Regularized generalized canonical correlation analysis for functional data

Tomasz Górecki, Mirosław Krzyśko, Waldemar Wołyński

Department of Mathematics and Computer Science
Adam Mickiewicz University, Poznań, Poland

XLIV Konferencja Statystyka Matematyczna
Będlewo 2-7.12.2018

In recent years methods for data representing functions or curves have received much attention. Such data are known in the literature as the functional data (Ramsay & Silverman, 2005). Examples of functional data can be found in several application domains, such as medicine, economics, meteorology and many others. In many applications there is need for using statistical methods for objects characterized by many features observed in many time points. Such data are called the multivariate functional data.

In this presentation we focused at relations between multiple sets of variables of multivariate functional data.

Let us assume that $\boldsymbol{X} \in L_2^p(I)$ is a random process, where $L_2(I)$ is a Hilbert space of square integrable functions on the interval I.
Additionally, we also assume that

$$\mathrm{E}(\boldsymbol{X}(t)) = \mathbf{0}, \ \ t \in I.$$

We will further assume that each component X_g of the process $\boldsymbol{X}$ can be represented by a finite number of basis functions $\{\varphi_e\}$:

$$X_g(t) = \sum_{e=0}^{B_g} \alpha_{ge}\varphi_e(t), s \in I, g = 1, 2, ..., p.$$

The degree of smoothness of function X_g depends on the value B_g (a small values cause more smoothing of the functions).

We introduce the following notation:

$$\boldsymbol{\alpha} = (\alpha_{10}, ..., \alpha_{1B_1}, ..., \alpha_{p0}, ..., \alpha_{pB_p})^\top,$$

$$\boldsymbol{\Phi}(s) = \begin{bmatrix} \boldsymbol{\varphi}_{B_1}^\top(t) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\varphi}_{B_2}^\top(t) & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\varphi}_{B_p}^\top(t) \end{bmatrix},$$

where $\boldsymbol{\varphi}_{B_1}, ..., \boldsymbol{\varphi}_{B_p}$ are orthonormal basis functions of space $L_2(I)$.

Using the above matrix notation the random process $\boldsymbol{X}$ can be represented as

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\alpha}. \tag{1}$$

This means that the realizations of process $\boldsymbol{X}$ are in finite dimensional subspace $\mathcal{L}_2^p(I)$ of $L_2^p(I)$.

We can estimate the vector $\boldsymbol{\alpha}$ on the basis of n independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the random process $\boldsymbol{X}$ (functional data) using eg. maximum likelihood method.

Details of the process of transformation of discrete data to functional data can be found eg. in Ramsay and Silverman (2005).

Canonical correlation analysis (Hotelling, 1936) is the study of the linear relations between two sets of variables. Let $\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1p})^\top$ and $\boldsymbol{X}_2 = (X_{21}, \ldots, X_{2q})^\top$ denote 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 $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$.

Let $\boldsymbol{X}^\top = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)$ has the covariance matrix of the form

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

The first pair of canonical variables (U_{11}, U_{21}) is defined via the pair of linear combinations of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$:

$$U_{11} = \boldsymbol{l}_{11}^\top \boldsymbol{X}_1,\ \ U_{21} = \boldsymbol{l}_{21}^\top \boldsymbol{X}_2$$

that maximize the correlation between U_1 and U_2, i.e. maximize

$$\mathrm{Corr}(U_1, U_2) = \mathrm{Corr}(\boldsymbol{l}_1^\top \boldsymbol{X}_1, \boldsymbol{l}_2^\top \boldsymbol{X}_2) \tag{2}$$

subject to U_1 and U_2 having unit variances.

Remaining canonical variables (U_{1j}, U_{2j}) maximize (2) subject to having unit variances and being uncorrelated with (U_{1k}, U_{2k}), $k < j$.

