DOI: 10.2478/bile-2020-0001

Biometrical Letters
Vol. 57 (2020), No. 1, 1-12

Generalized canonical correlation analysis for functional data

Tomasz Górecki[1], **Mirosław Krzyśko**[2], **Waldemar Wołyński**[1]

[1]Faculty of Mathematics and Computer Science, Adam Mickiewicz University,
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland,
e-mail: tomasz.gorecki@amu.edu.pl, wolynski@amu.edu.pl
[2]Interfaculty Institute of Mathematics and Statistics, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz, Nowy Świat 4, 62-800 Kalisz, Poland, e-mail: mkrzysko@amu.edu.pl

Summary

There is a growing need to analyze data sets characterized by several sets of variables observed on the same set of individuals. Such complex data structures are known as multiblock (or multiple-set) data sets. Multiblock data sets are encountered in diverse fields including bioinformatics, chemometrics, food analysis, etc. Generalized Canonical Correlation Analysis (GCCA) is a very powerful method to study this kind of relationships between blocks. It can also be viewed as a method for the integration of information from $K > 2$ distinct sources (Takane and Oshima-Takane 2002). In this paper, GCCA is considered in the context of multivariate functional data. Such data are treated as realizations of multivariate random processes. GCCA is a technique that allows the joint analysis of several sets of data through dimensionality reduction. The central problem of GCCA is to construct a series of components aiming to maximize the association among the multiple variable sets. This method will be presented for multivariate functional data. Finally, a practical example will be discussed.

Key words: multivariate functional data, generalized canonical correlation analysis

1. Introduction

For studying interrelations between two sets of variables, Canonical Correlation Analysis (CCA), proposed by Hotelling (1936), is often used. It consists in determining a linear transformation of the original variables from both sets into two new sets of variables not correlated within the sets but most highly correlated between them. Pairs of the corresponding new variables are called canonical variates, and the coefficients of correlation within the pairs are called canonical correlations.

Carroll (1968a, 1968b) proposed a generalized canonical correlation analysis. In generalized canonical correlation analysis, $K > 2$ sets of variables are analyzed simultaneously. The central problem of GCCA is to construct a series of components aiming to maximize the association among the multiple variable sets. Although several generalizations of canonical correlation analysis have been proposed, some of which are discussed and compared in Kettenring (1971) and Gower (1989), Carroll's approach has some attractive properties that make the method well suited to the analysis of multiple-set data (van de Velden 2011):

1. Computationally, the method is straightforward and its solution is based on an eigen-analysis.
2. The method is closely related to several well-known multivariate techniques such as principal component analysis, partial least squares and multivariate linear regression.
3. When the number of data sets $K = 2$, Carroll's GCCA reduces to the usual canonical correlation analysis.

In recent years, methods for representing data by functions have received much attention. Such data are known in the literature as functional data (Ramsay and Silverman 2005, Horváth and Kokoszka 2012). Examples of functional data can be found in various application domains, such as medicine, economics, meteorology and many others. Canonical correlation analysis for one-dimensional functional data was described by Leurgans et al. (1993) and Ramsay and Silverman (2005, Chapter 11). Despite its usefulness, Functional Canonical Correlation Analysis (FCCA) is limited to the analysis of two functional data sets. An extension of one-dimensional FCCA to the analysis of more than two sets of functional data was proposed by Hwang et al. (2012, 2013). Note that Hotelling's classic method assumes that the considered objects of two sets of data are characterized by many variables, whereas the referenced papers on functional data consider one-dimensional data. There is a discrepancy between the assumptions associated with the classical method and functional methods. Canonical Correlation Analysis for Multivariate Functional data (MFCCA) was described by Górecki et al. (2017, 2018). This paper considers a Generalized Canonical Correlation Analysis for Multivariate Functional data (MFGCCA).

