Cooley, Lohnes, Harris, Kendall, Stuart i Rao podali nie tylko teoretyczne aspekty korelacji kanonicznej, ale również praktyczne jej zastosowania. Główne dzie-

[1] T. W. Anderson: *An Introduction to Multivariate Statistical Analysis*, Wiley, New York 1958, s. 288—306.

dziny dotychczasowych zastosowań analizy kanonicznej to: psychologia, geografia, antropologia, botanika, nauki rolnicze i ekonomiczne.[2]

Zainteresowanie w naszym kraju analizą kanoniczną i jej wykorzystaniem do badań empirycznych datuje się od drugiej połowy lat siedemdziesiątych. Teoria tej problematyki omówiona jest między innymi w opracowaniach A. Krzyśki[3] i M. Nowosadzkiego, natomiast wyniki praktycznych zastosowań zawarte są w pracach B. Głębockiego[4], S. Mejzy[5] i W. Ratajczaka[6], dotyczących badań produkcji rolniczej, zootechnicznych i w geografii ekonomicznej. W pracach tych zastosowanie analizy kanonicznej pozwoliło zbadać związki pomiędzy zmiennymi mierzącymi poziom urbanizacji a zmiennymi mierzącymi poziom uprzemysłowienia w układzie gmin województwa poznańskiego, jak również współzależności pomiędzy rozwojem ekonomicznym, środowiskiem geograficznym i kształtem powiatów województwa poznańskiego a rozwojem ich sieci drogowej i kolejowej.

Niniejsze opracowanie nie opiera się na badaniach empirycznych. Ma ono charakter teoretyczny. Celem tego opracowania jest przedstawienie możliwości wykorzystania korelacji kanonicznej do wyboru optymalnego zbioru zmiennych objaśniających do modelu ekonometrycznego.

Możliwość zastosowania korelacji kanonicznej do wyboru zmiennych objaśniających do modelu ekonometrycznego w początkowej fazie jego budowy zaproponował J. Greń[7]. Podał on ogólną ideę tej metody w wymienionym aspekcie. W niniejszym opracowaniu — poza przypomnieniem propozycji J. Grenia — przedstawiono dalsze uwagi dotyczące uzyskania ostatecznego rozwiązania, tzn. ustalenia zbioru zmiennych objaśniających do modelu ekonometrycznego.

Załóżmy, że dysponujemy dużym zbiorem potencjalnych zmiennych, które można by użyć w modelu jako zmienne objaśniające. Nie chcemy jednak wprowadzać do modelu wszystkich zmiennych potencjalnych (zda-

[2] M. Nowosadzki: *Analiza kanoniczna i analiza redundacji*, Piąte Colloquium Metodologiczne z Agro-biometrii, Warszawa 1975, s. 230—252.

[3] M. Krzyśko: *Analiza zmiennych kanonicznych i korelacji kanonicznych* [w:] *Analiza regresji w geografii*, pr. zb. pod red. Z. Chojnickiego, PAN, Warszawa—Poznań 1980, s. 55—68.

[4] B. Głębocki: *Czynniki kształtujące przestrzenną strukturę produkcyjną rolnictwa*, Uniwersytet im. A. Mickiewicza, Poznań 1979.

[5] S. Mejza: *Korelacje kanoniczne i ich zastosowania w badaniach rolniczych*, Piąte Colloquium Metodologiczne z Agro-Biometrii, PAN, 1975, s. 254—274.

[6] W. Ratajczak: *Zastosowanie analizy kanonicznej w badaniach geograficznych*, pr. zbiorowa pod red. Z. Chojnickiego nt. „Analiza regresji w geografii", PAN, Warszawa—Poznań, 1980, s. 69—81.

[7] Propozycja ta została zgłoszona na seminarium naukowym poświęconym problemowi doboru zmiennych do modelu, które odbyło się w Zakopanem w kwietniu 1979 r.

rza się, że nadmierna liczba zmiennych objaśniających występująca w modelu poza kłopotami natury numerycznej utrudnia merytoryczne zinterpretowanie uzyskanych wyników). Musimy więc dokonać wyboru zmiennych spośród wszystkich kandydatek.

