ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Д.А. Борзых

О МЕТОДЕ КАНОНИЧЕСКИХ КОРРЕЛЯЦИЙ

Препринт WP2/2016/01
Серия WP2
Количественный анализ в экономике

Москва
2016

Редактор серии WP2
«Количественный анализ в экономике»
В.А. Бессонов

Борзых, Д. А.
О методе канонических корреляций [Электронный ресурс]: препринт WP2/2016/01 / Д. А. Борзых ; Нац. исслед. ун-т «Высшая школа экономики». – Электрон. текст. дан. (500 Кб). – М. : Изд. дом Высшей школы экономики, 2016. – (Серия WP2 «Количественный анализ в экономике»). – 18 с.

Предложен подход к изложению метода канонических корреляций, основанный на одном экстремальном свойстве квадратичных форм. Традиционно при изложении метода канонических корреляций используется метод множителей Лагранжа. На симулированных данных показано, что метод канонических корреляций может успешно применяться для выявления скрытых зависимостей между двумя группами переменных. Данная техника может оказаться исключительно полезной на предварительном этапе анализа данных в ситуации, когда на роль зависимой переменной для регрессии одновременно претендует достаточно большое количество показателей.

Ключевые слова: канонические корреляции, канонические переменные, канонические нагрузки

Классификация JEL: C38

Борзых Д.А., Национальный исследовательский университет «Высшая школа экономики»; E-mail: dborzykh@hse.ru

Препринты Национального исследовательского университета
«Высшая школа экономики» размещаются по адресу: http://www.hse.ru/org/hse/wp

1. Введение

Исследователи, работающие в области прикладной статистики и эконометрики, далеко не всегда имеют теоретическую модель, которую им предстоит оценить по имеющимся данным. Напротив, в большинстве случаев они сталкиваются с задачей, называемой “data mining”. Классическим примером такой ситуации является случай, когда исследователи располагают некоторым обширным набором данных и не имеют четко поставленной задачи, а имеют достаточно размытую цель – выявить какие-либо зависимости между имеющимися переменными. В этом случае может оказаться полезной техника, основанная на методе канонических корреляций.

Метод канонических корреляций был разработан Хотеллингом [Hotelling, 1936] для выявления корреляционных связей между двумя группами случайных величин и является естественным обобщением определения множественной корреляции (см. [Мхитарян, Трошин, 2003]). Отличие множественной корреляции от метода канонических корреляций состоит в том, что множественная корреляция является мерой связи лишь *одной* случайной величины с набором других случайных величин.

Подробное изложение метода канонических корреляций можно найти, например, в [Андерсон, 1963; Дубров, Мхитарян, Трошин, 2003; Hamilton, 1994; Hardoon, Szedmak, Shawe-Taylor, 2004]. Более современное изложение содержится в [Ulyanov, Fujikoshi, Shimizu, 2010].

Обоснование метода канонических корреляций в приведенных выше источниках опирается на метод множителей Лагранжа. В данной работе в разделе 2 мы предлагаем альтернативный подход к изложению метода канонических корреляций, основанный на одном экстремальном свойстве квадратичных форм, содержащемся в лемме 2. Данный подход проще в техническом отношении по сравнению с методом множителей Лагранжа и несколько быстрее приводит к цели (для сравнения см., например, [Андерсон, 1963, с. 390–398]).

В разделе 3 работы на симулированных данных показано, что метод канонических корреляций успешно справляется с задачей обнаружения скрытых зависимостей между двумя группами переменных.

2. Теория

Итак, пусть заданы два случайных вектора $X=(X_1,\dots,X_n)'$ и $Y=(Y_1,\dots,Y_m)'$ с конечными вторыми моментами и ковариационными матрицами

$$\Sigma_{XX}:=\mathrm{V}[X]\in\mathbb{R}^{n\times n},\quad \Sigma_{YY}:=\mathrm{V}[Y]\in\mathbb{R}^{m\times m},$$

$$\Sigma_{XY}:=\operatorname{cov}(X,Y)\in\mathbb{R}^{n\times m},\quad \Sigma_{YX}:=\operatorname{cov}(Y,X)\in\mathbb{R}^{m\times n}.$$

Будем предполагать, что матрицы Σ_{XX}, Σ_{YY} и $\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ являются положительно определенными. Обозначим через $p := \min(n,m)$.

Определение. *Первой парой канонических коэффициентов* для случайных векторов X и Y называется пара $(\alpha_{(1)}, \beta_{(1)})$, являющаяся решением экстремальной задачи

$$\begin{cases}\operatorname{corr}(\alpha' X, \beta' Y) \to \max\limits_{(\alpha,\,\beta)\in\mathbb{R}^n\times\mathbb{R}^m}\\ \mathrm{V}[\alpha' X] = \mathrm{V}[\beta' Y] = 1;\end{cases}$$

случайные величины $X_{(1)} := \alpha'_{(1)}X$ и $Y_{(1)} := \beta'_{(1)}Y$ называются *первой парой канонических случайных величин*, а число $\mathcal{C}_1(X,Y) := \operatorname{corr}(\alpha'_{(1)}X, \beta'_{(1)}Y)$ называется *первой канонической корреляцией*.

k-й парой канонических коэффициентов ($2 \le k \le p$) для случайных векторов X и Y называется пара $(\alpha_{(k)}, \beta_{(k)})$, являющаяся решением экстремальной задачи

$$\begin{cases}\operatorname{corr}(\alpha' X, \beta' Y) \to \max\limits_{(\alpha,\,\beta)\in\mathbb{R}^n\times\mathbb{R}^m}\\ \mathrm{V}[\alpha' X] = \mathrm{V}[\beta' Y] = 1,\\ \operatorname{cov}(\alpha' X, \alpha'_{(i)}X) = 0 \quad \text{для всех } i \in \{1,\ldots,k-1\},\\ \operatorname{cov}(\beta' Y, \beta'_{(j)}Y) = 0 \quad \text{для всех } j \in \{1,\ldots,k-1\};\end{cases}$$

случайные величины $X_{(k)} := \alpha'_{(k)}X$ и $Y_{(k)} := \beta'_{(k)}Y$ называются *k-й парой канонических случайных величин*, а число $\mathcal{C}_k(X,Y) := \operatorname{corr}(\alpha'_{(k)}X, \beta'_{(k)}Y)$ – *k-й канонической корреляцией*.

