STATISTICS IN TRANSITION new series, March 2018
Vol. 19, No. 1, pp. 75–85, DOI 10.21307/stattrans-2018-005

CANONICAL CORRELATION ANALYSIS IN THE CASE OF MULTIVARIATE REPEATED MEASURES DATA

Mirosław Krzyśko[1], **Wojciech Łukaszonek**[2], **Waldemar Wołyński**[3]

ABSTRACT

In this paper, we present, in the real example, canonical variables applicable in the case of multivariate repeated measures data under the following assumptions: (1) multivariate normality for the vector of observations and (2) Kronecker product structure of the positive definite covariance matrix. These variables are especially useful when the number of observations is not large enough to estimate the covariance matrix, and thus the traditional canonical variables fail. Computational schemes for maximum likelihood estimates of required parameters are also given.

Key words: canonical correlation analysis, repeated measures data (doubly multivariate data), Kronecker product covariance structure, maximum likelihood estimates.

1. Introduction

Suppose that we have a sample of n objects characterized by $(p+q)$-variables $X_1,\ldots,X_p,X_{p+1},\ldots,X_{p+q}$ measured in T different time-points or physical conditions. Such data are often referred to in the statistical literature as multivariate repeated data or doubly multivariate data. Analysis of such data is complicated by the existence of correlation among the measurements of different variables as well as correlation among measurements taken at different time points. If we take observations on $(p+q)$-variables at T time-points, then these observations can be represented as $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q}$, where $\boldsymbol{x}_i$ are T-vectors. Let $\mathrm{Cov}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ be the covariance between the T-vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. When we choose $\mathrm{Cov}(\boldsymbol{x}_i,\boldsymbol{x}_j) = (\sigma_{ij}\boldsymbol{V})$, $i,j = 1\ldots,p,p+1,\ldots,p+q$, and assume normality, then the distribution of the $(p+q)$-random vectors is

$$\mathrm{vec}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q}) \sim N_{(p+q)T}(\boldsymbol{\mu},\boldsymbol{\Omega}),$$

where $\boldsymbol{\mu} = \mathrm{vec}(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_{p+q})$.

The covariance matrix $\boldsymbol{\Omega}$ is positive definite. Its estimate $\hat{\boldsymbol{\Omega}}$ is positive definite with probability 1 if and only if $n > (p+q)T$ (see, e.g., Giri (1996), p. 93).

[1]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: mkrzysko@amu.edu.pl

[2]Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Poland. E-mail: w.lukaszonek@g.pl

[3]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: wolynski@amu.edu.pl

Hence, estimation of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ will require a very large sample, which may not always be feasible. Hence, we assume $\boldsymbol{\Omega}$ to be of the form:

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{V},$$

where $\boldsymbol{\Sigma}$ is a $(p+q) \times (p+q)$ positive definite matrix and $\boldsymbol{V}$ is $T \times T$ positive definite matrix and $\boldsymbol{\Sigma} \otimes \boldsymbol{V}$ is the Kronecker product of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$. In this case the estimates of the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are positive definite with probability 1 if and only if $n > \max(p+q, T)$.

The matrix $\boldsymbol{\Sigma}$ represents the covariance between all $(p+q)$-variables on a given object and for a given time-point. Likewise, $\boldsymbol{V}$ represents the covariance between repeated measures on a given object and for a given variable. The above covariance structure is subject to an implicit assumption that for all variables the correlation structure between repeated measures remains the same, and that covariance between variables does not depend on time and remains constant for all time-points.

