APPLICATION OF TWO-SET MULTIVARIATE STATISTICAL METHODS TO THE CZECH REPUBLIC ARRIVAL TOURISM DATA

Lukáš Malec - Miloslav Malec

Abstract

In this study are described the principles of canonical correlation analysis as classical multivariate approach and partial least squares (also called robust canonical analysis or canonical covariance) methods in the two-set case. The possibilities of their numerical realization, based on spectral, resp. singular decomposition theories are studied together with a functional relation between canonical correlation analysis and partial least squares. Based on matrix decomposition theory we can conclude to paths of eigenvalues and eigenvectors, resp. singular values and singular vectors which describe the relation between both approaches. This also opens the possibility of regularization of canonical correlation analysis if input matrices are singular. The application part is concentrated on arrival tourism data of non-residents and residents to the Czech regions. The paths of selected parameters are expressed together with first-order eigenvectors and their interpretation. A brief comparison of canonical correlation and partial least squares methods from the point of view of applicability in specific situations is also included.

Key words: multivariate analysis, regularization, generalized eigenproblem, tourism

JEL Code: C33, L83

Introduction

Multivariate statistical techniques of canonical correlation analysis (CCA) and partial least squares (PLS) belong to the group of commonly applied methods and deal with processing, resp. assessing data matrices from various research disciplines such as pattern recognition, psychometry, biometry and chemical engineering. Note that, PLS methodologies are still rarely applied in econometry. CCA and partial least squares methods are up to now evolved and generalized and the number of original works published in a high-quality proceedings and journals proves their importance. The programs are published along with the theoretical background for numerical realization of such methods.

The first reference concerning CCA originates in a 1936 seminar paper by H. Hotelling (Hotelling, 1936), long before H. Wold published the basic concept of PLS in 1975 (Wold, 1975). The PLS method (described below) is one of the ranges of approaches to PLS often called robust canonical analysis, or canonical covariance, closely related to canonical correlation. The basic difference in interpretation purposes of both methods is that CCA aims to find correlations between sets while partial least squares methodology deals with covariance structures. Generally, a significant part of the PLS methods use iteration approaches for solving optimization tasks, including non-linear ones.

Contrary to partial least squares, the numerical realization of CCA is often problematic and unstable (Ewerbring & Luk, 1989; Yamamoto et al., 2008; Hardoon, Szedmak & Shawe-Taylor, 2004). The case of singularity or near-singularity of data caused by both small-sample problem and/or collinearity within individual data sets (statistically, the case of violation of input assumptions) are the frequent reasons of disrupting numerical computations of canonical correlation analysis. To overcome this problem many solutions were proposed of which the most used is regularized canonical analysis (rCCA) (Vinod, 1976). Because the manually predefined change of computational matrices (by regularization) can significantly decrease the stability of the classical CCA algorithm and also due to statistically difficult interpretation of such results, PLS algorithm is set as a competing approach by some authors (Wegelin, 2000; Yamamoto et al., 2008).

The methods in this study are applied to the Czech Republic arrival tourism data. We divide visitors of individual regions to non-residents and residents and study the basic relations between both data sets. The main idea is to reveal similarity of profiles (after standardizing the data) between arrivals on the fundamental year scale.

The data presented are complicated in the nature. They suffer from the case of small-sample problem and also from collinearity within sets. Our attention is concentrated on the PLS approach, which has been negligible adopted in the tourism literature with seldom publications. The PLS method is fully the method of choice in this study. Only limited notes dealing with matrix analysis theory to formulate the relationship between CCA and partial least squares are still available whose basics are noted here. In the applications, we use significant eigenvalues of both methods, but we explain only first eigenvectors. According to (Wegelin, 2000), many PLS approaches give the same results for first-order analyses.

Chapters 1 and 2 consist of mathematical descriptions of CCA, rCCA and PLS methods and their relationships using paths of eigenvalues and eigenvectors in dependency on

parameter. Chapters 3 and 4 describe the experiment with arrival tourism data divided into non-residents and residents. Concluding remarks are also considered.

1 Algebraic approach to CCA and PLS

Let us introduce two random vectors $(x_1, x_2, ..., x_r)$, $(y_1, y_2, ..., y_s)$. The task of CCA method is finding a linear combination of the elements of this vectors in such a way that resulting random variables (latent variables) reach a maximum correlation (Krzanowski, 2000). The random sample of observations x_i, y_i of range n is usually available, which we consider here as standardized. We mark corresponding matrices as X of type (n, r) and Y of type (n, s), columns are variables, rows correspond to observations.