Покажем, что для любой пары случайных векторов $X = (X_1,\ldots,X_n)'$ и $Y = (Y_1,\ldots,Y_m)'$ с заданными выше свойствами существуют все p пар канонических коэффициентов $(\alpha_{(1)}, \beta_{(1)}),\ \ldots\ ,(\alpha_{(p)}, \beta_{(p)})$. Для дальнейшего нам потребуются две леммы.

Лемма 1 (неравенство Коши – Буняковского). *Для любых векторов $x, y \in \mathbb{R}^{n\times 1}$ имеет место неравенство*

$$|x'y| \le \sqrt{x'x}\sqrt{y'y}\,,$$

причем равенство в этом неравенстве достигается в том и только в том случае, когда векторы x и y являются линейно зависимыми.

Данное утверждение является широко известным фактом из курса линейной алгебры.

Лемма 2 (об экстремальном свойстве квадратичных форм). *Пусть $A \in \mathbb{R}^{n\times n}$ – симметрическая матрица и $q_1,\ldots,q_n \in \mathbb{R}^{n\times 1}$ – ее ортонормированный базис из собственных векторов, отвечающих собственным значениям $\lambda_1 \ge \ldots \ge \lambda_n$. Тогда для квадратичной формы* $f(x) = x'Ax$ справедливы следующие утверждения:

(i) $f(x) \le f(q_1) = \lambda_1$ *для любых векторов $x \in \mathbb{R}^{n\times 1}$ с $x'x = 1$*,

(ii) $f(x) \le f(q_k) = \lambda_k$ *для любого* $2 \le k \le n$ *и всех векторов* $x \in \mathbb{R}^{n\times 1}$ с $x'x = 1$, $q_1'x = \ldots = q_{k-1}'x = 0$.

Доказательство. Введем обозначения

$$\Lambda := \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix} \qquad \text{и} \qquad Q := [q_1 \cdots q_n] \in \mathbb{R}^{n\times n}.$$

Заметим, что для любого вектора $x \in \mathbb{R}^{n\times 1}$, для которого $x'x = 1$, имеет место равенство

$$\sum_{i=1}^{n} (q_i'x)^2 = 1. \tag{1}$$

В самом деле, $1 = x'x = x'QQ'x = (Q'x)'Q'x = \sum_{i=1}^{n} (q_i'x)^2$. Здесь мы воспользовались ортогональностью матрицы Q, из которой вытекает соотношение $QQ' = I$. Тогда для матрицы A имеет место разложение $A = Q\Lambda Q'$, с помощью которого получаем представление

$$f(x) = x'Ax = x'Q\Lambda Q'x = (Q'x)'\Lambda Q'x = \sum_{i=1}^{n} \lambda_i (q_i'x)^2 . \tag{2}$$

Докажем утверждение (i). Принимая во внимание равенство (1), замечаем, что квадратичная форма $f(x)$ представляет собой выпуклую комбинацию чисел $\lambda_1, \ldots, \lambda_n$ с коэффициентами $(q_1'x)^2, \ldots, (q_n'x)^2$. Отсюда, а также из того, что λ_1 является наибольшим среди чисел $\lambda_1, \ldots, \lambda_n$, следует, что $f(x)$ принимает наибольшее значение, равное λ_1. Для этого надо положить $(q_1'x)^2 = 1$, $(q_2'x)^2 = \ldots = (q_n'x)^2 = 0$. Данные условия выполняются, например, при $x = q_1$. Итак, мы показали, что для любого вектора $x \in \mathbb{R}^{n\times 1}$, для которого $x'x = 1$, имеет место соотношение

$$f(x) \le f(q_1) = \lambda_1 .$$

Доказательство утверждения (ii) дословно повторяет пункт (i) после учета условий $q_1'x = \ldots = q_{k-1}'x = 0$ в равенствах (1) и (2). В этом случае квадратичная форма $f(x)$ принимает вид

$$f(x) = \sum_{i=k}^{n} \lambda_i (q_i'x)^2 , \tag{3}$$

т.е. теперь $f(x)$ является выпуклой комбинацией чисел $\lambda_k, \ldots, \lambda_n$ с коэффициентами $(q_k'x)^2, \ldots, (q_n'x)^2$. Учитывая, что λ_k является наибольшим среди чисел $\lambda_k, \ldots, \lambda_n$, заключаем, что выпуклая комбинация (3) принимает наибольшее значение, равное λ_k. Это происходит, например, при $x = q_k$, поскольку в этом случае $q_k'x = 1$, $q_{k+1}'x = \ldots q_n'x = 0$. Таким образом, утверждение (ii) доказано. □

Теперь перейдем непосредственно к изложению способа нахождения канонических коэффициентов. Для определенности будем считать, что $m \le n$, т.е. $p = m$. Рассмотрим функцию

$$\rho = \mathrm{corr}(\alpha' X, \beta' Y) = \frac{\mathrm{cov}(\alpha' X, \beta' Y)}{\sqrt{\mathrm{V}[\alpha' X]}\sqrt{\mathrm{V}[\beta' Y]}} = \frac{\alpha' \mathrm{cov}(X, Y)\beta}{\sqrt{\alpha' \mathrm{V}[X]\alpha}\sqrt{\beta' \mathrm{V}[Y]\beta}} = \frac{\alpha' \Sigma_{XY}\beta}{\sqrt{\alpha' \Sigma_{XX}\alpha}\sqrt{\beta' \Sigma_{YY}\beta}}.$$