Zbiór zmiennych oznaczmy przez χ, natomiast zbiór zmiennych, które ostatecznie wejdą do modelu przez χ_A, a zbiór zmiennych pominiętych — przez χ_B. Zmienne ze zbioru χ_A będziemy nazywać zmiennymi aktywnymi, zaś zmienne ze zbioru χ_B — zmiennymi biernymi. Na tej podstawie zbiór zmiennych potencjalnych można zapisać jako sumę podzbiorów χ_A i χ_B, czyli:

$$\chi = \chi_A \cup \chi_B$$

gdzie: $\chi_A = \{X_i,\ i \in A\}$, $\chi_B = \{X_j,\ j \in B\}$.

Problem więc sprowadza się do odpowiedniego podziału zbioru χ na podzbiory χ_A i χ_B. Podział ten powinien być jednak tak dokonany, aby wybrane zmienne do modelu najlepiej wyjaśniały zmienność zmiennej objaśnianej. Co więcej — ze względu na brak dokładnego rozeznania, które ze zmiennych zbioru χ bezwzględnie powinny w modelu wystąpić — nie chcemy całkowicie rezygnować z wpływu zmiennych pomijanych. Wymagamy więc, aby zmienne podzbioru χ_A, poza informacjami, jakie same wnoszą do modelu, reprezentowały również informacje pochodzące od zmiennych pomijanych. Wydaje się, że odpowiedniego podziału zbioru χ na podzbiory χ_A i χ_B można dokonać przez wykorzystanie teorii korelacji kanonicznej.

KORELACJA KANONICZNA

Rozważmy wektor $\mathbf{x}$ zmiennych o $i+j$ składowych oraz podwektory $\mathbf{x}_A = [x_i]$ i $\mathbf{x}_B = [x_j]$ Utwórzmy dwie zmienne sztuczne u_A i v_B, będące kombinacjami liniowymi elementów wektorów $\mathbf{x}_A$ i $\mathbf{x}_B$, co można zapisać następująco:

$$\begin{aligned} u_A &= \sum_{i \in A} q_i x_i = \mathbf{q}^T \mathbf{x}_A \\ v_B &= \sum_{j \in B} h_j x_j = \mathbf{h}^T \mathbf{x}_B \end{aligned} \tag{2.1}$$

gdzie: $\mathbf{q} = [q_i]$, $\mathbf{h} = [h_j]$ — współczynniki powyższych kombinacji liniowych będą tak dobrane, aby współczynnik korelacji pomiędzy zmiennymi u_A i v_B był maksymalny.

Dla uzyskania jednoznacznych rozwiązań numerycznych wprowadza się dodatkowy warunek, a mianowicie taki, żeby współczynniki q_i i h_j

były tak dobrane, aby wariancje zmiennych u_A i v_B równały się jedności, czyli:

$$D^2(u_A)=1 \text{ i } D^2(v_B)=1 \tag{2.2}$$

Współczynnik korelacji pomiędzy zmiennymi u_A i v_B oznaczony przez $\varrho u_A v_B$ można wtedy wyrazić następująco:

$$\rho_{u_A v_B} = \frac{\operatorname{cov}(u_A, v_B)}{\sqrt{D^2(u_A) D^2(v_B)}} = \operatorname{cov}(u_A, v_B). \tag{2.3}$$

Zdefiniowane wzorem (2.1) zmienne u_A i v_B nazywamy zmiennymi kanonicznymi, a współczynnik korelacji pomiędzy tymi zmiennymi określony wzorem (2.3) nazywamy współczynnikiem korelacji kanonicznej. Współczynnik ten mierzy siłę związku pomiędzy zmiennymi kanonicznymi. Maksymalizując go chcemy zapewnić sobie wprowadzenie do modelu informacji nie tylko reprezentowanych przez zmienne, które zostaną w modelu uwzględnione, ale również — przez silne skorelowanie ich ze zmiennymi pomijanymi — informacje pochodzące od zmiennych nie uwzględnionych w modelu. Współczynnik $\varrho u_A v_B$ jest bowiem kryterium wyboru odpowiedniego podziału zbioru χ na podzbiory χ_A i χ_B, umożliwiającym ustalenie listy zmiennych objaśniających, które powinny wystąpić w modelu. Dokładniej problem wyboru omówiono w punkcie 3 niniejszego opracowania.