Algebraically, the CCA method wishes to finds vectors $\boldsymbol{u}_x \in R^r$, $\boldsymbol{u}_y \in R^s$ maximizing

$$\frac{\boldsymbol{u}_x' X'Y \boldsymbol{u}_y}{(\boldsymbol{u}_x' X'X \boldsymbol{u}_x)^{\frac{1}{2}} (\boldsymbol{u}_y' Y'Y \boldsymbol{u}_y)^{\frac{1}{2}}}. \tag{1}$$

Lagrange multipliers method guarantees the existence of numbers $\lambda_1 = \lambda_2 = \lambda$ such that stationary solution of (1) gives the system

$$\begin{aligned} X'Y\boldsymbol{u}_y &= \lambda X'X\boldsymbol{u}_x \\ Y'X\boldsymbol{u}_x &= \lambda Y'Y\boldsymbol{u}_y. \end{aligned} \tag{2}$$

Analogically to CCA, at PLS method we find vectors $\boldsymbol{u}_x \in R^r$, $\boldsymbol{u}_y \in R^s$ maximizing

$$\frac{\boldsymbol{u}_x' X'Y \boldsymbol{u}_y}{(\boldsymbol{u}_x' \boldsymbol{u}_x)^{\frac{1}{2}} (\boldsymbol{u}_y' \boldsymbol{u}_y)^{\frac{1}{2}}}. \tag{3}$$

Stationary solution is given by the system

$$\begin{aligned} X'Y\boldsymbol{u}_y &= \lambda \boldsymbol{u}_x \\ Y'X\boldsymbol{u}_x &= \lambda \boldsymbol{u}_y. \end{aligned} \tag{4}$$

We can rewrite the system (2) by matrix representation

$$\begin{pmatrix} 0 & X'Y \\ Y'X & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_x \\ \boldsymbol{u}_y \end{pmatrix} = \lambda \begin{pmatrix} X'X & 0 \\ 0 & Y'Y \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_x \\ \boldsymbol{u}_y \end{pmatrix}. \tag{5}$$

This is the generalized task of eigenvalues of a symmetric matrix. In regular case, we solve a generalized eigenproblem using its inversion transform to the standard eigenproblem and then using e.g. Cholesky decomposition to a standard task with the symmetric matrix (see below).

The problem arises when this matrix is singular, viz (Golub & Van Loan, 1996; sub-chapt. 7.7). Such a situation can be solved by regularization.

The system (4) can be rewritten as

$$\begin{pmatrix} 0 & X'Y \\ Y'X & 0 \end{pmatrix}\begin{pmatrix} \boldsymbol{u}_x \\ \boldsymbol{u}_y \end{pmatrix} = \lambda\begin{pmatrix} \boldsymbol{u}_x \\ \boldsymbol{u}_y \end{pmatrix}. \qquad (6)$$

The entries (4), resp. (6) are the problems for finding singular values and singular vectors of matrix $X'Y$, viz (Harville, 1997; Golub & Van Loan, 1996).[1]

Matrices $X'X$, $Y'Y$ are positive definite (thus regular) if $n \geq \max(r, s)$, $h(X) = r$, $h(Y) = s$. Moreover in such a case $h(X'Y) = \min(r, s) = h$. The task (1) maximizes a quantity $\boldsymbol{u}_x' X'Y \boldsymbol{u}_y$ on ellipsoids, task (3) on unit spheres in R^r, resp. R^s given by a subset of optimization problems (1) and (3). System (2) can also be transformed to an eigenproblem of symmetrical matrices using Cholesky decomposition (Hardoon, Szedmak & Shawe-Taylor, 2004) or using the square root of matrix (Ewerbring & Luk, 1989).

Let $X'X = T_x T_x'$, $Y'Y = T_y T_y'$ where T_y, T_y are lower triangular matrices with positive diagonal elements; using substitution $T_x' \boldsymbol{u}_x = \boldsymbol{v}_x$, $T_y' \boldsymbol{u}_y = \boldsymbol{v}_y$ we can rewrite (2) as

$$\begin{aligned} &\underbrace{T_x^{-1} X'Y T_y^{-1}}_{A}{}' \underbrace{T_y^{-1} Y'X T_x^{-1}}_{A'}{}' \boldsymbol{v}_x = \lambda^2 \boldsymbol{v}_x \\ &\underbrace{T_y^{-1} Y'X T_x^{-1}}_{A'}{}' \underbrace{T_x^{-1} X'Y T_y^{-1}}_{A}{}' \boldsymbol{v}_y = \lambda^2 \boldsymbol{v}_y. \end{aligned} \qquad (7)$$

This is the system for finding eigenvalues of symmetrical matrices. The problem (7) has h nonzero eigenvalues for which $1 \geq \lambda_1^2 \geq \lambda_2^2 \geq \ldots \geq \lambda_h^2 > 0$. Numbers $\lambda_i > 0$ correspond to eigenvalues of (2) which are realized in the given eigenvectors. System (7) can also be solved by singular decomposition of A, viz (Ewerbring & Luk, 1989; Golub & Van Loan, 1996).