Выполним замену переменных:

$$\gamma := \Sigma_{XX}^{1/2}\alpha, \qquad \delta := \Sigma_{YY}^{1/2}\beta.$$

В результате получим

$$\rho = \frac{\alpha' \Sigma_{XY}\beta}{\sqrt{\alpha' \Sigma_{XX}\alpha}\sqrt{\beta' \Sigma_{YY}\beta}} = \frac{\gamma' \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}{\sqrt{\gamma'\gamma}\sqrt{\delta'\delta}}. \tag{4}$$

Далее, согласно неравенству Коши – Буняковского скалярное произведение $\gamma' \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta$ векторов γ и $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta$ в числителе дроби (4) не превосходит произведения $\sqrt{\gamma'\gamma}\sqrt{(\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta)'\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}$, т.е. имеем оценку

$$\gamma' \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta \le \sqrt{\gamma'\gamma}\sqrt{\delta' \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}. \tag{5}$$

Обратим внимание, что в силу неравенства Коши – Буняковского данное неравенство обратится в равенство, если векторы γ и $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta$ взять пропорциональными друг другу, т.е. $\gamma = c \cdot \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta$, где c – произвольный коэффициент, который в дальнейшем будет выбран так, чтобы выполнялось условие $\mathrm{V}[\alpha' X] = 1$.

Тогда при таком выборе вектора γ неравенство (5) обращается в равенство, и формула (4) может быть переписана в виде

$$\rho = \frac{\gamma' \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}{\sqrt{\gamma'\gamma}\sqrt{\delta'\delta}} = \frac{\sqrt{\gamma'\gamma}\sqrt{\delta' \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}}{\sqrt{\gamma'\gamma}\sqrt{\delta'\delta}} = \sqrt{\frac{\delta' \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}{\delta'\delta}}.$$

Теперь рассмотрим квадратичную форму $f(\delta) = \delta' \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta$. Пусть $q_1, \ldots, q_m \in \mathbb{R}^{m\times 1}$ – ортонормированный базис из собственных векторов матрицы $A = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}$, отвечающих собственным значениям $\lambda_1 \ge \ldots \ge \lambda_m$. Тогда согласно пункту (i) леммы 2

$$\forall \delta \in \mathbb{R}^{m\times 1},\ \delta'\delta = 1 \qquad f(\delta) \le f(q_1).$$

Следовательно, для $\forall \delta \ne 0$

$$\rho = \sqrt{\frac{\delta' \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}{\delta'\delta}} = \sqrt{\frac{f(\delta)}{\delta'\delta}} \le \sqrt{\frac{f(q_1)}{q_1' q_1}},$$

откуда видно, что функция ρ принимает наибольшее значение, равное $\rho_1^* = \sqrt{\frac{f(q_1)}{q_1' q_1}}$, при $\delta = q_1$. Значит, первая каноническая корреляция равна

$$\mathcal{C}_1(X,Y)=\rho_1^*=\sqrt{\frac{q_1'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_1}{q_1'q_1}}=\sqrt{\frac{q_1'\lambda_1 q_1}{q_1'q_1}}=\sqrt{\lambda_1}\ .$$

Найдем теперь канонические коэффициенты и канонические переменные:

$$\alpha_{(1)}=\Sigma_{XX}^{-1/2}\gamma=\Sigma_{XX}^{-1/2}c\cdot\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta=c\cdot\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_1\,,$$

$$\beta_{(1)}=\Sigma_{YY}^{-1/2}\delta=\Sigma_{YY}^{-1/2}q_1\,,$$

$$X_{(1)}=\alpha_{(1)}'X=c\cdot q_1'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}X \quad \text{и} \quad Y_{(1)}=\beta_{(1)}'Y=q_1'\Sigma_{YY}^{-1/2}Y\,,$$

где коэффициент c определяется из условия $\mathrm{V}[X_{(1)}]=1$:

$$\mathrm{V}[X_{(1)}]=\mathrm{V}[c\cdot q_1'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}X]=c^2\cdot q_1'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_1=$$

$$=c^2\cdot q_1'\underbrace{\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_1}_{=\lambda_1 q_1}=c^2\cdot\lambda_1\underbrace{q_1'q_1}_{=1}=1\,,$$

т.е. $c=\lambda_1^{-1/2}$. Дисперсия случайной величины $Y_{(1)}$ равна единице:

$$\mathrm{V}[Y_{(1)}]=\mathrm{V}[q_1'\Sigma_{YY}^{-1/2}Y]=q_1'\Sigma_{YY}^{-1/2}\,\mathrm{V}[Y]\Sigma_{YY}^{-1/2}q_1'=q_1'\underbrace{\Sigma_{YY}^{-1/2}\Sigma_{YY}\Sigma_{YY}^{-1/2}}_{=I}q_1'=1\,.$$

Таким образом, *первой парой канонических коэффициентов* является набор

$$\alpha_{(1)}=\lambda_1^{-1/2}\cdot\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_1 \qquad \text{и} \qquad \beta_{(1)}=\Sigma_{YY}^{-1/2}q_1\,; \tag{6}$$

первая пара канонических случайных величин есть

$$X_{(1)}=\lambda_1^{-1/2}\cdot q_1'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}X \quad \text{и} \quad Y_{(1)}=q_1'\Sigma_{YY}^{-1/2}Y\,, \tag{7}$$

а *первая каноническая корреляция* равна

$$\mathcal{C}_1(X,Y)=\sqrt{\lambda_1}\ . \tag{8}$$

Аналогично, применяя пункт (ii) леммы 2, убеждаемся, что k*-й парой канонических коэффициентов* является

$$\alpha_{(k)}=\lambda_k^{-1/2}\cdot\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_k \qquad \text{и} \qquad \beta_{(k)}=\Sigma_{YY}^{-1/2}q_k\,; \tag{9}$$

k*-я пара канонических случайных величин* есть:

$$X_{(k)}=\lambda_k^{-1/2}\cdot q_k'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}X \qquad \text{и} \qquad Y_{(k)}=q_k'\Sigma_{YY}^{-1/2}Y\,, \tag{10}$$