The paper is organized as follows. Section 2 contains a review of the usual canonical correlation analysis. Section 3 presents the generalized canonical correlation analysis given by Carroll (1968a, 1968b). A process of transformation of discrete data into functional data is described in Section 4. A generalized canonical correlation analysis for multivariate functional data is presented in Section 5. Section 6 contains a real example of the proposed methodology. Concluding remarks are given in Section 7.

2. Canonical correlation analysis

Canonical correlation analysis (Hotelling 1936) is the study of the linear relations between two blocks of variables.

Let $\boldsymbol{X}_1 = (X_{11}, \dots, X_{1p_1})^\top$ and $\boldsymbol{X}_2 = (X_{21}, \dots, X_{2p_2})^\top$ denote blocks of random variables (random vectors) with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$. Without loss of generality we can assume that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. A superblock is defined as the concatenation of these blocks (Tenenhaus and Tenenhaus 2011, Tenenhaus et al. 2017b). Let the superblock $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$ have a covariance matrix of the form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Pairs of canonical variables (U_{1i}, U_{2i}), $i = 1, 2, \dots, s$, $s = \operatorname{rank}(\boldsymbol{\Sigma}_{12})$ are defined via the pairs of linear combinations of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$:

$$U_{1i} = \boldsymbol{l}_{1i}^\top \boldsymbol{X}_1, \ U_{2i} = \boldsymbol{l}_{2i}^\top \boldsymbol{X}_2$$

that maximize the correlation between U_{1i} and U_{2i}, i.e. that maximize

$$\operatorname{Corr}(U_{1i}, U_{2i}) = \operatorname{Corr}(\boldsymbol{l}_{1i}^\top \boldsymbol{X}_1, \boldsymbol{l}_{2i}^\top \boldsymbol{X}_2) \tag{1}$$

subject to U_{1i} and U_{2i} having unit variance while for $1 \leqslant i_1 < i_2 \leqslant s$ the vectors (U_{1i_1}, U_{2i_1}) and (U_{1i_2}, U_{2i_2}) are uncorrelated.

Let $\boldsymbol{U}_1 = (U_{11}, \dots, U_{1s})^\top$ be the vector containing the s canonical variables from $\boldsymbol{X}_1$, and $\boldsymbol{U}_2 = (U_{21}, \dots, U_{2s})^\top$ be the vector containing the s canonical variables from $\boldsymbol{X}_2$. Moreover, let $\boldsymbol{L}_1 = (\boldsymbol{l}_{11}, \dots, \boldsymbol{l}_{1s})$, $\boldsymbol{L}_2 = (\boldsymbol{l}_{21}, \dots, \boldsymbol{l}_{2s})$. Then

$$\operatorname{Var}\begin{bmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_1^\top \boldsymbol{\Sigma}_{11} \boldsymbol{L}_1 & \boldsymbol{L}_1^\top \boldsymbol{\Sigma}_{12} \boldsymbol{L}_2 \\ \boldsymbol{L}_2^\top \boldsymbol{\Sigma}_{21} \boldsymbol{L}_1 & \boldsymbol{L}_2^\top \boldsymbol{\Sigma}_{22} \boldsymbol{L}_2 \end{bmatrix},$$

$\boldsymbol{L}_1^\top \boldsymbol{\Sigma}_{11}\boldsymbol{L}_1 = \boldsymbol{I}_s$, $\boldsymbol{L}_2^\top \boldsymbol{\Sigma}_{22}\boldsymbol{L}_2 = \boldsymbol{I}_s$, and

$$\boldsymbol{L}_1^\top \boldsymbol{\Sigma}_{12}\boldsymbol{L}_2 = \boldsymbol{P},$$

where $\boldsymbol{P}$ is the diagonal matrix with diagonal elements $\rho_1, \ldots, \rho_s$ (called the canonical correlation coefficients), so that $\rho_1^2 \geqslant \cdots \geqslant \rho_s^2$, and $\rho_i = \mathrm{Corr}(U_{1i}, U_{2i})$, $i = 1, \ldots, s$.