Classification rules in the case of multivariate repeated measures data under the assumption of multivariate normality for classes and with compound symmetric correlation structure on the matrix $\boldsymbol{V}$ were given by Roy and Khattree (2005). Next, Roy and Khattree (2008) gave the solution of this problem for the case when the correlation matrix $\boldsymbol{V}$ has the first order autoregressive [AR(1)] structure. Srivastava and Naik (2008), and McCollum (2010) describe the structure of canonical correlation and canonical variables (Hotelling 1936) based on the variables $X_1, \ldots, X_p$ and $X_{p+1}, \ldots, X_{p+q}$ observed at T time-points. Srivastava et al. (2008) found the form of maximum likelihood estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ and using these estimates gave the test of the hypothesis that the covariance matrix $\boldsymbol{\Omega}$ has the form $\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{V}$ against the alternative that the covariance matrix is not of Kronecker product structure, when $n > (p+q)T$. Krzyśko and Skorzybut (2009) and Krzyśko et al. (2011) proposed some new classification rules applicable in the case when no structures whatsoever are imposed on $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ except that they are positive definite. Deręgowski and Krzyśko (2009) constructed principal components for this model. Application of principal component analysis for the different types of genotypes of blackcurrants is described in the paper by Krzyśko et al. (2010), while the use of principal components to analyse a data set obtained from the experiments with varieties of winter rye is described in the paper Krzyśko et al. (2014).

The aim of this paper is to examine the relationship between the characteristics of higher education and the quality of life and human capital characteristics observed in 2002-2014 in each of the 16 Polish provinces. A particular model of canonical analysis, described in Srivastava and Naik (2008), will be used as a research tool.

In Section 2 we characterize our data set. In Section 3 canonical analysis for doubly multivariate data, based on results of Srivastava and Naik (2008), and McCollum (2010), is presented in the case when no structures whatsoever are imposed on $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ except that they are positive definite. In Section 4 we present the computational schemes for maximum likelihood estimates of unknown parameters.

In Section 5 the analysis results are presented.

2. Characteristics of the data set

The data used are taken from the Local Data Bank (https://bdl.stat.gov.pl). Local Data Bank is Poland's largest database containing data with respect to economy, households, innovation, public finance, society, demographics and the environment. The analysis relates to 16 Polish provinces $(n = 16)$. On the graphs, the provinces are denoted by numbers as given in Table 1.

Table 1. Designations of provinces

1	Dolnośląskie	9	Podkarpackie
2	Kujawsko-Pomorskie	10	Podlaskie
3	Lubelskie	11	Pomorskie
4	Lubuskie	12	Śląskie
5	Łódzkie	13	Świętokrzyskie
6	Małopolskie	14	Warmińsko-Mazurskie
7	Mazowieckie	15	Wielkopolskie
8	Opolskie	16	Zachodniopomorskie

The analysed data cover a period of 13 years, from 2002 to 2014 $(T = 13)$. Each province was characterized by two vectors of features:

$\boldsymbol{X}_1 = (X_1, \ldots, X_7)'$ characteristics of higher education $(p = 7)$

X_1 – The number of universities per 1 million inhabitants,
X_2 – The number of students per 1000 inhabitants,
X_3 – The number of university graduates per 1000 inhabitants,
X_4 – The number of academic teachers per 1000 inhabitants,
X_5 – The number of professors per 100 000 inhabitants,
X_6 – The number of post-graduate students per 10 000 inhabitants,
X_7 – The number of doctoral students per 10 000 inhabitants,

and

$\boldsymbol{X}_2 = (X_8, \ldots, X_{15})'$ quality of life and human capital characteristics $(q = 8)$

X_8 – Infant mortality rate per 1000 live births,
X_9 – Incidence of pulmonary tuberculosis per 100 000 inhabitants,
X_{10} – GDP per capita,
X_{11} – The Registered Unemployment Rate,
X_{12} – The percentage of inhabitants working in industry,
X_{13} – The percentage of inhabitants with university education,
X_{14} – The percentage of people learning and further training at the age of 25-69,
X_{15} – Employed in Research & Development per 1000 inhabitants.