а k*-я каноническая корреляция* равна

$$\mathcal{C}_k(X,Y)=\mathrm{corr}(\alpha_{(k)}'X,\beta_{(k)}'Y)=\rho=\sqrt{\lambda_k}\ . \tag{11}$$

В самом деле, в силу пункта (ii) леммы 2 для любого $k\in\{2,\ldots,m\}$ имеем

$$\forall\delta\in\mathbb{R}^{m\times 1},\quad \delta'\delta=1,\quad q_1'\delta=\ldots=q_{k-1}'\delta=0 \qquad f(\delta)\le f(q_k)=\lambda_k\,.$$

Следовательно, $\forall\delta\ne 0$: $q_1'\delta=\ldots=q_{k-1}'\delta=0$

$$\rho=\sqrt{\frac{\delta'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}\delta}{\delta'\delta}}=\sqrt{\frac{f(\delta)}{\delta'\delta}}\le\sqrt{\frac{f(q_k)}{q_k'q_k}}$$

т.е. при $\delta = q_k$ на множестве $\delta \neq 0,\ q_1'\delta = \ldots = q_{k-1}'\delta = 0$ функция ρ принимает наибольшее значение, равное $\rho_k^* = \sqrt{\frac{f(q_k)}{q_k'q_k}}$. Значит, k-я каноническая корреляция равна

$$\mathcal{C}_k(X,Y) = \rho_k^* = \sqrt{\frac{f(q_k)}{q_k'q_k}} = \sqrt{\frac{q_k'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_k}{q_k'q_k}} = \sqrt{\frac{q_k'\lambda_k q_k}{q_k'q_k}} = \sqrt{\lambda_k}\ .$$

Здесь следует отметить, что в силу равенства

$$\operatorname{cov}(\beta'Y, \beta_{(j)}'Y) = \beta'\Sigma_{YY}\beta_{(j)} = \delta'\Sigma_{YY}^{-1/2}\Sigma_{YY}\Sigma_{YY}^{-1/2}q_j = \delta' q_j = 0\,, \qquad j = 1,\ldots,k-1\,,$$

условие $q_1'\delta = \ldots = q_{k-1}'\delta = 0$ равносильно условию

$$\operatorname{cov}(\beta'Y, \beta_{(1)}'Y) = \ldots = \operatorname{cov}(\beta'Y, \beta_{(k-1)}'Y) = 0\,.$$

Таким образом, для $\delta = q_k$ вектор $\beta_{(k)} = \Sigma_{YY}^{-1/2}\delta = \Sigma_{YY}^{-1/2}q_k$ удовлетворяет условию

$$\operatorname{cov}(\beta_{(k)}'Y, \beta_{(1)}'Y) = \ldots = \operatorname{cov}(\beta_{(k)}'Y, \beta_{(k-1)}'Y) = 0\,,$$

которое налагается на вектор $\beta_{(k)}$ в определении k-й канонической корреляции.

Оставшиеся условия k-й пары канонических коэффициентов проверяются просто:

$$\mathrm{V}[X_{(k)}] = \mathrm{V}[\alpha_{(k)}'X] = \lambda_k^{-1}q_k'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_k =$$

$$= \lambda_k^{-1}q_k'\underbrace{\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_k}_{=\lambda_k q_k} = \lambda_k^{-1}q_k'\lambda_k q_k = 1\,,$$

$$\mathrm{V}[Y_{(k)}] = \mathrm{V}[\beta_{(k)}'Y] = q_k'\underbrace{\Sigma_{YY}^{-1/2}\Sigma_{YY}\Sigma_{YY}^{-1/2}}_{=I}q_k = 1\,,$$

и, наконец, при $j \in \{1,\ldots,k-1\}$

$$\operatorname{cov}(X_{(k)}, X_{(j)}) = \operatorname{cov}(\alpha_{(k)}'X, \alpha_{(j)}'X) = \alpha_{(k)}'\Sigma_{XX}\alpha_{(j)} =$$

$$= \lambda_k^{-1/2}q_k'\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX}\lambda_j^{-1/2}\cdot\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_j =$$

$$= \lambda_k^{-1/2}\lambda_j^{-1/2}q_k'\underbrace{\Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2}q_j}_{=\lambda_j q_j} = \lambda_k^{-1/2}\lambda_j^{1/2}\underbrace{q_k'q_j}_{=0} = 0\,.$$

Пример 1. Пусть $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5, \varepsilon_6$ – независимые стандартные нормальные случайные величины. Рассмотрим два случайных вектора

$$X = (\varepsilon_1,\ \varepsilon_1+\varepsilon_2,\ \varepsilon_1+\varepsilon_2+\varepsilon_3)' \qquad \text{и} \qquad Y = (\varepsilon_1,\ \varepsilon_2+\varepsilon_4,\ \varepsilon_3+\varepsilon_5+\varepsilon_6)'\,.$$

Найдем все три пары канонических коэффициентов $(\alpha_{(1)},\beta_{(1)}),\ (\alpha_{(2)},\beta_{(2)}),\ (\alpha_{(3)},\beta_{(3)})$, три набора канонических случайных величин $(X_{(1)},Y_{(1)})$, $(X_{(2)},Y_{(2)})$, $(X_{(3)},Y_{(3)})$, а также три канонические корреляции $\mathcal{C}_1(X,Y)$, $\mathcal{C}_2(X,Y)$ и $\mathcal{C}_3(X,Y)$. Ковариационные матрицы случайных векторов X и Y равны

$$\Sigma_{XX} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}, \quad \Sigma_{XY} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad \Sigma_{YX} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Sigma_{YY} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Далее, находим

$$\Sigma_{XX}^{-1} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \qquad \Sigma_{YY}^{-1/2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{3}} \end{bmatrix},$$

$$A = \Sigma_{YY}^{-1/2}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1/2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