If we write $\boldsymbol{U} = \boldsymbol{L}^\top \boldsymbol{X}$, where $\boldsymbol{L} = (\boldsymbol{L}_1^\top, \boldsymbol{L}_2^\top)^\top$, then

$$\mathrm{Var}(\boldsymbol{U}) = \boldsymbol{L}^\top \boldsymbol{\Sigma}\boldsymbol{L} = \boldsymbol{L}_1^\top \boldsymbol{\Sigma}_{11}\boldsymbol{L}_1 + \boldsymbol{L}_2^\top \boldsymbol{\Sigma}_{22}\boldsymbol{L}_2 + 2\boldsymbol{L}_1^\top \boldsymbol{\Sigma}_{12}\boldsymbol{L}_2,$$

and the problem of maximizing the expression (1) is equivalent to the problem of maximizing

$$\phi(\boldsymbol{L}) = \mathrm{tr}(\boldsymbol{L}^\top \boldsymbol{\Sigma}\boldsymbol{L}),$$

subject to $\boldsymbol{L}^\top \boldsymbol{D}\boldsymbol{L} = \boldsymbol{I}_s$, where

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

3. Generalized canonical correlation analysis

Now, we consider a generalized version of canonical correlation analysis (Carroll 1968a, 1968b) that allows the analysis of several blocks of variables simultaneously.

Let $\boldsymbol{X}_k = (X_{k1}, \ldots, X_{kp_k})^\top$ denote the blocks of random variables (random vectors) with zero mean vectors and covariance matrices $\boldsymbol{\Sigma}_{kk}$, $k = 1, \ldots, K$. Moreover, let the superblock $\boldsymbol{X}$ have the form $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_K^\top)^\top$, and

$$\mathrm{Var}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1K} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & & \boldsymbol{\Sigma}_{2K} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\Sigma}_{K1} & \boldsymbol{\Sigma}_{K2} & \cdots & \boldsymbol{\Sigma}_{KK} \end{bmatrix}.$$

Now we seek vectors of canonical variables $(U_{1i}, \ldots, U_{Ki})$, $i = 1, 2, \ldots, s$, $s = \min\limits_{k \neq j} \mathrm{rank}(\boldsymbol{\Sigma}_{kj})$ being linear combinations of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ respectively. For $i = 1, 2, \ldots, s$, the canonical variables maximize the sum of the correlations between each pair of them, i.e. they maximize

$$\sum_{k,j=1,k<j}^{K} \mathrm{Corr}(U_{ki}, U_{ji}) \tag{2}$$

subject to U_{ki} having unit variances, $k = 1, 2, \ldots, K$. Moreover, for $1 \leqslant i_1 < i_2 \leqslant s$ the vectors $(U_{1i_1}, \ldots, U_{Ki_1})$ and $(U_{1i_2}, \ldots, U_{Ki_2})$ are uncorrelated.

Let us write $U_{ki} = \boldsymbol{l}_{ki}^\top \boldsymbol{X}_k$, $\boldsymbol{U}_k = (U_{k1}, \ldots, U_{ks})^\top$, $\boldsymbol{L}_k = (\boldsymbol{l}_{k1}, \ldots, \boldsymbol{l}_{ks})$, $k = 1, 2, \ldots, K$, $i = 1, \ldots, s$. Then

$$\boldsymbol{U}_k = \boldsymbol{L}_k^\top \boldsymbol{X}_k, \ k = 1, 2, \ldots, K.$$

Moreover, let $\boldsymbol{U} = \boldsymbol{L}^\top \boldsymbol{X}$, where $\boldsymbol{L} = (\boldsymbol{L}_1^\top, \ldots, \boldsymbol{L}_K^\top)^\top$. We have

$$\operatorname{Var}(\boldsymbol{U}) = \boldsymbol{L}^\top \boldsymbol{\Sigma} \boldsymbol{L} = \sum_{k=1}^{K} \boldsymbol{L}_k^\top \boldsymbol{\Sigma}_{kk} \boldsymbol{L}_k + 2 \sum_{k,j,k<j}^{K} \boldsymbol{L}_k^\top \boldsymbol{\Sigma}_{kj} \boldsymbol{L}_j.$$