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

$$\mathrm{Var}(U) = \boldsymbol{l}^\top \boldsymbol{\Sigma} \boldsymbol{l} = \boldsymbol{l}_1^\top \boldsymbol{\Sigma}_{11} \boldsymbol{l}_1 + \boldsymbol{l}_2^\top \boldsymbol{\Sigma}_{22} \boldsymbol{l}_2 + 2\boldsymbol{l}_1^\top \boldsymbol{\Sigma}_{12} \boldsymbol{l}_2, \qquad (3)$$

and the problem of maximizing the expression (2) is equivalent to the problem of maximizing (3) subject to $\mathrm{Var}(U_1) = \boldsymbol{l}_1^\top \boldsymbol{\Sigma}_{11} \boldsymbol{l}_1 = 1$, and $\mathrm{Var}(U_2) = \boldsymbol{l}_2^\top \boldsymbol{\Sigma}_{22} \boldsymbol{l}_2 = 1$.

Now, we consider the generalized version of canonical correlation analysis (Carroll, 1968), that allows to analyze several sets of variables simultaneously.

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip_i})^\top$ denote random vectors with zero mean vector and covariance matrices $\boldsymbol{\Sigma}_{ii}$, $i = 1, \ldots, K$. Moreover, let $\boldsymbol{X}^\top = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_K^\top)$, 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 for K canonical variables $U_{11}, \ldots, U_{K1}$ being the linear combination of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ respectively, that maximize the sum of correlations between each pair of canonical variables, and having unit variances.

Denote $U_i = \boldsymbol{l}_i^\top \boldsymbol{X}_i$, $i = 1, \ldots, K$, $\boldsymbol{l}^\top = (\boldsymbol{l}_l, \ldots, \boldsymbol{l}_K)$, and $U = \boldsymbol{l}^\top \boldsymbol{X}$. Then the main problem of generalized canonical correlation analysis may be formulated as maximize $\mathrm{Var}(U)$ subject to $\mathrm{Var}(U_i) = 1$, $i = 1, \ldots, K$. Note that the problem of maximizing $\mathrm{Var}(U)$ is equivalent to the problem of maximizing

$$\sum_{i,j=1,i<j}^{K} \mathrm{Cov}(U_i, U_j) = \sum_{i,j=1,i<j}^{K} \boldsymbol{l}_i^\top \boldsymbol{\Sigma}_{ij} \boldsymbol{l}_j$$

subject to

$$\mathrm{Var}(U_i) = \boldsymbol{l}_i^\top \boldsymbol{\Sigma}_{ii} \boldsymbol{l}_i = 1, \; i = 1, \ldots, K.$$

In the case of random processes, we define the K canonical variables $U_1, \ldots, U_K$ as a dot product, i.e.

$$U_i =< \boldsymbol{l}_i, \boldsymbol{X}_i >= \int_I \boldsymbol{l}_i^\top(t)\boldsymbol{X}_i(t)dt,$$

where $\boldsymbol{l}_i \in \mathcal{L}_2^p(I)$, $i = 1, \ldots, K$.
In this case, we may assume (Ramsay and Silverman (2005)) that the vector weight function $\boldsymbol{l}_i$ and the process $\boldsymbol{X}_i$ are in the same space, i.e. the function $\boldsymbol{l}_i$ can be written in the form

$$\boldsymbol{l}_i(t) = \boldsymbol{\Phi}_i(t)\boldsymbol{\lambda}_i, \tag{4}$$

where $\boldsymbol{\lambda}_i \in \mathbb{R}^{B_{i1}+\ldots+B_{ip_i}}$.

Hence

$$U_i = < \boldsymbol{l}_i, \boldsymbol{X}_i > = \boldsymbol{\lambda}_i^\top [\int_I \boldsymbol{\Phi}^\top(t)\boldsymbol{\Phi}(t)dt]\boldsymbol{\alpha}_i = \boldsymbol{\lambda}_i^\top \boldsymbol{\alpha}_i,$$

where $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_i$ are vectors occurring in the representations (1) and (4) of process $\boldsymbol{X}_i$ and function $\boldsymbol{l}_i$, $i = 1, \ldots, K$.

So our problem may be reduced to the problem involving only random vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_i$.