Векторы $q_1 = (1,0,0)'$, $q_2 = (0,1,0)'$ и $q_3 = (0,0,1)'$ образуют ортонормированный базис из собственных векторов матрицы A, отвечающих собственным значениям $\lambda_1 = 1$, $\lambda_2 = \frac{1}{2}$ и $\lambda_3 = \frac{1}{3}$. Тогда согласно формулам (6)–(11) получаем:

$$\alpha_{(1)} = (1,0,0), \quad \beta_{(1)} = (1,0,0), \quad X_{(1)} = \varepsilon_1, \quad Y_{(1)} = \varepsilon_1, \quad \mathcal{C}_1(X,Y) = 1;$$

$$\alpha_{(2)} = (-1,1,0), \quad \beta_{(2)} = (0,\tfrac{1}{\sqrt{2}},0), \quad X_{(2)} = \varepsilon_2, \quad Y_{(2)} = \tfrac{\varepsilon_2+\varepsilon_4}{\sqrt{2}}, \quad \mathcal{C}_2(X,Y) = \tfrac{1}{\sqrt{2}};$$

$$\alpha_{(3)} = (0,-1,1), \quad \beta_{(3)} = (0,0,\tfrac{1}{\sqrt{3}}), \quad X_{(3)} = \varepsilon_3, \quad Y_{(3)} = \tfrac{\varepsilon_3+\varepsilon_5+\varepsilon_6}{\sqrt{3}}, \quad \mathcal{C}_3(X,Y) = \tfrac{1}{\sqrt{3}}.$$

Для интерпретации результатов применения канонического анализа в следующем параграфе нам потребуется определение так называемых *матриц нагрузок*.

Определение. Пусть $X = (X_1,\dots,X_n)'$ и $Y = (Y_1,\dots,Y_m)'$ – два случайных вектора и $p = \min(n,m)$. *Матрицей нагрузок* вектора X называется корреляционная матрица

$$\begin{bmatrix} \operatorname{corr}(X_1, X_{(1)}) & \cdots & \operatorname{corr}(X_1, X_{(p)}) \\ \vdots & \ddots & \vdots \\ \operatorname{corr}(X_n, X_{(1)}) & \cdots & \operatorname{corr}(X_n, X_{(p)}) \end{bmatrix}$$

вектора X и вектора канонических переменных $U = (X_{(1)},\dots,X_{(p)})'$. Аналогично определяется матрица нагрузок вектора Y.

Перейдем теперь к вопросу оценивания канонических корреляций и канонических переменных по данным. Пусть $(X(1),Y(1))$, … , $(X(T),Y(T))$ – случайная выборка, содержащая T независимых наблюдений, где $X(t)$ и $Y(t)$ – $(n\times 1)$– и $(m\times 1)$ – случайные векторы соответственно. Тогда, используя оценки ковариационных матриц

$$\hat{\Sigma}_{XX} = \frac{1}{T}\sum_{t=1}^{T}(X(t)-\bar{X})\cdot(X(t)-\bar{X})', \qquad \hat{\Sigma}_{YY} = \frac{1}{T}\sum_{t=1}^{T}(Y(t)-\bar{Y})\cdot(Y(t)-\bar{Y})',$$

$$\hat{\Sigma}_{XY} = \frac{1}{T}\sum_{t=1}^{T}(X(t)-\bar{X})\cdot(Y(t)-\bar{Y})', \qquad \hat{\Sigma}_{YX} = \frac{1}{T}\sum_{t=1}^{T}(Y(t)-\bar{Y})\cdot(X(t)-\bar{X})',$$

где $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X(t)$, $\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y(t)$, а также формулы (6), (8), (9) и (11), получаем оценки $\hat{\mathcal{C}}_1(X,Y),\dots,\hat{\mathcal{C}}_p(X,Y)$ канонических корреляций и канонических коэффициентов.

Для тестирования гипотезы $H_0: \mathcal{C}_k(X,Y)=0$ [1] о незначимости k-го коэффициента канонической корреляции $(k=1,\ldots,p)$ против альтернативной гипотезы $H_1: \mathcal{C}_k(X,Y)\neq 0$ о его статистической значимости может быть использована тестовая статистика

$$S=-\left(T-k-\tfrac{n+m+1}{2}+\sum\nolimits_{i=1}^{k-1}\hat{\mathcal{C}}_i^2(X,Y)\right)\cdot\sum\nolimits_{i=k}^{p}\ln(1-\hat{\mathcal{C}}_i^2(X,Y)),$$

имеющая асимптотическое χ^2-распределение с $(n-k+1)\cdot(m-k+1)$ степенями свободы (см. [Дубров, Мхитарян, Трошин, 2003, с. 278]).

3. Примеры применения

Теперь мы переходим к изложению одного из способов интерпретации результатов применения данного инструментария на основе двух примеров. Чтобы иметь контролируемый эксперимент, будем работать с симулированными данными. Начнем с более простого примера.

Пример 2. Сгенерируем выборку

$$X(t)=(X_1(t),X_2(t),X_3(t))', \qquad Y(t)=(Y_1(t),Y_2(t),Y_3(t))', \qquad t=1,\ldots,T,$$

объема $T=500$ наблюдений, в которой случайные векторы $(X_1(t),X_2(t),X_3(t),Y_1(t),Y_2(t),Y_3(t))$, $t=1,\ldots,T$ независимы в совокупности и при каждом фиксированном t имеют многомерное нормальное распределение с нулевым математическим ожиданием и ковариационной матрицей

$$\begin{bmatrix} 1 & 0 & 0 & 0.9 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0.7 & 0 \\ 0 & 0 & 1 & 0 & 0 & -0.6 \\ 0.9 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 1 & 0 \\ 0 & 0 & -0.6 & 0 & 0 & 1 \end{bmatrix}.$$