Similarly to the classical case, the main problem of generalized canonical correlation analysis may be formulated as that of maximizing

$$\phi(\boldsymbol{L}) = \operatorname{tr}(\boldsymbol{L}^\top \boldsymbol{\Sigma} \boldsymbol{L}),$$

subject to

$$\boldsymbol{L}^\top \boldsymbol{D} \boldsymbol{L} = \boldsymbol{I}_s, \tag{3}$$

where $\boldsymbol{D}$ is a block diagonal matrix formed with the matrices $\boldsymbol{\Sigma}_{kk}$ as the kth diagonal block.

This leads to a generalized eigenequation of the form:

$$\boldsymbol{\Sigma} \boldsymbol{L} = \boldsymbol{D} \boldsymbol{L} \boldsymbol{\Delta}^2,$$

where $\boldsymbol{\Delta}^2$ is the diagonal matrix consisting of the s largest generalized eigenvalues of $\boldsymbol{\Sigma}$ with respect to the matrix $\boldsymbol{D}$, and $\boldsymbol{L}$ is the matrix of the corresponding generalized eigenvectors.

The details of methods for solving the maximizing problem in generalized canonical correlation analysis can be found in Takane et al. (2008) and Makos and D'enza (2016).

4. The functional data

Now, assume that we wish to analyze several multi-dimensional random processes $\boldsymbol{X}_k(t) = (X_{k1}(t), \ldots, X_{kp_k}(t))^\top \in L_2^{p_k}(I)$, $t \in I$, $k = 1, \ldots, K$, where $L_2(I)$ is the Hilbert space of square-integrable functions. Moreover, assume that the lth component of the vector $\boldsymbol{X}_k(t)$ can be represented by a finite number of orthonormal basis functions $\{\varphi_b(t)\}$, where $\varphi_b(t) \in L_2(I),\ t \in I$. Omitting the index k, we may write:

$$X_l(t) = \sum_{b=0}^{B_l} c_{lb}\varphi_b(t),\ t \in I,\ l = 1, \dots, p, \tag{4}$$

where $c_{l0}, c_{l1}, \dots, c_{lB_l}$ are the unknown coefficients.

Let $\boldsymbol{c} = (c_{10}, \dots, c_{1B_1}, \dots, c_{p0}, \dots, c_{pB_p})^\top$ and

$$\boldsymbol{\Phi}(t) = \begin{bmatrix} \boldsymbol{\varphi}_1^\top(t) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}_2^\top(t) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\varphi}_p^\top(t) \end{bmatrix},$$

where $\boldsymbol{\varphi}_l(t) = (\varphi_0(t), ..., \varphi_{B_l}(t))^\top$, $l = 1, ..., p$. Using the above matrix notation, process $\boldsymbol{X}(t)$ can be represented as:

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c}, \tag{5}$$

where $\mathrm{E}(\boldsymbol{c}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{c}) = \boldsymbol{\Sigma}_c$. This means that the realizations of the process $\boldsymbol{X}(t)$ lie in a finite-dimensional subspace of $L_2^p(I)$.

We can estimate the vector $\boldsymbol{c}$ on the basis of n independent realizations $\boldsymbol{x}_1(t), \boldsymbol{x}_2(t), \dots, \boldsymbol{x}_n(t)$ of the random process $\boldsymbol{X}(t)$ (functional data). As a method of estimation we use the least squares method.

