УДК 37.012:311.2

ВИКОРИСТАННЯ КАНОНІЧНОГО КОРЕЛЯЦІЙНОГО АНАЛІЗУ У ПЕДАГОГІЧНИХ ДОСЛІДЖЕННЯХ

Лупан І.В., Халецька З.П., Чеча В.О.

У статті розглянуто особливості та призначення канонічного кореляційного аналізу, а також наведено приклади його використання у педагогічних дослідженнях.
Ключові слова: канонічна кореляція, множина змінних, канонічні корені.

В статье рассмотрены особенности и назначение канонического корреляционного анализа, а также примеры его применения в педагогических исследованиях.
Ключевые слова: каноническая корреляция, множество переменных, канонические корни.

The article covers characteristic features and purpose of canonical correlation analysis and gives examples of its application in pedagogical research.
Key words: canonical correlation, set of variables, canonical roots.

Канонічний кореляційний аналіз – один із методів багатовимірного аналізу даних [6]. Це найбільш узагальнена форма аналізу кореляцій, яка дозволяє досліджувати взаємозв'язок між двома множинами змінних, на відміну від факторного аналізу, який застосовують для встановлення зв'язків усередині однієї множини змінних.

Метод канонічного аналізу відносно молодий. Уперше його ідею було опубліковано американським економістом Гарольдом Хотеллінгом (H.Hotelling) у журналі Біометрика у 1936 р. [1]. Однак активно теорія канонічного аналізу розроблялася вже у 70-ті рр. XX ст. [2, 3, 4, 5] з розвитком відповідного програмного забезпечення. На сьогоднішній день канонічний аналіз використовується у маркетингових, економічних, природничих, медичних дослідженнях [7, 8, 9, 10], проте залишається ще деякою мірою екзотичним маловикористовуваним методом, про що свідчить майже повна відсутність присвяченої йому літератури українською та російською мовами. Однак інтерес дослідників до канонічного аналізу зростає, оскільки реалізація його у відомих статистичних пакетах, зокрема у пакеті Statistica, дозволяє користуватися інструментарієм канонічного аналізу, не переймаючись обчислювальною складністю методу.

Вхідні дані для канонічного аналізу можна схематично представити як два масиви з однаковою кількістю рядків N (об'єктів дослідження) та різною (у загальному випадку, але не менше двох для кожної множини) кількістю стовпців-змінних (рис. 1).

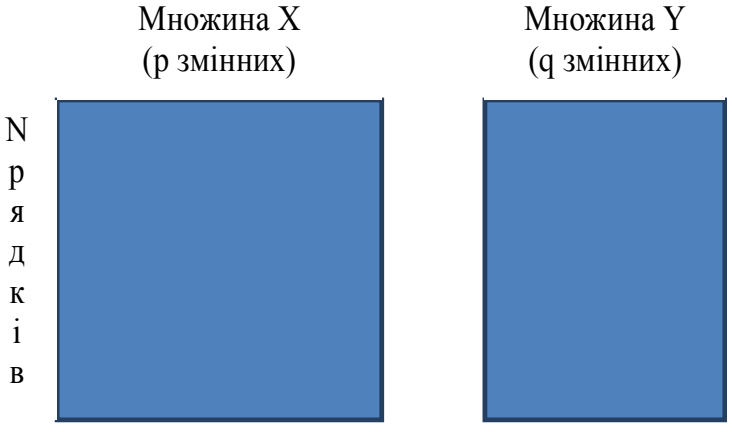

Рис. 1

У термінах математики задача полягає у пошуку лінійної комбінації (*a straight line axis*) *p* змінних та лінійної комбінації *q* змінних таким чином, щоб їхня кореляція була максимальною.

Канонічна кореляція – це кореляція між канонічними змінними U та V, де $U=a_1y_1+\ldots+a_qy_q$ та $V= b_1x_1+\ldots+b_px_p$. Тобто $r = \dfrac{\operatorname{cov}(V, U)}{\sigma_u \cdot \sigma_v}$.

Значущість канонічної кореляції визначають за критерієм Бартлетта.

Природа цієї кореляції залежить від виду (моделі) асоціації між двома множинами даних та всередині кожної множини. Ці зв'язки можна виразити, об'єднавши множини даних та обчисливши коефіцієнти кореляції (*product moment correlation coefficients*) для кожної пари змінних (рис. 2).