В частности, это означает, что

$$\operatorname{corr}(X_1(t),Y_1(t))=0.9, \qquad \operatorname{corr}(X_2(t),Y_2(t))=0.7, \qquad \operatorname{corr}(X_3(t),Y_3(t))=-0.6. \tag{12}$$

Решим теперь обратную задачу: на основе полученной выборки попытаемся восстановить корреляционные зависимости (12). Для этого по данной выборке рассчитаем канонические корреляции и матрицы нагрузок. Канонические корреляции получились следующими: $\hat{\mathcal{C}}_1(X,Y)=0.91$, $\hat{\mathcal{C}}_2(X,Y)=0.73$ и $\hat{\mathcal{C}}_3(X,Y)=0.61$. Все они оказались значимыми на любом разумном уровне значимости. Матрицы нагрузок приведены ниже:

---

[1] Заметим, что в силу соотношения $\mathcal{C}_1(X,Y)\geq\ldots\geq\mathcal{C}_p(X,Y)\geq 0$ из равенства $\mathcal{C}_k(X,Y)=0$ следует, что $\mathcal{C}_{k+1}(X,Y)=\ldots=\mathcal{C}_p(X,Y)=0$.

Таблица 1. Матрица нагрузок вектора X

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$
X_1	**0.99**	0.06	0.01
X_2	–0.13	**0.98**	0.13
X_3	0.01	0.24	**–0.97**

Таблица 2. Матрица нагрузок вектора Y

	$Y_{(1)}$	$Y_{(2)}$	$Y_{(3)}$
Y_1	**0.99**	0.04	0.02
Y_2	–0.06	**0.97**	0.20
Y_3	0.01	–0.24	**0.97**

Для выявления взаимосвязей между группами переменных X_1, X_2, X_3 и Y_1, Y_2, Y_3 воспользуемся эвристическим приемом, широко применяемым в социально-экономической литературе (см., например, [Sherry, Henson, 2005; Thompson, 1991]). В контексте данного примера опишем предлагаемый подход.

Как видно из первых столбцов табл. 1 и 2, переменная X_1 наиболее коррелирована с канонической переменной $X_{(1)}$ ($\mathrm{corr}(X_1, X_{(1)}) = 0.99$), в свою очередь, переменная Y_1 наиболее коррелирована с канонической переменной $Y_{(1)}$ ($\mathrm{corr}(Y_1, Y_{(1)}) = 0.99$). Тогда, учитывая значимую корреляцию между связующими каноническими переменными $X_{(1)}$ и $Y_{(1)}$ ($\hat{\mathcal{C}}_1(X,Y) = \mathrm{corr}(X_{(1)}, Y_{(1)}) = 0.91$), выдвигаем предположение о наличии значимой положительной корреляции между переменными X_1 и Y_1. Обратим внимание, что данное предположение соответствует действительности – выборка была сгенерирована с соблюдением условия $\mathrm{corr}(X_1(t), Y_1(t)) = 0.9$. Логика описанного выше подхода схематически проиллюстрирована на рис. 1.

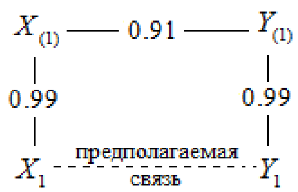

Рис. 1. Графическое изображение взаимосвязей между переменными при построении первой канонической корреляции (над ребрами графа указаны соответствующие корреляции)

Дальнейший анализ вполне аналогичен предыдущему с тем отличием, что изучению подвергаются столбцы матриц нагрузок с последующими номерами. Итак, анализируя вторые столбцы в табл. 1 и 2, видим, что переменная X_2 наиболее коррелирована с канонической переменной $X_{(2)}$ ($\mathrm{corr}(X_2, X_{(2)}) = 0.98$), переменная Y_2 наиболее коррелирована с канонической переменной $Y_{(2)}$ ($\mathrm{corr}(Y_2, Y_{(2)}) = 0.97$). Тогда, принимая во внимание значимую корреляцию между каноническими переменными $X_{(2)}$ и $Y_{(2)}$ ($\hat{\mathcal{C}}_2(X,Y) = \mathrm{corr}(X_{(2)}, Y_{(2)}) = 0.73$), выдвигаем предположение о значимой положительной корреляции между переменными X_2 и Y_2. Как и выше, данное предположение оказывается верным, поскольку $\mathrm{corr}(X_2(t), Y_2(t)) = 0.7$.

Наконец, из третьих столбцов рассматриваемых матриц нагрузок видим, что переменная X_3 наиболее (отрицательно) коррелирована с канонической переменной $X_{(3)}$ ($\mathrm{corr}(X_3, X_{(3)}) = -0.97$), переменная Y_3 наиболее коррелирована с канонической переменной $Y_{(3)}$ ($\mathrm{corr}(Y_3, Y_{(3)}) = 0.97$). Учитывая при этом значимую корреляцию между каноническими переменными $X_{(3)}$ и $Y_{(3)}$ ($\hat{\mathcal{C}}_3(X,Y) = \mathrm{corr}(X_{(3)}, Y_{(3)}) = 0.61$), предполагаем наличие отрицательной корреляционной связи между переменными X_3 и Y_3. Это предположение также оказывается справедливым, так как $\mathrm{corr}(X_3(t), Y_3(t)) = -0.6$.

Обратим внимание, что анализируя каждый следующий столбец матриц нагрузок, мы выявляли все более слабую корреляционную связь – информация о самой тесной связи содержалась в первом столбце, менее тесная связь была обнаружена при изучении второго столбца, а информация о самой слабой – в третьем.

Рассмотрим следующий пример, приближенный к реальному статистическому исследованию.