3. Canonical analysis for doubly multivariate data

Let $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q}) = (\boldsymbol{X}_1,\boldsymbol{X}_2)$, where $\boldsymbol{X}_1 = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p)$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q})$ and let

$$\begin{aligned}\mathrm{Var}(\mathrm{vec}(\boldsymbol{X})) &= \mathrm{Var}\begin{pmatrix}\mathrm{vec}(\boldsymbol{X}_1)\\ \mathrm{vec}(\boldsymbol{X}_2)\end{pmatrix} = \boldsymbol{\Omega} = \boldsymbol{\Sigma}\otimes\boldsymbol{V}\\ &= \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\otimes\boldsymbol{V} = \begin{bmatrix}\boldsymbol{\Sigma}_{11}\otimes\boldsymbol{V} & \boldsymbol{\Sigma}_{12}\otimes\boldsymbol{V}\\ \boldsymbol{\Sigma}_{21}\otimes\boldsymbol{V} & \boldsymbol{\Sigma}_{22}\otimes\boldsymbol{V}\end{bmatrix},\end{aligned}$$

where $\boldsymbol{\Sigma}_{11}$ is $p\times p$ matrix.

In this model, the basis of canonical analysis are the eigenvalues and the eigenvectors of the matrices $\boldsymbol{A} = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\otimes\boldsymbol{I}_T$ and $\boldsymbol{B} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\otimes\boldsymbol{I}_T$ (Srivastava and Naik, 2008).

One of the main reasons for the use of the Kronecker product is a simple relationship between the eigenvalues and the eigenvectors $\boldsymbol{A}$ and $\boldsymbol{I}_T$ and $\boldsymbol{A}\otimes\boldsymbol{I}_T$ (see, e.g., Lancaster and Tismenetsky (1985), p. 412; Ortega (1987), p. 237). If $\alpha_1,\ldots,\alpha_p$ are the eigenvalues of $\boldsymbol{A}$ and $\beta_1,\ldots,\beta_T$ are the eigenvalues of $\boldsymbol{I}_T$, then eigenvalues of $\boldsymbol{A}\otimes\boldsymbol{I}_T$ are the pT numbers $\alpha_r\beta_s$, $r = 1,\ldots,p$, $s = 1,\ldots,T$. If $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{A}$ corresponding to the eigenvalues α, and $\boldsymbol{w}$ is an eigenvector of $\boldsymbol{I}_T$ corresponding to the eigenvalues β, then an eigenvector $\boldsymbol{\gamma}$ of $\boldsymbol{A}\otimes\boldsymbol{I}_T$ associated with $\alpha\beta$ is $\boldsymbol{\gamma} = \boldsymbol{u}\otimes\boldsymbol{w} = (\boldsymbol{u}_1\boldsymbol{w}',\boldsymbol{u}_2\boldsymbol{w}',\ldots,\boldsymbol{u}_p\boldsymbol{w}')'$.

Eigenvalues of matrix $\boldsymbol{I}_T$ are identical and equal to one. The corresponding eigenvectors are of the form:

$$\boldsymbol{w}_j = (0,\ldots,0,1,0,\ldots,0)',$$

where the only nonzero element is 1 in the $(T+1-j)$th position, $j = 1,2,\ldots,T$. Hence, the nonzero eigenvalues ρ_i^2 of matrix $\boldsymbol{A}\otimes\boldsymbol{I}_T$ are equal

$$\underbrace{\alpha_1,\ldots,\alpha_1}_{T\text{ times}},\underbrace{\alpha_2,\ldots,\alpha_2}_{T\text{ times}},\ldots,\underbrace{\alpha_k,\ldots,\alpha_k}_{T\text{ times}},$$

where $k = \mathrm{rank}(\boldsymbol{\Sigma}_{12}) \le \min(p,q)$.

The corresponding eigenvectors are of the form:

$$\boldsymbol{\gamma}_{ij} = \boldsymbol{u}_i\otimes\boldsymbol{w}_j,\ i = 1,\ldots,k,\ j = 1,\ldots,T.$$

The lth element of the vector $\boldsymbol{\gamma}_{ij}$ has the form:

$$\gamma_{ij,l} = \begin{cases} u_{il}, & \text{if } l = i(T+1-j),\\ 0, & \text{if } l \ne i(T+1-j),\end{cases}$$

$i = 1,\ldots,k$, $j = 1,\ldots,T$, $l = 1,\ldots,p$.