Змінні	X_1 ... X_p	Y_1 ... Y_q
X_1 … X_p	$\boldsymbol{R_{xx}}$	$\boldsymbol{R_{xy}}$
Y_1 … Y_q	$\boldsymbol{R_{yx}}$	$\boldsymbol{R_{yy}}$

Рис. 2

Другим кроком є складання канонічного рівняння: для пошуку загальних моделей або канонічних кореляцій спочатку слід знайти латентні корені канонічного рівняння:

$$\left|M - \lambda I\right| = 0,$$

де $M = R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}$; л – латентний корінь матриці M (вектор власних значень), який визначає розмір загальних моделей; I – одинична матриця; R_{yy}^{-1} – лінійний оператор для множини Y; R_{xx}^{-1} – лінійний оператор для множини X; R_{yx} та R_{xy} – матриці кореляцій між множинами X та Y. Добутки $R_{yy}^{-1}R_{yx}$ та $R_{xx}^{-1}R_{xy}$ утворюють матриці косинусів між двома множинами після обертання за допомогою лінійних операторів відповідно R_{yy}^{-1}та R_{xx}^{-1}.

Канонічні корені є квадратними коренями з латентних коренів: $r_i = \sqrt{\lambda_i}$.

У разі канонічного аналізу основною статистичною гіпотезою (H_0) є гіпотеза про відсутність зв'язку між двома множинами даних. Її перевіряють за допомогою функції λ-Уілкса, яка апроксимується розподілом χ^2 з *pq* степенями вільності:

$$\Lambda = \prod_{i=1}^{\min(p,q)}(1-\lambda_i), \qquad \chi^2 = -\left[(N-1)-0{,}5\,(p+q+1)\right]\ln\Lambda$$

У разі, коли нульову гіпотезу доводиться відкинути й визнати зв'язок між множинами значущим, перший латентний корінь вилучають та перевіряють значущість решти коренів.

Після визначення числа значущих канонічних коренів виникає питання про інтерпретацію кожного значущого кореня. Одним із способів тлумачення значення кожного канонічного кореня є дослідження вагових коефіцієнтів, що відповідають кожній множині даних. Їх також називають канонічними вагами.

Канонічні ваги (позначимо їх A та B) – показують внесок кожної змінної у кожну загальну модель.

Під час аналізу, зазвичай, користуються тим, що чим більша приписана вага (тобто абсолютне значенні ваги), тим більший внесок відповідної змінної у значенні канонічної змінної. Для проведення більш детального порівняльного аналізу, як правило, розглядаються стандартизовані змінні, тобто *z*-перетворені змінні з нульовим середнім і одиничним стандартним відхиленням. Дослідження канонічних ваг дозволяє зрозуміти значення кожного канонічного кореня, побачити, як конкретні змінні в кожній множині впливають на канонічну змінну:

$$(M-\lambda I)B = 0, \qquad A = \frac{R_{xx}^{-1}\cdot R_{xy}\cdot B}{\sqrt{\lambda\cdot I}}.$$

Ще одним способом інтерпретації канонічних коренів є дослідження звичайних кореляцій між канонічними змінними (або факторами) та змінними з кожної множини. Ці кореляції також називаються канонічними навантаженнями факторів. Вважається, що змінні, сильно корельовані з канонічною змінною, мають з нею багато спільного. Тому під час опису значення канонічної змінної слід виходити в основному з реального значення таких сильно корельованих змінних. Такий спосіб інтерпретації канонічних змінних схожий на метод, який використовується у факторному аналізі. Іноді канонічні ваги для змінної виявляються близькими до нуля, а відповідні їм навантаження дуже великими. Також можлива ситуація, коли навпаки канонічні ваги великі, а навантаження малі. У таких випадках висновок може виявитися досить суперечливим.