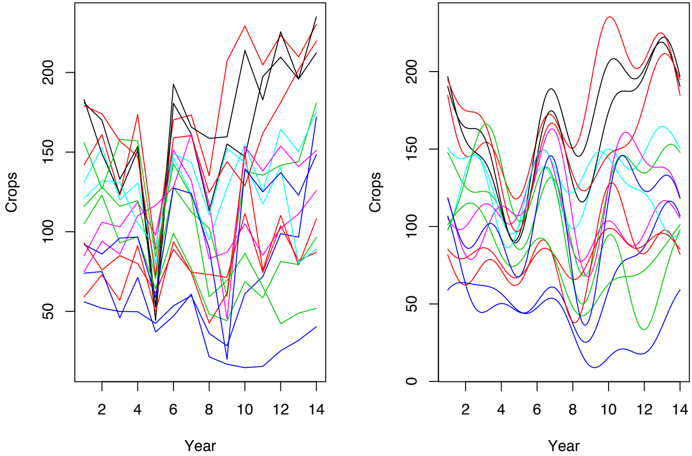
Example – Polish regions

As a real example we used agriculture data about Polish regions available at Central Statistical Office (Poland) website (http://stat.gov.pl/). We have crops (in quintals per hectare) from 2003-2016 (14 years and 16 voivodeships). Data set (in total 30 variables) is split into three natural blocks:

- Section 1 (9 variables): wheat, rye, barley, oat, triticale, buckwheat, millet, potatoes and sugar beet.
- Section 2 (6 variables): legume fodder, clover, lucerne, serradella, field crops, root fodder.
- Section 3 (15 variables): cabbage, cauliflower, onion, carrot, cucumbers, tomatoes, apples, pears, plums, cherries, sweet cherries, strawberries, raspberries, currants, gooseberry.

Example – Polish regions

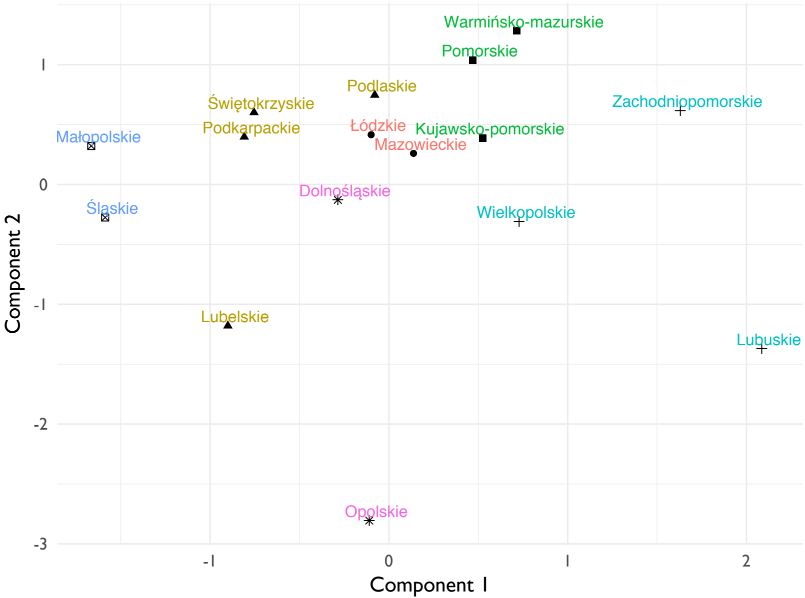
During the smoothing process we used Fourier basis with 9 components (eg. apples – discrete data on the left and smoothed data on the right).



In the next step we applied described earlier method.

We used packages `RGCCA` and `fda` from `R` free software environment.

Example – Polish regions



- CARROLL, J.D. (1968). 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.
- HOTELLING, H. (1936). Relations between two sets of variates. Biometrika 28:321–377.
- HWANG, H., JUNG, K., TAKANE, Y. (2011). Functional multiple-set canonical correlation analysis. Psychometrika 73(4):753–775.
- MARKOS, A., D'ENZA, A.I. (2016). Incremental generalized canonical correlation analysis. In: *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, Second Edition, Springer.
- TENENHAUS, A., GUILLEMOT, V. (2017). RGCCA: Regularized and sparse generalized canonical correlation analysis for multiblock data. `http://cran.project.org/web/packages/RGCCA/index.html`.
- TENENHAUS, A., TENENHAUS, M. (2011). Regularized generalized canonical correlation analysis. Psychometrika 76:257–284.
- TENENHAUS M., TENENHAUS A., GROENEN P. (2017), Regularized generalized canonical correlation analysis: A framework for sequential multiblock component methods. Psychometrika 82(3):737–777.