Typically, data are recorded at discrete moments in time. Let x_{lj} denote an observed value of the feature X_l, $l = 1, 2, \dots, p$ at the jth time point t_j, where $j = 1, 2, ..., J$. Then our data consist of the pJ pairs (t_j, x_{lj}). These discrete data can be smoothed by continuous functions $x_l(t)$, and I is a compact set such that $t_j \in I$, for $j = 1, ..., J$. Details of the process of transformation of discrete data to functional data can be found in e.g. Ramsay and Silverman (2005) or Górecki et al. (2018).

5. Generalized canonical correlation analysis for functional data

In the case of random processes, we define the canonical variables $U_{1i}, \dots, U_{Ki}$ as dot products, i.e.

$$U_{ki} = \langle \boldsymbol{l}_{ki}, \boldsymbol{X}_k \rangle = \int_I \boldsymbol{l}_{ki}^\top(t)\boldsymbol{X}_k(t)dt,$$

where $\boldsymbol{l}_{ki}(t) \in L_2^{p_k}(I)$, $k = 1, \ldots, K$. Let the vector weight function $\boldsymbol{l}_{ki}(t)$ and the process $\boldsymbol{X}_k(t)$ be in the same space, i.e. the function $\boldsymbol{l}_{ki}(t)$ can be written in the form

$$\boldsymbol{l}_{ki}(t) = \boldsymbol{\Phi}_k(t)\boldsymbol{\lambda}_{ki}, \tag{6}$$

where $\boldsymbol{\lambda}_{ki} \in \mathbb{R}^{B_{k1}+\ldots+B_{kp_k}+p_k}$.

Hence

$$\langle \boldsymbol{l}_{ki}, \boldsymbol{X}_k \rangle = \boldsymbol{\lambda}_{ki}^\top \left[\int_I \boldsymbol{\Phi}_k^\top(t)\boldsymbol{\Phi}_k(t)dt \right] \boldsymbol{c}_k = \boldsymbol{\lambda}_{ki}^\top \boldsymbol{c}_k,$$

where $\boldsymbol{c}_k$ and $\boldsymbol{\lambda}_{ki}$ are vectors occurring in the representations (5) and (6) of process $\boldsymbol{X}_k(t)$ and function $\boldsymbol{l}_{ki}(t)$, $k = 1, \ldots, K$. Thus, our problem may be reduced to a problem involving only random vectors $\boldsymbol{c}_k$ and $\boldsymbol{\lambda}_{ki}$. Let

$$\boldsymbol{\Lambda}_k = (\boldsymbol{\lambda}_{k1}, \ldots, \boldsymbol{\lambda}_{ks}), \quad \boldsymbol{U}_k = \boldsymbol{\Lambda}_k^\top \boldsymbol{c}_k, \ k = 1, 2, \ldots, K,$$

and

$$\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1^\top, \ldots, \boldsymbol{\Lambda}_K^\top)^\top, \quad \boldsymbol{c} = (\boldsymbol{c}_1^\top, \ldots, \boldsymbol{c}_K^\top)^\top.$$

Then

$$\boldsymbol{U} = \boldsymbol{\Lambda}^\top \boldsymbol{c},$$

and the case of functional data reduces to the vector data considered in Section 3.

Note that the canonical weight functions $\boldsymbol{l}_{ki}(t)$, $k = 1, \ldots, K$ do not give any meaningful information about the data and clearly demonstrate the need for a technique involving smoothing. A straightforward way of introducing smoothing is to modify the constraints (3) by adding roughness penalty terms (Górecki et al. 2018) to give:

$$\mathrm{Var}(U_{ki}) + \lambda\, \mathrm{PEN}_2(\boldsymbol{l}_{ki}) = 1,$$

where the roughness function PEN_2 is the integrated squared second derivative

$$\mathrm{PEN}_2(\boldsymbol{l}_{ki}) = \int_I \left(\frac{\partial^2 \boldsymbol{l}_{ki}(t)}{\partial t^2} \right)^\top \frac{\partial^2 \boldsymbol{l}_{ki}(t)}{\partial t^2} dt = \boldsymbol{\lambda}_{ki}^\top \boldsymbol{R}_k \boldsymbol{\lambda}_{ki},$$

where

$$\boldsymbol{R}_k = \int_I \left(\frac{\partial^2 \boldsymbol{\Phi}_k(t)}{\partial t^2} \right)^\top \frac{\partial^2 \boldsymbol{\Phi}_k(t)}{\partial t^2} dt, \ k = 1, \ldots, K.$$