Obecnie przedstawimy proces wyznaczania maksymalnego współczynnika korelacji kanonicznej dla jednego z m możliwych podziałów zbioru χ na podzbiory χ_A i χ_B.

Jeżeli dysponujemy macierzą $\mathbf{x}$ obserwacji na zmiennych potencjalnych i r-tym podziałem tej macierzy na bloki $\mathbf{x}_A$ i $\mathbf{x}_B$ oraz wektorami zmiennych kanonicznych tego podziału, to za Theilem możemy podać, że [8]:

$$\begin{aligned} D^2(u_A^{(r)}) &= u_A^{(r)T} u_A^{(r)} = q_r^T X_A^T X_A q_r = 1 \\ D^2(v_B^{(r)}) &= v_B^{(r)T} v_B^{(r)} = h_r^T X_B^T X_B h_r = 1 . \end{aligned} \tag{2.4}$$

Natomiast współczynnik korelacji kanonicznej można zapisać następująco:

$$\rho_{u_A^{(r)} v_B^{(r)}} = u_A^{(r)T} v_B^{(r)} = q_r^T X_A^T X_B h_r . \tag{2.5}$$

Aby otrzymać maksymalny współczynnik $\varrho u_A^{(r)} v_B^{(r)}$, należy zmaksymalizować prawą stronę wyrażenia (2.5) przy warunkach (2.4). Problem

[8] H. Theil: *Zasady ekonometrii*, PAN, Warszawa 1979, s. 323.

ten — jak wiadomo — jest poszukiwaniem maksimum warunkowego funkcji Lagrange'a, która w naszym wypadku przyjmuje następującą postać:

$$F(\mathbf{q}_r, \mathbf{h}_r) = \mathbf{q}_r^T \mathbf{X}_A^T \mathbf{X}_B \mathbf{h}_r - \frac{1}{2}\lambda(\mathbf{q}_r^T \mathbf{X}_A^T \mathbf{X}_A \mathbf{q}_r - 1) - \frac{1}{2}\mu(\mathbf{h}_r^T \mathbf{X}_B^T \mathbf{X}_B \mathbf{h}_r - 1) \quad (2.6)$$

gdzie: λ i μ są mnożnikami Lagrange'a.

Obliczając pochodne cząstkowe funkcji (2.6) względem wektorów $\mathbf{q}_r$ i $\mathbf{h}_r$ i przyrównując je do wektora zerowego otrzymujemy:

$$\begin{aligned} \frac{\partial F}{\partial \mathbf{q}_r} &= \mathbf{X}_A^T \mathbf{X}_B \mathbf{h}_r - \lambda \mathbf{X}_A^T \mathbf{X}_A \mathbf{q}_r = \mathbf{o} \\ \frac{\partial F}{\partial \mathbf{h}_r} &= \mathbf{X}_B^T \mathbf{X}_A \mathbf{q}_r - \mu \mathbf{X}_B^T \mathbf{X}_B \mathbf{h}_r = \mathbf{o}. \end{aligned} \quad (2.7)$$

Wykorzystując (2.4) i (2.7), można wykazać, że:

$$\lambda = \mu = \rho_{u_A(r) v_B(r)} . \quad (2.8)$$

Z kolei wykorzystując (2.7) i (2.8) i dokonując odpowiednich przekształceń, otrzymujemy dwa równania o następującej postaci:

$$\begin{aligned} [(\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \mathbf{X}_B (\mathbf{X}_B^T \mathbf{X}_B)^{-1} \mathbf{X}_B^T \mathbf{X}_A - \rho^2_{u_A(r) v_B(r)} \mathbf{I}]_{\mathbf{q}_r} &= \mathbf{o} \\ [(\mathbf{X}_B^T \mathbf{X}_B)^{-1} \mathbf{X}_B^T \mathbf{X}_A (\mathbf{X}_A^T \mathbf{X}_A)^{-1} \mathbf{X}_A^T \mathbf{X}_B - \rho^2_{u_A(r) v_B(r)} \mathbf{I}]_{\mathbf{h}_r} &= \mathbf{o}. \end{aligned} \quad (2.9)$$