Канонічні кореляції не містять інформації про те, яку частину дисперсії кожен канонічний корінь пояснює в досліджуваних змінних. Однак, розглядаючи навантаження канонічних факторів, можна зробити висновок про частку, що пояснюється дисперсією. Якщо піднести ці кореляції до квадрату, отримані числа будуть відображати частку дисперсії, що пояснюється кожною змінною. Для кожного кореня можна обчислити середнє значення цих часток. При цьому отримаємо середню частку дисперсії, поясненої в множині на підставі відповідної канонічної змінної. Інакше кажучи, можна обчислювати середню частку дисперсії, отриманої кожним коренем. Канонічна кореляція у разі піднесення до квадрату дає частку дисперсії, загальну для сум по кожній множині. Якщо помножити цю частку на частку видобутої дисперсії, то отримаємо міру надмірності множини змінних, тобто величину, що показує, наскільки надлишковою є одна множинна змінних, якщо її співставити з іншою множиною. Відзначимо також, що можна

обчислити надлишковість першої множини змінних у разі заданої другої множини і надлишковість другої множини змінних у разі заданої першої множини. Оскільки послідовно добуті канонічні корені не корелюють між собою, то можна просто підсумувати надлишковість за усіма (або тільки за значущими) коренями, отримавши при цьому загальний коефіцієнт надлишковості.

Наведемо припущення канонічного аналізу, виконання яких забезпечує отримання достовірних і обґрунтованих результатів [2]:

· застосування критерію значущості під час аналізу канонічної кореляції засноване на припущенні, що змінні у вибірці мають багатовимірний нормальний розподіл;

· для отримання достовірних оцінок навантажень канонічних факторів рекомендують використовувати як мінімум у 20 разів більше спостережень, ніж число змінних, використаних в аналізі, якщо потрібно інтерпретувати тільки найбільш значущий корінь. Для отримання достовірних оцінок для двох канонічних коренів, рекомендують, спираючись на дослідження за допомогою методу Монте-Карло, використовувати в 40–60 разів більше спостережень, ніж число досліджуваних змінних;

· наявність викидів може мати великий вплив на значення коефіцієнтів кореляції: чим менший розмір вибірки, тим більший вплив викидів;

· ще одним припущенням є вимога, щоб змінні в обох множинах не були повністю надлишковими;

· за Дж.Стевенсом за наявності великих кореляцій між даними (наприклад, R>0.7), навіть малі розміри вибірки (наприклад, n = 50) дозволяють у більшості випадків виявити кореляції.

З метою вивчення можливостей застосування канонічного аналізу під нашим керівництвом було виконано дипломну роботу [11]. Її результати переконали в тому, що використання канонічного аналізу має перспективу і в педагогічних дослідженнях. Як приклад такого застосування нами було проведено дослідження зв'язку успішності студентів фізико-математичного факультету з дисциплін різних блоків.

Під час формування базової множини даних виникли деякі проблеми:

1) для проведення дослідження даних успішності студентів одного випуску виявилося недостатньо, довелося збирати дані за декілька років: фактично у дослідженні було використано результати успішності з усіх вивчених дисциплін випускників факультету (бакалаврів) за 2005–2011 роки (розмір вибірки склав 142 особи);

2) в аналізі були використані дані лише з тих дисциплін, з яких студенти вибірки мали екзамени. На жаль, у зв'язку зі змінами навчальних планів змінилися форми звітності з деяких дисциплін, що не дозволило використати їх у дослідженні;

3) основними недоліками базової множини змінних є те, що результати навчання студентів представлені у порядковій шкалі (у той час, як канонічний аналіз потребує числових даних) та нормальність розподілу базових змінних не завжди безсумнівна.

Остання проблема досить суттєва, однак у даному випадку довелося не брати її до уваги, оскільки нас цікавив не стільки результат канонічного аналізу, як вивчення можливостей його використання. Зауважимо лише, що коректні результати канонічний аналіз дає у разі використання великих вибірок (не менше 100 спостережень), з числовими даними (шкала вимірювання інтервальна чи відношень), які мають нормальний розподіл.

Фрагмент вхідних даних для канонічного аналізу представлено на рис. 3.

Data: бак MI* (31v by 142c)

Матінф (2)

	1 Рік	2 N	4 Історія України	5 Філософія	12 Педагогіка та ОПМ Заг	13 Психол огія	18 Математичний аналіз Заг	19 Алгебра і т.ч. 3	20 Геометрія 2	24 МВМ Заг	29 Педагогічна практика
65	2007	65	4	3	3	3	4	3	3	3	5
66	2007	66	3	3	3	3	3	3	3	3	3
67	2008	67	4	4	5	4	4	4	4	5	5
68	2008	68	4	4	4	4	4	4	4	4	5
69	2008	69	3	5	4	4	4	4	4	4	5
70	2008	70	3	3	3	4	3	3	3	3	4
71	2008	71	3	5	4	5	4	3	3	4	5
72	2008	72	3	5	3	5	4	3	4	4	5
73	2008	73	5	5	4	4	5	4	4	5	5
74	2008	74	5	4	3	5	4	4	3	3	4
75	2008	75	5	5	5	5	5	5	5	5	5
76	2008	76	4	4	5	5	5	5	5	5	5
77	2008	77	5	5	5	5	5	5	5	5	5
78	2008	78	4	4	4	5	4	4	4	5	5
79	2008	79	4	4	4	5	5	4	4	4	5
80	2008	80	3	4	3	3	3	4	3	3	3
81	2009	81	5	4	4	4	4	5	4	3	4
82	2009	82	4	4	4	4	4	4	4	4	5