Therefore, the aim of MFGCCA is to obtain the matrix $\boldsymbol{\Lambda}$ which maximizes

$$\phi(\boldsymbol{\Lambda}) = \operatorname{tr}(\boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}_c \boldsymbol{\Lambda})$$

subject to $\boldsymbol{\Lambda}^\top(\boldsymbol{D} + \lambda \boldsymbol{R})\boldsymbol{\Lambda} = \boldsymbol{I}_s$, where $\boldsymbol{D}$ is a block diagonal matrix formed of the matrices $\boldsymbol{\Sigma}_{ckk}$, $\boldsymbol{R}$ is a diagonal matrix formed of the matrices $\boldsymbol{R}_k$, and $\lambda > 0$ is a penalty parameter.

The coordinates of the projection of the lth realization $\boldsymbol{x}_{1l}(t), \ldots, \boldsymbol{x}_{Kl}(t)$ of the processes $\boldsymbol{X}_1(t), \ldots, \boldsymbol{X}_K(t)$ onto the space spanned by the superblock components are equal to

$$\hat{\boldsymbol{U}}_l = \hat{\boldsymbol{\Lambda}}^\top \hat{\boldsymbol{c}}_l,$$

where $\hat{\boldsymbol{c}}_l = (\hat{\boldsymbol{c}}_{1l}^\top, \ldots, \hat{\boldsymbol{c}}_{Kl}^\top)^\top$ and $\hat{\boldsymbol{\Lambda}}$ are the estimators of the vectors $\boldsymbol{c}_l$ and the matrix $\boldsymbol{\Lambda}$, $l = 1, 2, \ldots, n$ respectively.

We may summarize the MFGCCA algorithm in the following way.

MFGCCA algorithm

- Given: data set $\{(t_j, \boldsymbol{x}_{kjl}) : \ k = 1, \ldots, K, \ j = 1, \ldots, J, \ l = 1, \ldots, n\}$, where $\boldsymbol{x}_{kjl} \in \mathbb{R}^{p_k}$.
- Compute the estimators $\hat{\boldsymbol{c}}_{kl}$, $k = 1, \ldots, K$, $l = 1, \ldots, n$, using the least square method to get the functional data $\boldsymbol{x}_{1l}(t), \ldots, \boldsymbol{x}_{Kl}(t)$, $l = 1, \ldots, n$.
- Choose the penalty parameter $\lambda > 0$.
- Solve the eigenproblem with the matrices $\boldsymbol{\Lambda}$, $\boldsymbol{D}$ and $\boldsymbol{R}$.
- Project the functional data onto the space spanned by the superblock components.

6. Illustrative example

To present the described methodology in practice we use agriculture data for Polish regions available at the Statistics Poland website (http://stat.gov.pl). We have crop yields (in quintals per hectare) from 2003–2016 ($J = 14$ years and $n = 16$ regions). The data set ($p = 30$ variables in total) is split (by the Polish government) into $K = 3$ blocks:

- Block 1 ($p_1 = 9$ variables): wheat, rye, barley, oat, triticale, buckwheat, millet, potatoes and sugar beet.

- Block 2 ($p_2 = 6$ variables): legume fodder, clover, lucerne, serradella, field crops, root fodder.
- Block 3 ($p_3 = 15$ variables): cabbage, cauliflower, onion, carrot, cucumbers, tomatoes, apples, pears, plums, cherries, sweet cherries, strawberries, raspberries, currants, gooseberry.