Пример 3. Сгенерируем выборку

$$X(t) = (X_1(t), \dots, X_{15}(t))', \qquad Y(t) = (Y_1(t), \dots, Y_{10}(t))', \qquad t = 1, \dots, T,$$

объема $T = 500$ наблюдений. Предполагается, что случайные величины $X_1(t), \dots, X_{15}(t), Y_4(t), \dots, Y_{10}(t), \varepsilon_1(t), \dots, \varepsilon_3(t)$ независимы и имеют нормальное стандартное распределение, а величины $Y_1(t)$, $Y_2(t)$ и $Y_3(t)$ рассчитываются по формулам:

$$Y_1(t) = 1 \cdot X_1(t) + 2 \cdot X_2(t) + 3 \cdot X_3(t) + \varepsilon_1(t), \tag{13}$$

$$Y_2(t) = 4 \cdot X_4(t) + 5 \cdot X_5(t) + 6 \cdot X_6(t) + \varepsilon_2(t), \tag{14}$$

$$Y_3(t) = 7 \cdot X_7(t) + 8 \cdot X_8(t) + 9 \cdot X_9(t) + \varepsilon_3(t). \tag{15}$$

Кроме этого предполагается, что случайные векторы $(X_1(t), \dots, X_{15}(t), Y_1(t), \dots, Y_{10}(t))$, $t = 1, \dots, T$ независимы в совокупности.

На основе сгенерированных данных попытаемся выявить заложенные при генерации зависимости (13)–(15). Для этого рассчитаем канонические корреляции и матрицы нагрузок. Канонические корреляции получились следующими: $\hat{\mathcal{C}}_1(X,Y) = 0.99$, $\hat{\mathcal{C}}_2(X,Y) = 0.98$,

$\hat{C}_3(X,Y)=0.96$, $\hat{C}_4(X,Y)=0.22$, $\hat{C}_5(X,Y)=0.20$, $\hat{C}_6(X,Y)=0.19$, $\hat{C}_7(X,Y)=0.12$, $\hat{C}_8(X,Y)=0.11$, $\hat{C}_9(X,Y)=0.06$, $\hat{C}_{10}(X,Y)=0.03$. При этом только первые три канонические корреляции оказались значимыми (на любом разумном уровне значимости). Ниже приведены матрицы нагрузок.

Таблица 3. Матрица нагрузок вектора X

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$
X_1	0.03	0.01	**0.27**
X_2	0.04	0.04	**0.54**
X_3	-0.03	-0.04	**0.80**
X_4	-0.06	**0.46**	0.00
X_5	-0.06	**0.60**	0.01
X_6	0.03	**0.64**	-0.02
X_7	**0.54**	0.04	0.07
X_8	**0.56**	-0.03	-0.02
X_9	**0.65**	0.00	-0.07
X_{10}	-0.01	-0.07	-0.01
X_{11}	0.04	-0.02	0.07
X_{12}	0.01	0.06	-0.06
X_{13}	-0.04	0.01	-0.02
X_{14}	-0.01	0.04	0.01
X_{15}	-0.03	-0.02	0.04

Таблица 4. Матрица нагрузок вектора Y

	$Y_{(1)}$	$Y_{(2)}$	$Y_{(3)}$
Y_1	0.02	-0.00	**0.99**
Y_2	-0.03	**0.99**	-0.00
Y_3	**0.99**	0.00	-0.02
Y_4	-0.00	-0.05	0.02
Y_5	-0.02	-0.02	0.00
Y_6	-0.01	-0.10	-0.02
Y_7	0.08	0.01	-0.01
Y_8	-0.00	0.02	0.03
Y_9	0.07	0.00	-0.03
Y_{10}	0.01	-0.00	0.06

Анализируя первые столбцы табл. 3 и 4, видим, что переменные X_7, X_8 и X_9 наиболее коррелированы с канонической переменной $X_{(1)}$ ($\mathrm{corr}(X_7, X_{(1)})=0.54$, $\mathrm{corr}(X_8, X_{(1)})=0.56$ и $\mathrm{corr}(X_9, X_{(1)})=0.65$), в свою очередь, переменная Y_3 наиболее коррелирована с канонической переменной $Y_{(1)}$ ($\mathrm{corr}(Y_3, Y_{(1)})=0.99$). Тогда значимая корреляция между связующими каноническими переменными $X_{(1)}$ и $Y_{(1)}$ ($\hat{C}_1(X,Y)=\mathrm{corr}(X_{(1)}, Y_{(1)})=0.99$) позволяет выдвинуть предположение о статистической зависимости переменной Y_3 от переменных X_7, X_8 и X_9 (а это и есть зависимость (15), заложенная на этапе генерации данных). Данная зависимость схематически изображена на рис. 2.

$$Y_3 \overset{0.99}{\text{———}} Y_{(1)} \overset{0.99}{\text{———}} X_{(1)} \begin{array}{l} \overset{0.54}{\diagup}\ X_7 \\ \text{-}0.56\text{-}\ X_8 \\ \underset{0.65}{\diagdown}\ X_9 \end{array}$$

Рис. 2. Графическое изображение зависимости Y_3 от X_7, X_8 и X_9
(над ребрами графа указаны значения соответствующих корреляций)

Вторые столбцы табл. 3 и 4 показывают, что переменные X_4, X_5 и X_6 наиболее коррелированы с канонической переменной $X_{(2)}$ ($\mathrm{corr}(X_4, X_{(2)}) = 0.46$, $\mathrm{corr}(X_5, X_{(2)}) = 0.60$ и $\mathrm{corr}(X_6, X_{(2)}) = 0.64$), а переменная Y_2 наиболее коррелирована с канонической переменной $Y_{(2)}$ ($\mathrm{corr}(Y_2, Y_{(2)}) = 0.99$). Следовательно, учитывая значимую корреляцию между связующими каноническими переменными $X_{(2)}$ и $Y_{(2)}$ ($\hat{\mathcal{C}}_2(X,Y) = \mathrm{corr}(X_{(2)}, Y_{(2)}) = 0.98$), выдвигаем предположение о статистической зависимости переменной Y_2 от переменных X_4, X_5 и X_6 (а это и есть зависимость (14), заложенная на этапе генерации данных). Данная зависимость схематически изображена на рис. 3.