Рис. 3

Канонічний аналіз виконувався засобами статистичного пакету Statistica. У дослідженні успішності на спеціальності "Математика" отримані результати виявилися цілком очікуваними і підтвердили апріорні гіпотези про наявність статистичного зв'язку між базовими математичними дисциплінами, що вивчаються на перших курсах (мат. аналіз, алгебра, геометрія – ліва множина) та дисциплінами, що вивчаються на старших курсах (математична логіка і теорія алгоритмів, теорія ймовірностей і математична статистика, методи обчислень – права множина).

Canonical Analysis Summary (бак MI)
Canonical R: ,78177
Chi?(21)=151,07 p=0,0000

N=139	Left Set	Right Set
No. of variables	7	3
Variance extracted	71,1106%	100,000%
Total redundancy	39,3013%	50,6631%
Variables: 1	Математичний аналіз 1	Мат. логіка і теорія алгоритмів
2	Математичний аналіз 2	Теорія ймов.
3	Математичний аналіз 3	Чис. Методи
4	Математичний аналіз 4	
5	Математичний аналіз Заг	
6	Алгебра і т.ч. 3	
7	Геометрія 2	

Рис. 4

На рис. 4 представлено основні результати канонічного аналізу: як бачимо виявлено досить сильний значущий зв'язок (R=0,78 при p<0,001). При цьому також цікаво з'ясувати, що задаючи значення змінних з лівої множини, можна пояснити 50,7% дисперсії змінних, що відповідають правій множині (тобто за успішністю вивчення базових математичних дисциплін можна прогнозувати успішність опанування дисциплін, що вивчатимуться на старших курсах). Вплив змінних правої множини на ліву не цікавий у контексті даного дослідження, бо не підлягає інтерпретуванню.

Оскільки множини складаються відповідно з п'яти та трьох змінних, матимемо три латентні (рис. 5) та три канонічні (рис. 6) корені, з яких значущим є лише один (p<0,01):

Root	Eigenvalues (бак MI)		
	Root 1	Root 2	Root 3
Value	0,611	0,163	0,017

Рис. 5

Root Removed	Chi-Square Tests with Successive Roots Removed (бак MI)					
	Canonicl R	Canonicl R-sqr.	Chi-sqr.	df	p	Lambda Prime
0	0,782	0,611	151,070	21	0,000	0,320
1	0,404	0,163	25,910	12	0,011	0,822
2	0,132	0,017	2,326	5	0,802	0,983

Рис. 6

Root Variable	Factor Structure, left set (бак MI		
	Root 1	Root 2	Root 3
Математичний аналіз 1	0,648	-0,299	-0,246
Математичний аналіз 2	0,761	0,265	0,089
Математичний аналіз 3	0,732	0,274	-0,008
Математичний аналіз 4	0,810	-0,024	0,051
Математичний аналіз Заг	0,763	0,224	-0,192
Алгебра і т.ч. 3	0,961	0,140	0,083
Геометрія 2	0,845	0,067	-0,376

Рис. 7

Root Variable	Factor Structure, right set (бак MI)		
	Root 1	Root 2	Root 3
Мат. логіка і теорія алг.	0,915	0,344	-0,209
Теорія ймов.	0,853	-0,357	-0,381
Чис. Методи	0,920	-0,057	0,389

Рис. 8

Подальший аналіз дозволяє з'ясувати, що усі вибрані змінні, як з лівої, так і з правої множини, досить сильно корелюють саме з першим канонічним коренем. Кореляція з другим та третім коренями значно слабша (рис. 7, 8).