In the first step we transformed discrete data into functional data. During the smoothing process we used a Fourier basis with 9 ($B_l = 8,\ l = 1, 2, \ldots, 30$) components. In the next step we applied the method described earlier. A graphical display of the regions is presented in Figure 1. We projected the data onto the space spanned by the first two superblock components. The space spanned by the global components is viewed as a compromise space that integrates all the modalities and facilitates the visualization of the results and their interpretation. In Figure 1 we can see that regions from the same macroregion are quite close to each other. This is reasonable, because regions from the same macroregion have similar soil, temperature and precipitation conditions. The only exception is the Opole region, which is completely different from the other regions. This may be expected, as this region is the best for agriculture in Poland. The climate in the region is characterized by a warm summer, mild and short winter, early spring and long mild autumn. 62% of the province's area consists of fertile brown and clay soils and fluvisols. The high quality of soils, lowland terrain and mild climate are conducive to the development of agriculture.

Calculations were performed using R 3.6.1 (R Core Team 2019), with the RGCCA (Tenenhaus and Guillemot 2017a) and fda (Ramsay et al. 2018) packages.

7. Concluding remarks

In this paper we have presented a technique for analyzing a multivariate functional multiblock data set. We propose an extension of functional canonical correlation analysis to the analysis of more than two sets of multivariate functional data. MFGCCA is a very attractive method for the analysis of such data sets. The proposed method has proved useful in investigating Polish agricultural regions.

Several points have been passed over in this paper but will be investigated in future research:

- GCCA can be viewed as a special case of the more general Regularized

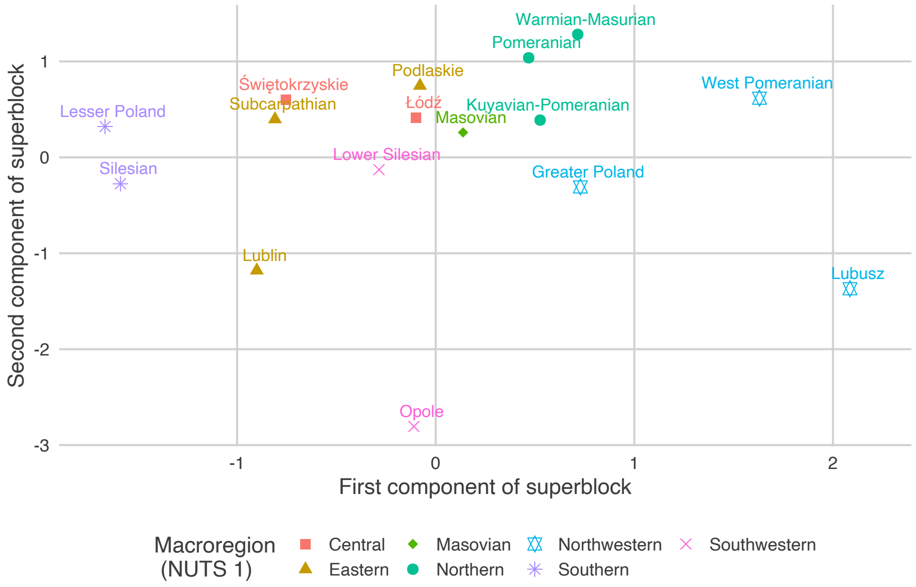

Figure 1. Graphical display of the regions obtained by crossing the first two components of the superblock

Generalized Canonical Correlation Analysis (RGCCA), which is a generalization of regularized canonical correlation analysis to three or more sets of variables (Tenenhaus and Tenenhaus 2011). It ought to be possible to translate the whole RGCCA framework to functional data.

- GCCA captures only linear relations between blocks of variables. To assess nonlinear relations, a kernel extension of GCCA (Tenenhaus et al. 2015) for functional data should be developed.
- Sparse GCCA (Tenenhaus et al. 2014 and Löfstedt et al. 2018) was recently proposed to address the issue of variable selection. It seems that it should be possible to utilize this idea in the case of functional data.