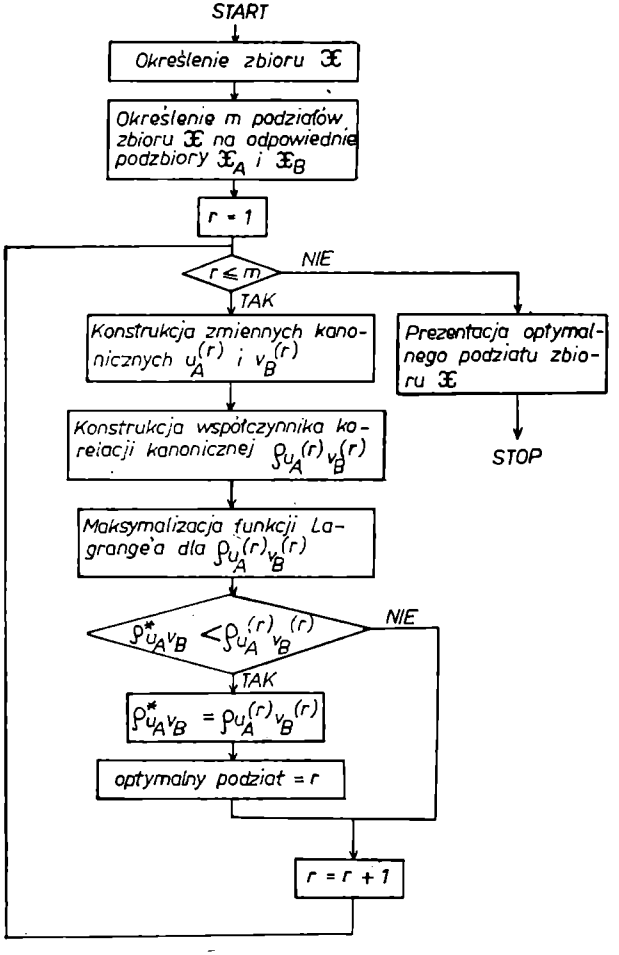
Jak wynika z (2.9) $\varrho^2 u_A(r) v_B(r)$ jest wartością własną odpowiednich macierzy — wartością, której pierwiastek jest współczynnikiem korelacji kanonicznej. Natomiast $\mathbf{q}_r$ i $\mathbf{h}_r$ są wektorami własnymi tych samych, odpowiednich macierzy. Aby więc uzyskać największy współczynnik korelacji kanonicznej, wybieramy największy pierwiastek wielomianu charakterystycznego, występującego w równaniu charakterystycznym, które jest wyznacznikiem (2.9) porównanym do zera. Największemu pierwiastkowi przyporządkowane będą odpowiednie wektory spełniające warunek (2.2).

PROCEDURA ZASTOSOWANIA KORELACJI KANONICZNEJ DO WYBORU ZMIENNYCH OBJAŚNIAJĄCYCH

Cały proces wykorzystania analizy kanonicznej do wyboru zmiennych objaśniających można przedstawić w postaci poniższego schematu blokowego. Schemat ten przedstawia kolejność czynności zmierzających do

wyboru ostatecznego, optymalnego podziału zbioru χ na podzbiory χ_A i χ_B.

Jak wynika ze schematu przedstawionego na rycinie, poszukiwanie maksymalnego współczynnika korelacji kanonicznej przebiega przez wszystkie $r \leq m$ podziałów zbioru χ na odpowiednie podzbiory χ_A i χ_B. Należy jednak podkreślić, że interesują nas tylko te podziały, które zapewniają co najmniej dwuelementowe podzbiory χ_A i χ_B. Takie bowiem podzbiory umożliwiają konstrukcję zmiennych kanonicznych.