Візуально результати канонічного аналізу подають у вигляді графіка розсіювання значущих канонічних змінних. На рис. 9 представлено графік розсіювання першої канонічної змінної (перший канонічний корінь) та графік лінійної регресії для нього. Як видно з графіка, зв'язок між канонічними змінними можна вважати лінійним, однак діаграма розсіювання надто розмита, тобто зв'язок хоч і значущий, але значно послаблений викидами.

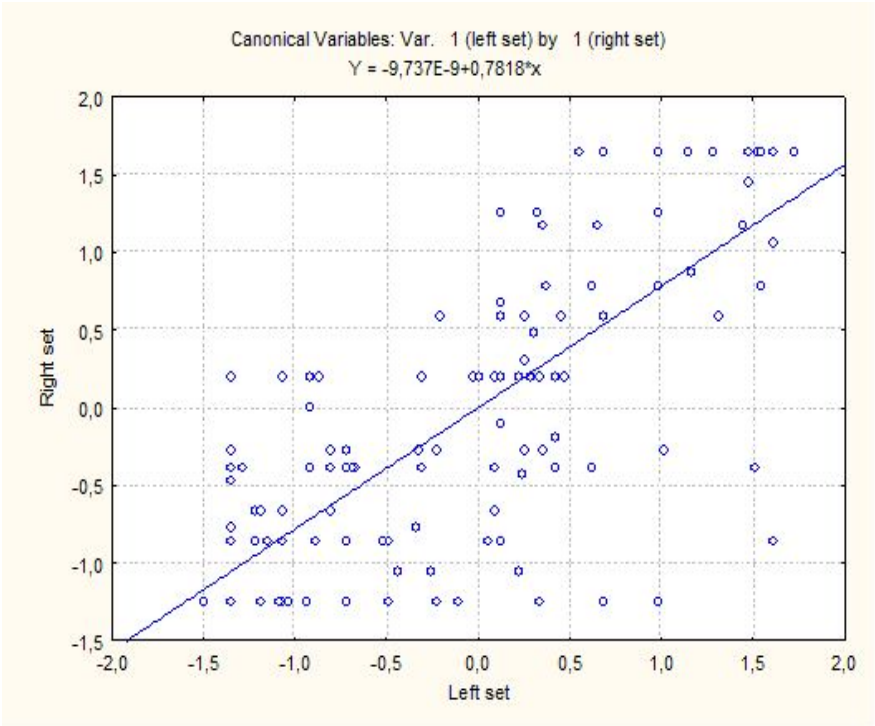

Рис. 9

Інакше виглядає діаграма розсіювання під час порівняння множини гуманітарних та множини математичних дисциплін (рис. 10). Точки навколо прямої регресії розташовані більш компактно.

При цьому і коефіцієнт канонічної кореляції виявився несподівано більшим: R=0,90 (рис. 11).

Основним завданням даного дослідження було ознайомлення з основними результатами канонічного кореляційного аналізу та їхньою інтерпретацією в контексті предметної галузі. З оглядом на педагогічне дослідження слід зазначити, що у наведених прикладах дані не повною мірою відповідають вимогам, які висуваються до даних у канонічному аналізі, тому стосовно зв'язку між різними блоками навчальних дисциплін рано ще робити остаточні висновки. Однак у сучасній педагогічній практиці є чимало питань, для дослідження яких інструментарій канонічного аналізу може стати у нагоді.