All of these possibilities warrant future theoretical and empirical work.

References

Carroll, J.D. (1968a): Generalization of canonical correlation analysis to three or more sets of variables. Proceedings of the 76th Annual Convention of the American Psychological Association 3:227–228.

Carroll, J.D. (1968b): Equations and tables for a generalization of canonical correlation analysis to three or more sets of variables. Unpublished companion paper to Carroll (1968a).

Gower, J.C. (1989): Generalized canonical analysis. R. Coppi and S. Bolasco, Editors, Multiway Data Analysis. North Holland, Amsterdam, 221–232.

Górecki, T., Krzyśko, M., Wołyński, W. (2017): Correlation analysis for multivariate functional data. Data Science, Studies in Classification, Data Analysis, and Knowledge Organization 243–258.

Górecki, T., Krzyśko, M., Waszak, Ł., Wołyński, W. (2018): Selected statistical methods of data analysis for multivariate functional data. Statistical Papers 59(1): 153–182.

Hotelling, H. (1936): Relations between two sets of variates. Biometrika 28(3/4): 321–377.

Horváth, L., Kokoszka, P. (2012): Inference for Functional Data with Applications. Springer. New York.

Hwang, H., Jung, K., Takane, Y., Woodward, T.S. (2012): Functional multiple-set canonical correlation analysis. Psychometrika 77(1): 48–64.

Hwang, H., Jung, K., Takane, Y., Woodward, T.S. (2013): A unified approach to multiple-set canonical correlation analysis and principal components analysis. British Journal of Mathematical and Statistical Psychology 66: 308–321.

Kettenring, J.R. (1971): Canonical analysis of several sets of variables. Biometrika 58(3): 433–451.

Leurgans, S.E., Moyeed, R.A., Silverman, B.W. (1993): Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society. Series B 55(3): 725—740.

Löfstedt, T. Hadj-Selem, F., Guillemot, V., Philippe, C., Raymond, N., Duchesney, E., Frouin, V., Tenenhaus, A. (2018): A general multiblock method for structured variable selection. arXiv:1610.09490v1 [stat.ML]

Markos, A., D'enza, A.I. (2016): Incremental generalized canonical correlation analysis. Analysis of Large and Complex Data, Studies in Classification, Data Analysis, and Knowledge Organization: 185–194.

Ramsay, J.O., Silverman, B.W. (2005): Functional Data Analysis, 2nd edition. Springer, New York.

Ramsay, J.O. Wickham, H., Graves, S., Hooker, G. (2018): fda: Functional Data Analysis. R package version 2.4.8. `https://CRAN.R-project.org/package=fda`

R Core Team (2019): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`

Takane, Y., Hwang, H., Abdi, H. (2008): Regularized multiple-set canonical correlation analysis. Psychometrika 73(4): 753–775.

Takane, Y., Oshima-Takane, Y. (2002): Nonlinear generalized canonical correlation analysis by neural network models. Measurement and Multivariate Analysis: 183–190.

Tenenhaus, A., Guillemot, V. (2017a): RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis for Multiblock Data. R package version 2.1.2. `https://CRAN.R-project.org/package=RGCCA`

Tenenhaus, A., Philippe, C., Frouin, V. (2015): Kernel generalized canonical correlation analysis. Computational Statistics & Data Analysis 90(C): 114–131.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.A., Grill, J., Frouin, V. (2014): Variable selection for generalized canonical correlation analysis. Biostatistics 15(3): 569–83.

Tenenhaus, A., Tenenhaus, M. (2011): Regularized generalized canonical correlation analysis. Psychometrika 76(2): 257–284.

Tenenhaus, M., Tenenhaus, A., Groenen, P. (2017b): Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. Psychometrika 82(3): 737–777.

Van de Velden. M. (2011): On generalized canonical correlation analysis. Proc. 58th World Statistical Congress. Dublin, 758–765.