Nonzero values $\rho_1 \ge \rho_2 \ge \cdots \ge \rho_{kT}$, which are positive square roots of $\rho_1^2 \ge \rho_2^2 \ge$

$\cdots \geq \rho_{kT}^2$, are called the canonical correlations. The variables

$$U_{ij} = \boldsymbol{\gamma}_{ij}' \mathrm{vec}(\boldsymbol{X}_1),\ i = 1,\ldots,k,\ j = 1,\ldots,T,$$

are called the canonical variables in the space $\boldsymbol{X}_1$.

Similarly, if $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_{kT}^2$ are non-zero eigenvalues of the matrix $\boldsymbol{B}$ and $\boldsymbol{\psi}_{ij}$, $i = 1,\ldots,k,\ j = 1,\ldots,T$ are the corresponding eigenvectors, then the variables

$$V_{ij} = \boldsymbol{\psi}_{ij}' \mathrm{vec}(\boldsymbol{X}_2),\ i = 1,\ldots,k,\ j = 1,\ldots,T,$$

are called the canonical variables in the space $\boldsymbol{X}_2$.

In practice, the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are replaced by their estimators.

4. Maximum likelihood estimators of $\boldsymbol{\mu}$, Σ, and $\boldsymbol{V}$

We will use a different estimation method from the method used in Srivastava and Naik (2008), namely we will select the estimators obtained in Srivastava et al. (2008). For estimating the unknown parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$ we require n observations on the $T \times (p+q)$-matrix $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p,\boldsymbol{x}_{p+1},\ldots,\boldsymbol{x}_{p+q})$. These n observation matrices will be denoted by

$$\boldsymbol{X}_j = (\boldsymbol{x}_{1j},\ldots,\boldsymbol{x}_{pj},\boldsymbol{x}_{p+1,j},\ldots,\boldsymbol{x}_{p+q,j}),\ j = 1,\ldots,n.$$

Let

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j,\ \boldsymbol{X}_{j,c} = \boldsymbol{X}_j - \bar{\boldsymbol{X}},\ j = 1,\ldots,n.$$

We consider a model, denoted by I, described as follows: all observations $\boldsymbol{X}_j$ are independent and $\mathrm{vec}(\boldsymbol{X}_j) \sim N_{(p+q)T}(\boldsymbol{\mu},\boldsymbol{\Sigma}\otimes\boldsymbol{V})$, where $\boldsymbol{\Sigma}$ is $(p+q)\times(p+q)$ positive definite covariance matrix and $\boldsymbol{V}$ is $T \times T$ positive definite covariance matrix, $j = 1,\ldots,n$, $n > \max(p+q,T)$. The maximum likelihood estimation equations are of the form (Srivastava et al. 2008):

$$\hat{\boldsymbol{\mu}} = \mathrm{vec}(\bar{\boldsymbol{X}}) \tag{1}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{nT}\sum_{j=1}^{n} \boldsymbol{X}_{j,c}' \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}_{j,c}, \tag{2}$$

$$\hat{\boldsymbol{V}} = \frac{1}{n(p+q)}\sum_{j=1}^{n} \boldsymbol{X}_{j,c} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_{j,c}'. \tag{3}$$

In this case no explicit maximum likelihood estimates (MLEs) of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are available. The MLEs of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are obtained by solving simultaneously and iteratively the equations (2) and (3). This is the so-called "flip-flop" algorithm.

In this model (model I), if $n > \max(p+q, T)$ then the maximum likelihood estimation equations given by (2) and (3) will always converge to the unique maximum (Srivastava et al. 2008).

The following iterative steps are suggested to obtain the maximum likelihood estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$.

Step 1. Get the initial covariance matrix $\boldsymbol{V}$ in the form

$$\hat{\boldsymbol{V}} = \frac{1}{n(p+q)} \sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'. \tag{4}$$

Step 2. On the basis of the initial covariance matrix $\hat{\boldsymbol{V}}$ compute the matrix $\hat{\boldsymbol{\Sigma}}$ given by (2).

Step 3. Compute the matrix $\hat{\boldsymbol{V}}$ from equation (3) using the matrix $\hat{\boldsymbol{\Sigma}}$ obtained in Step 2.