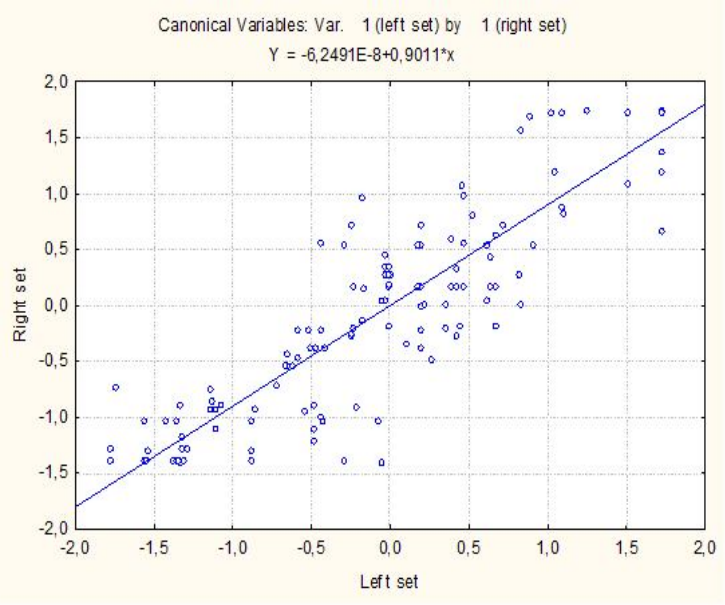

Рис. 10

N=139	Canonical Analysis Summary (бак MI) Canonical R: ,90111 Chi?(36)=255,92 p=0,0000 Left Set	Right Set
No. of variables	6	6
Variance extracted	100,000%	100,000%
Total redundancy	48,7539%	56,6741%
Variables: 1	Історія України	Математичний аналіз Заг
2	Філософія	Алгебра і т.ч. 3
3	Основи економічної теорії	Геометрія 2
4	Політологія	Мат. логіка і теорія алгоритмів
5	Ділова українська мова (УМ за ПС)	Теорія ймов.
6	Іноземна мова	Чис. Методи

Рис. 11

Література

1. Hotelling H. Relations between two sets of variates [Електронний ресурс] / H. Hotelling. – Biometrika (1936) 28 (3–4). – 321–377 pp. – Режим доступу:
http://www.csulb.edu/~jchang9/OnlinePapers/relations_between_two_sets_of_variates.pdf. – Назва з екрану.

2. Clark D. Understanding canonical correlation analysis, 1975, 18–24. [Електронний ресурс] / D. Clark. – Режим доступу:
http://www.qmrg.org.uk/files/2008/12/3-understanding-canonical-correlation-analysis1.pdf. – Назва з екрану.

3. Krus D. J., et al. Rotation in canonical analysis. Educational and Psychological Measurement, 36. – 1976. – P. 725–730 [Електронний ресурс]. – Реж. доступу:
http://www.visualstatistics.net/Statistics/Rotation in CA/Rotation in CA.htm. – Назва з екрану.

4. Stevens J. Applied multivariate statistics for the social sciences L. Erlbaum Associates Inc. / J. Stevens. – Hillsdale, NJ, USA,1986 [Електронний ресурс]. – Режим доступу:
http://books.google.com/books/about/Applied_multivariate_statistics_for_the.html?id=mK0MtyWa7-QC. – Назва з екрану.

5. Liang K. H. K-fold crossvalidation in canonical analysis [Електронний ресурс] / K. H. Liang, D. J. Krus, J. M. Webb. – Multivariate Behavioral Research, 30. – 1995. – P. 539–545. – Режим доступу:
www.visualstatistics.net/Statistics/K-fold CA/K-fold CA.doc. – Назва з екрану.

6. Канонический корреляционный анализ [Електронний ресурс] – Режим доступу:
http://www.statsoft.ru/home/portal/applications/academic/kanon.htm. – Назва з екрану.

7. Bayyurt N. Comparative efficiency measurement of Turkish and Chinese manufacturing firms [Ел. ресурс] / N. Bayyurt, G. Duzu. – Режим доступу:
http://ces.epoka.edu.al/icme/30.pdf. – Назва з екрану.

8. Диденко Н. И. Методы анализа процессов в мировой экономике [Електронний ресурс] / Н. И. Диденко. – СПб., 2007. – Режим доступу:
http://window.edu.ru/window_catalog/files/r61538/Metodi_analiza.pdf. – Назва з екрану.
9. Ефимов В. М. Многомерный анализ биологических данных : учебное пособие [Електронний ресурс] / В. М. Ефимов, В. Ю. Ковалева. – Горно-Алтайск : РИО ГАГУ, 2007. – 75 с. – Режим доступу:
http://bib.tiera.ru/ShiZ/Homelab/spec131/ Efimov V.M., Kovaleva V.Yu. Mnogomernyj analiz biologicheskih dannyh. (2007).pdf. – Назва з екрану.
10. Гальченко В. Я. Анализ взаимосвязи между синдромами функциональных заболеваний кишечника и факторами риска их развития у детей раннего возраста средствами канонического анализа / В. Я. Гальченко, А. М. Полковниченко, Л. М. Полковниченко та ін. // Український медичний альманах. – 2011. – Т. 14, № 1 (додаток). – С. 72–76.
11. Чеча В. О. Дослідження можливостей методу канонічного аналізу : дипломна робота бакалавра / В. О. Чеча. – Кіровоград, 2011. – 49 с. (рукопис).