$$Y_2 \overset{0.99}{\text{———}} Y_{(2)} \overset{0.98}{\text{———}} X_{(2)} \begin{array}{l} \overset{0.46}{\diagup}\ X_4 \\ \text{-}0.60\text{-}\ X_5 \\ \underset{0.64}{\diagdown}\ X_6 \end{array}$$

Рис. 3. Графическое изображение зависимости Y_2 от X_4, X_5 и X_6
(над ребрами графа указаны значения соответствующих корреляций)

Из третьих столбцов табл. 3 и 4 видно, что переменные X_1, X_2 и X_3 наиболее коррелированы с канонической переменной $X_{(3)}$ ($\mathrm{corr}(X_1, X_{(3)}) = 0.27$, $\mathrm{corr}(X_2, X_{(3)}) = 0.54$ и $\mathrm{corr}(X_3, X_{(3)}) = 0.80$), в свою очередь, переменная Y_1 наиболее коррелирована с канонической переменной $Y_{(3)}$ ($\mathrm{corr}(Y_1, Y_{(3)}) = 0.99$). Тогда, учитывая значимую корреляцию между связующими каноническими переменными $X_{(3)}$ и $Y_{(3)}$ ($\hat{\mathcal{C}}_3(X,Y) = \mathrm{corr}(X_{(3)}, Y_{(3)}) = 0.96$), выдвигаем предположение о статистической зависимости переменной Y_1 от переменных X_1, X_2 и X_3 (а это и есть зависимость (13), заложенная на этапе генерации данных). Данная зависимость схематически изображена на рис. 4.

$$\begin{tikzcd}
 & & & X_1 \\
Y_1 \arrow[r, dash, "0.99"] & Y_{(3)} \arrow[r, dash, "0.96"] & X_{(3)} \arrow[ur, dash, "0.27"] \arrow[r, dash, "0.54"] \arrow[dr, dash, "0.80"'] & X_2 \\
 & & & X_3
\end{tikzcd}$$

Рис. 4. Графическое изображение зависимости Y_1 от X_1, X_2 и X_3
(над ребрами графа указаны значения соответствующих корреляций)

Как показывает приведенный пример, предлагаемый вниманию подход позволил заподозрить все зависимости (13)–(15) и при этом не указал на какие-либо ложные зависимости, которых не было при генерации данных.

Таким образом, примеры 2 и 3 свидетельствую в пользу того, что метод канонических корреляций может быть использован для выявления скрытых зависимостей между двумя группами переменных.

4. Заключение

В работе предложен альтернативный подход к изложению метода канонических корреляций, основанный на одном экстремальном свойстве квадратичных форм. Данный подход в техническом отношении проще по сравнению с традиционно применяемым методом множителей Лагранжа.

В работе приведены расчеты, которые показывают, что метод канонических корреляций может успешно применяться для выявления скрытых зависимостей между двумя группами переменных.

В приведенных расчетах (см. табл. 1, 2 и 3, 4) «значимые» канонические нагрузки видны «невооруженным глазом». В общем случае необходимо решающее правило, согласно которому каноническая нагрузка признается значимой. В литературе встречается эвристическое правило (см. [Sherry, Henson, 2005, p. 44]): каноническая нагрузка значима, если по модулю она превосходит критическое значение 0.45. Тем не менее в литературе формального обоснования данного или какого-либо еще решающего правила автору найти не удалось. Это могло бы быть предметом дальнейшего исследования.

Литература

Андерсон Т. Введение в многомерный статистический анализ. М.: Физматгиз, 1963.

Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: учебник. М.: Финансы и статистика, 2003.

Hamilton J.D. Time Series Analysis. Princeton University Press, 1994.

Hardoon D.R., Szedmak S., Shawe-Taylor J. Canonical Correlation Analysis: An Overview with Application to Learning Methods // Neural Computation. 2004. No. 16 (12).

Hotelling H. Relations Between Two Sets of Variates // Biometrika. 1936. No. 28 (3, 4).

Sherry A., Henson R. Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer //Journal of Personality Assessment. 2005. No. 84 (1).

Simonson D.G., Stowe J.D., Watson C.J. A Canonical Correlation Analysis of Commercial Bank Asset/Liability Structures // The Journal of Financial and Quantitative Analysis. 1983. No. 18 (1).

Thompson B. A Primer on the Logic and Use of Canonical Correlation Analysis // Measurement and Evaluation in Counseling and Development. 1991. No. 24 (2).

Ulyanov V.V., Fujikoshi Y., Shimizu R. Multivariate Statistics: High-Dimensional and Large-Sample Approximations. New Jersey: John Wiley & Son, 2010.

Borzykh, D. A.
On the method of canonical correlations [Electronic resource] : Working paper WP2/2016/01 / D. A. Borzykh ; National Research University Higher School of Economics. – Electronic text data (500 Kb). – Moscow : Higher School of Economics Publ. House, 2016. – (Series WP2 “Quantitative Analysis of Russian Economy”). – 18 p. (In Russian.).

The approach to the presentation of the method of canonical correlations based on one extremal property of quadratic forms is proposed. Traditionally, when describing the method of canonical correlation the method of Lagrange multipliers is used. On simulated data it is shown that canonical correlation analysis can be successfully applied for an identification of the hidden interdependencies between two groups of variables. This technique can be extremely useful at the preliminary stage of data analysis in a situation where a sufficiently large number of indicators can pretend to the role of the dependent variable in the regression.

Keywords: canonical correlations, canonical variables, canonical loadings

JEL classification: C38

Borzykh D.A., National Research University Higher School of Economics

Борзых Дмитрий Александрович

О методе канонических корреляций

Зав. редакцией оперативного выпуска *А.В. Заиченко*
Технический редактор *Ю.Н. Петрина*

Изд. № 1957
Национальный исследовательский университет
«Высшая школа экономики»