Step 4. Repeat Steps 2 and 3 until convergence is attained. We have selected the following stopping rule. Compute two matrices: (a) a matrix of difference between two successive solutions of $\hat{\boldsymbol{\Sigma}}$, and (b) a matrix of difference between two successive solutions of $\hat{\boldsymbol{V}}$. Continue the iteration procedure until the maxima of the absolute values of the elements of the matrices in (a) and (b) are smaller than the prespecified quantities.

As noted in the literature (see, e.g., Galecki (1994); Naik and Rao (2001)), since

$$(c\boldsymbol{\Sigma}) \otimes (c^{-1}\boldsymbol{V}) = \boldsymbol{\Sigma} \otimes \boldsymbol{V},$$

all the parameters of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ are not defined uniquely. But in our case, estimation of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{V}$ is not an aim in itself.

The resulting estimates are used to construction the canonical variables. Canonical variables considered in this paper are functions of $\hat{\boldsymbol{\Sigma}} \otimes \hat{\boldsymbol{V}}$, instead of $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{V}}$ separately, and hence only $\hat{\boldsymbol{\Sigma}} \otimes \hat{\boldsymbol{V}}$ needs to be unique under the model (I) with $\boldsymbol{\Sigma} > 0$ and $\boldsymbol{V} > 0$.

5. Analysis results

We are interested in the relationship between the vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. We build estimators of the matrices $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ and $\boldsymbol{\Sigma}_{12}$, and we then find the non-zero eigenvalues $\hat{\alpha}_k$ and corresponding eigenvectors $\hat{\boldsymbol{u}}_k$ of the matrix $\hat{\boldsymbol{A}} = \hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}\hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}$, and the non-zero eigenvalues $\hat{\alpha}_k$ and corresponding eigenvectors $\hat{\boldsymbol{v}}_k$ of the matrix $\hat{\boldsymbol{B}} = \hat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\boldsymbol{\Sigma}}_{21}\hat{\boldsymbol{\Sigma}}_{11}^{-1}\hat{\boldsymbol{\Sigma}}_{12}$, where $\hat{\boldsymbol{\Sigma}}_{21} = \hat{\boldsymbol{\Sigma}}_{12}'$, $k = 1, \ldots, 7 = \min(p, q)$.

The non-zero eigenvalues $\hat{\alpha}_k$ are shown in Figure 1.

The characteristics of higher education and quality of life and human capital characteristics are moderately correlated with the canonical correlation coefficient $\hat{\rho}_1 = 0.38$.

Eigenvectors $\hat{\boldsymbol{u}}_k$ of the matrix $\hat{\boldsymbol{A}}$ are shown in Table 2, and eigenvectors $\hat{\boldsymbol{v}}_k$ of the matrix $\hat{\boldsymbol{B}}$ are shown in Table 3.

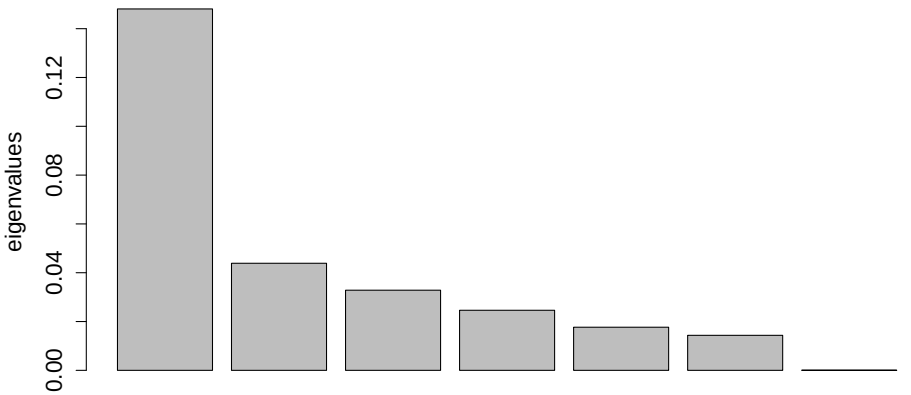

Figure 1: The non-zero eigenvalues $\hat{\alpha}_k$

Table 2. Eigenvectors $\hat{\boldsymbol{u}}_k$ of the matrix $\hat{\boldsymbol{A}}$