Schemat blokowy wykorzystania analizy kanonicznej do wyboru zmiennych objaśniających

Block scheme of the application of canonical analysis to the selection of explanatory variables

Ostateczny wybór optymalnego podziału rozważanego zbioru następuje po zbadaniu maksymalnych współczynników korelacji kanonicznej dla m podziałów. Traktując bowiem współczynnik korelacji kanonicznej jako kryterium wyboru optymalnego podziału zbioru χ na podzbiory χ_A i χ_B, wybieramy ze wszystkich m maksymalnych współczynników korelacji ten, który jest największy. Kryterium to możemy zapisać następująco:

$$\rho^*_{u_A v_B} = \max_{A,B} \left(\max_{q,h} \rho_{u_A v_B}\right) = \max_{A,B} \rho'_{u_A v_B} \tag{3.1}$$

gdzie:

$$\rho'_{u_A v_B} = \max_{q,h} \rho_{u_A v_B}$$

Wydaje się, że powyższe kryterium maksymalnego współczynnika korelacji kanonicznej może zapewnić wybór optymalnego (najlepszego) podziału zbioru zmiennych potencjalnych na podzbiór zmiennych aktywnych i podzbiór zmiennych biernych. Należy sądzić, że zmienne ostatecznie wprowadzone do modelu w myśl kryterium (3.1) dobrze objaśniać będą zmienną objaśnianą i dobrze zastępować zmienne pominięte. Taki sposób podejścia może pozwolić na zredukowanie dużej liczby zmiennych potencjalnych do zbioru zmiennych aktywnych, zachowując jednocześnie — przez wprowadzenie analizy kanonicznej — oddziaływanie zmiennych biernych. Trzeba jednak zdawać sobie sprawę z faktu, że strona rachunkowa dojścia do ostatecznego rozwiązania jest czasochłonna i skomplikowana. Zbadanie maksimum funkcji (2.6) dla wszystkich m podziałów dużego zbioru χ zmusza do korzystania z techniki komputerowej. Ponadto wymaga również znajomości odpowiednich programów obliczeniowych. Powyższy fakt sprawia, że analiza kanoniczna budzi pewne kontrowersje. Należy jednak zaznaczyć, że obecny poziom techniki komputerowej jest taki, że nawet czasochłonne i skomplikowane numerycznie zadania mogą być zadowalająco rozwiązane, o czym świadczą cytowane w tym opracowaniu publikacje.

РЕЗЮМЕ

В статье представлена возможность применения канонической корреляции для выбора объяснимых переменных в эконометрической модели. Вступительная часть работы посвящена общим принципам деления большого множества потенциальных переменных на подмножество переменных, входящих в модель, и на подмножество пропущенных переменных.

Вторая часть работы посвящена критерию выбора соответствующего деления множества потенциальных переменных. Таким критерием есть максимальный

коэффициент канонической корреляции между двумя каноническими переменными, из которых одна является линейной комбинацией переменных, принятых во внимание в модели, а другая — линейной комбинацией пропущенных переменных. Нам кажется, что максимализируя коэффициент канонической корреляции между этими переменными, можно будет произвести такой выбор потенциальных переменных, при котором введенные в модель переменные будут хорошо объяснять объяснимые переменные. Больше того — если мы их сильно скоррелируем с пропущенными переменными, то они будут учитывать информацию, содержащуюся в переменных, не учтенных в модели.

В третьей части статьи представлена целая процедура получения окончательного, оптимального деления множества потенциальных переменных на множество переменных, введенных в модель, и на множество пропущенных переменных. Следует, однако, добавить, что процедура максимализации коэффициента канонической корреляции, являющегося критерием выбора, должна охватывать все соответствующие деления множества потенциальных переменных. Этот процесс сложен в нумерическом отношении, он требует использования компьютерной вычислительной техники.

SUMMARY

The article presents a possibility of applying canonical correlation to the selection of explanatory variables for an econometric model. The introductory section deals with the general principles of the division of a large set of potential variables into a subset of variables included in the model and a subset of variables omitted.

The second part concerns the selection criterion for an appropriate division of the potential variables set. This criterion is provided by the maximum coefficient of canonical correlation between two canonical variables, of which one is a linear combination of variables included in the model, while the other is a linear combination of variables omitted. It seems that the maximization of the canonical correlation coefficient between these variables may ensure the selection of such a division of the potential variables set that the variables introduced into the model will interpret well the variable explained. Moreover, by their strong correlation with the variables omitted, they will take into account the information contained in variables not included in the model.

The third part of the article presents the whole procedure of reaching the final, optimal division of the potential variables set into the set of variables introduced into the model and the set of variables omitted. However, it should be added that the procedure of the maximization of the canonical correlation coefficient, which is the selection criterion, must cover all the appropriate divisions of the potential variables set. It is a process numerically complex and requires the application of computer calculation techniques.