	$\hat{\boldsymbol{u}}_1$	$\hat{\boldsymbol{u}}_2$	$\hat{\boldsymbol{u}}_3$	$\hat{\boldsymbol{u}}_4$	$\hat{\boldsymbol{u}}_5$	$\hat{\boldsymbol{u}}_6$	$\hat{\boldsymbol{u}}_7$
1	-0.0703	0.0564	-0.0274	0.0037	0.1379	-0.0045	-0.0086
2	-0.0159	0.0451	0.0054	-0.0045	-0.0140	-0.0130	0.0055
3	-0.0015	-0.1640	0.0371	0.0010	0.0312	-0.0068	0.0044
4	0.1027	0.1848	0.3343	0.4695	-0.4270	-0.0586	-0.1198
5	-0.6819	-0.9531	-0.7931	-0.8710	0.8034	0.9929	0.9859
6	-0.0857	-0.1140	-0.0884	0.0079	-0.0762	-0.0032	-0.0082
7	-0.7155	0.1108	0.4992	-0.1446	-0.3825	0.1021	-0.1163

Table 3. Eigenvectors $\hat{\boldsymbol{v}}_k$ of the matrix $\hat{\boldsymbol{B}}$

	$\hat{\boldsymbol{v}}_1$	$\hat{\boldsymbol{v}}_2$	$\hat{\boldsymbol{v}}_3$	$\hat{\boldsymbol{v}}_4$	$\hat{\boldsymbol{v}}_5$	$\hat{\boldsymbol{v}}_6$	$\hat{\boldsymbol{v}}_7$
1	-0.5126	-0.2076	0.1924	0.1080	0.2955	0.3694	0.2669
2	-0.0164	0.0178	0.0785	-0.0231	0.1125	-0.0218	0.1561
3	0.0008	0.0002	0.0004	-0.0004	0.0001	0.0002	0.0001
4	-0.4536	0.4884	0.3444	-0.5500	-0.4824	0.3384	-0.3656
5	0.0428	-0.2987	-0.1727	-0.2816	-0.2534	0.0425	0.3910
6	-0.0445	-0.7583	0.2534	-0.2278	0.1240	-0.1116	-0.7477
7	0.5255	0.2089	-0.7506	0.2628	0.4246	0.8400	-0.1666
8	0.5013	-0.1010	0.4252	0.6964	-0.6383	0.1696	0.1762

Eigenvalues of matrix $\boldsymbol{I}_{13}$ are identical and equal to one. The corresponding eigenvectors are of the form $\boldsymbol{w}_j = (0,\ldots,1,\ldots,0)'$, where the only non-zero element is the number 1 in the $(14-j)$th position, $j = 1,\ldots,13$.

Hence, the eigenvalues of matrix $\hat{\boldsymbol{A}} \otimes \boldsymbol{I}_{13}$ are equal

$$\underbrace{\hat{\rho}_1^2,\ldots,\hat{\rho}_1^2}_{13 \text{ times}},\underbrace{\hat{\rho}_2^2,\ldots,\hat{\rho}_2^2}_{13 \text{ times}},\ldots,\underbrace{\hat{\rho}_7^2,\ldots,\hat{\rho}_7^2}_{13 \text{ times}}.$$

The corresponding eigenvectors are of the form:

$$\hat{\boldsymbol{\gamma}}_{ij} = \hat{\boldsymbol{u}}_i \otimes \boldsymbol{w}_j,\ i = 1,\ldots,7,\ j = 1,\ldots,13.$$

The lth element of the vector $\hat{\boldsymbol{\gamma}}_{ij}$ has the form:

$$\hat{\gamma}_{ij,l} = \begin{cases} \hat{u}_{il}, & \text{if } l = i(14-j), \\ 0, & \text{if } l \neq i(14-j), \end{cases}$$

$i = 1,\ldots,7,\ j = 1,\ldots,13,\ l = 1,\ldots,7.$

Replacing the vectors $\hat{\boldsymbol{u}}_i$ by $\hat{\boldsymbol{v}}_i$ we obtain similar results for the matrix $\hat{\boldsymbol{B}} \otimes \boldsymbol{I}_{13}$.

If we choose the system $(U_{1,13}, V_{1,13})$, then the location of the various provinces in this system include data from 2002 (see Fig. 2). If, however, we choose the system (U_{11}, V_{11}), then the location of the various provinces in this system includes data from 2014 (see Fig. 3). It results from the vectors $\boldsymbol{\gamma}_{1j}$ and $\boldsymbol{\psi}_{1j}$, $j = 1,\ldots,T$.

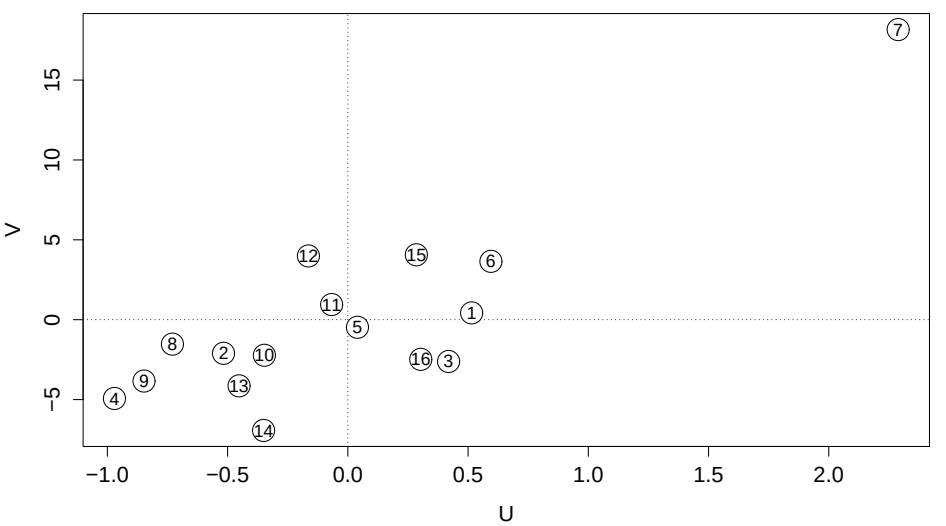

Figure 2: The location of the various provinces in 2002

For example

$$\boldsymbol{\gamma}_{11} = (0,\ldots,u_{11},0,\ldots,u_{12},\ldots,0,\ldots,u_{17})',$$

and

$$\boldsymbol{\gamma}_{1,13} = (u_{11},\ldots,0,\ldots,u_{12},\ldots,0,\ldots,u_{17},\ldots,0)'.$$

Note that all of these systems of canonical variables correspond to the same value of the canonical correlation coefficient $\hat{\rho}_1 = 0.38$.

In Fig. 2, on the one hand, one can see provinces with bad (low) values of characteristics of higher education and bad (low) values of quality of life and human capital characteristics such as Lubuskie (4), Podkarpackie (9) and Opolskie (8), and on the other hand, one can see provinces with high values of characteristics of higher education and good (high) values of quality of life and human capital characteristics such as Mazowieckie (7), Małopolskie (6), and Dolnośląskie (1). The absolute leader is Mazowieckie (7) province. Comparing 2014 to 2002 in Fig. 3, a change in the location of some provinces can be observed. For example, Opolskie (8) and Pomorskie (11) provinces improved their position, but the position of Zachodniopomorskie (16) and Warmińsko-Mazurskie (14) deteriorated.

During the calculations we used R (R Core Team (2017)) software. The R source code is available on request at the third co-author.

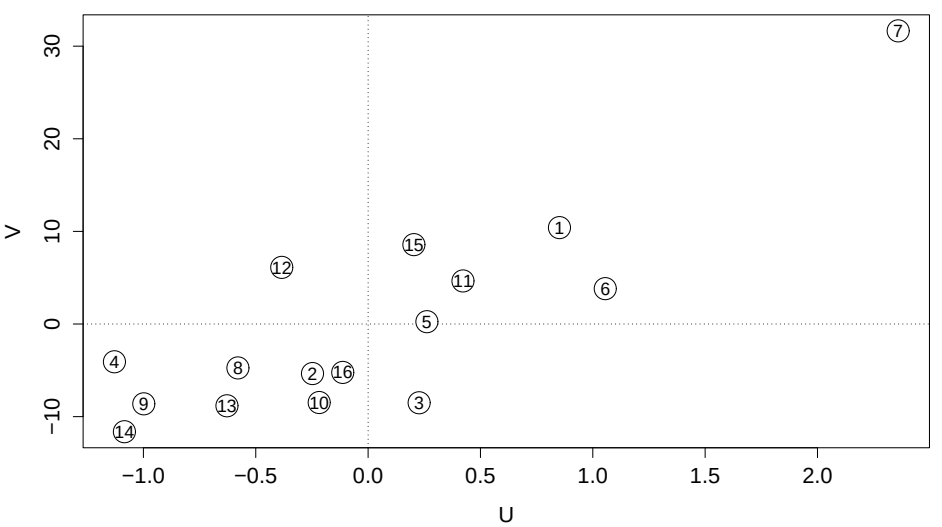

Figure 3: The location of the various provinces in 2014

REFERENCES

DERĘGOWSKI, K., KRZYŚKO, M., (2009). Principal component analysis in the case of multivariate repeated measures data, Biometrical Letters, 46 (2), pp. 163–172.

GALECKI, A. T., (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis, Communications in Statistics – Theory and Methods, 23, pp. 3105-3119.

GIRI, N. C., (1996). Multivariate Statistical Analysis, Marcel Dekker, New York.

HOTELLING, H., (1936). Relations between two sets of variates, Biometrika, 28, pp. 321–377.

KRZYŚKO, M., SKORZYBUT, M., (2009). Discriminant analysis of multivariate repeated measures data with a Kronecker product structured covariance matrices, Statistical papers, 50, 817-835.

KRZYŚKO, M., MĄDRY, W., PLUTA, S., SKORZYBUT, M., WOŁYŃSKI, W., (2010). Analysis of multivariate repeated measures data, Colloquium Biometricum, 40, pp. 117–133.

KRZYŚKO, M., SKORZYBUT, M., WOŁYŃSKI, W., (2011). Classifiers for doubly multivariate data, Discussiones Mathematicae. Probability and Statistics, 31, pp. 5–27.

KRZYŚKO, M., ŚMIAŁOWSKI, T., WOŁYŃSKI, W., (2014). Analysis of multivariate repeated measures data using a MANOVA model and principal components, Biometrical Letters, 51 (2), pp. 103–124.

LANCASTER, P., TISMENETSKY, M., (1985). The Theory of Matrices, Second Edition: With Applications. Academic Press, Orlando.

MCCOLLUM, R., (2010). Canonical correlation analysis for longitudinal data. Ph.D. thesis, Old Dominion University.

NAIK, D. N., RAO, S., (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix, J. Appl. Statist., 28, pp. 91–105.

ORTEGA, J. M., (1987). Matrix Theory: A Second Course. Plenum Press, New York.

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

ROY, A., KHATTREE, R., (2005). On discrimination and classification with multivariate repeated measures data, Journal of Statistical Planning and Inference, 134, pp. 462–485.

ROY, A., KHATTREE, R., (2008). Classification rules for repeated measures data from biomedical research. In: Khattree, R., Naik, D. N. (eds) Computational methods in biomedical research, Chapman and Hall/CRC, pp. 323–370.

SRIVASTAVA, J., Naik, D. N., (2008). Canonical correlation analysis of longitudinal data, Denver JSM 2008 Proceedings, Biometrics Section, pp. 563–568.

SRIVASTAVA, M.S., VON ROSEN, T., VON ROSEN, D., (2008). Models with a Kronecker product covariance structure: estimation and testing, Math. Methods Stat., 17 (4), pp. 357–370.