The case of singular or near-singular matrices $X'X$, $Y'Y$ in CCA is operated by various ways. One basic step is the regularization approach where $X'X, Y'Y$ matrices in the denominator of (1) are removed by $X'X + \delta I_x$, $Y'Y + \delta I_y$, where I_x, I_y are identity matrices and δ is a small nonzero parameter. In this study, the other solution is used, i.e. the convex combination as $(1-\delta)X'X + \delta I_x$, $(1-\delta)Y'Y + \delta I_y$. The notation of regularized task and its solution is realized in accordance with the original solution.

[1] The results of matrix theory are presented in (Harville, 1997), the solution of its numerical realization in (Golub & Van Loan, 1996).

2 Functional relation of CCA and PLS

Let $A = \begin{pmatrix} 0 & X'Y \\ Y'X & 0 \end{pmatrix}$, $B_1(\delta) = (1-\delta)X'X + \delta I_x$, $B_2(\delta) = (1-\delta)Y'Y + \delta I_y$. Then

$$B(\delta) = \begin{pmatrix} B_1(\delta) & 0 \\ 0 & B_2(\delta) \end{pmatrix} \tag{8}$$

where $\delta \in \langle 0, 1 \rangle$ is the parameterization of the abscissa (in the dimension of matrices) with the boundary points A and $I = \begin{pmatrix} I_x & 0 \\ 0 & I_y \end{pmatrix}$. Stationary solution of maximization task of type (1) in dependency on parameter δ is given by the system

$$A\begin{pmatrix} \boldsymbol{u}_x(\delta) \\ \boldsymbol{u}_y(\delta) \end{pmatrix} = \lambda(\delta)B(\delta)\begin{pmatrix} \boldsymbol{u}_x(\delta) \\ \boldsymbol{u}_y(\delta) \end{pmatrix}. \tag{9}$$

If matrix $B(0)$ is positive definite then $B(\delta)$ is positive definite at $\langle 0, 1 \rangle$. This task for $\delta = 0$ gives CCA, for $\delta = 1$ it gives the partial least squares.

The description of functional relation between CCA and PLS means to assign spaces of continuous, smooth, resp. analytical functions in whose lie paths of vector functions $\lambda(\delta)$, $\boldsymbol{u}_x(\delta)$, $\boldsymbol{u}_y(\delta)$. Those paths are solutions of system (9).

In this study we describe only smooth paths of system (9). For this reason we rewrite (9) into the form by using symmetrical matrices

$$\begin{aligned} &T_x(\delta)^{-1}X'YT_y(\delta)^{-1}{}'\,T_y(\delta)^{-1}Y'XT_x(\delta)^{-1}{}'\,\boldsymbol{v}_x(\delta) = \lambda^2\boldsymbol{v}_x(\delta) \\ &T_y(\delta)^{-1}Y'XT_x(\delta)^{-1}{}'\,T_x(\delta)^{-1}X'YT_y(\delta)^{-1}{}'\,\boldsymbol{v}_y(\delta) = \lambda^2\boldsymbol{v}_y(\delta). \end{aligned} \tag{10}$$

Due to (Harville, 1997; theorems of sub-chapt. 15.8), the elements of matrices in (10) are smooth functions at $\langle 0, 1 \rangle$. According to (Harville, 1997; sub-chapt. 21.15) the path

$$\left(\lambda_i(\delta), \boldsymbol{u}_{jx}(\delta), \boldsymbol{u}_{jy}(\delta)\right) \tag{11}$$

of system (10) is the smooth function at $\langle 0, 1 \rangle$ if eigenvalue $\lambda_i(\delta)$ is simple at $\langle 0, 1 \rangle$. The existence of analytic paths does not require the existence of simple eigenvalues, viz (Bunse-Gerstner at al., 1991).

In the case of a small-sample problem, and in the case of strong collinearity, the matrices $X'X$, resp. $Y'Y$ are singular or near-singular. This situation is solvable by rCCA.

We choose small $\delta > 0$. The result of (10) for $\delta = \delta_0$ removes the solution for $\delta = 0$ which is not unique. The functional relation is studied at $\langle \delta_0, 1 \rangle$. The structure of the spectrum given by (5) in the case when $X'X$, resp. $Y'Y$ are singular is described, e.g. in (Golub & Van Loan, 1996).

3 Experimental

3.1 Database considered

In this study, we used arrival tourism data into 14 Czech regions.[2] The main purpose is to find a similarity in the annual profiles between arrivals of non-residents and residents during the years 2003 – 2011, and to reveal the basic features of the functional relation of CCA and partial least squares method. We use the Czech statistical office database (CSO database, 2013). Those statistics are investigated at all collective tourist accommodation establishments using a questionnaire survey. An effective Directive 95/57/EC of 23 November 1995 on the collection of statistical information in the field of tourism (Council Directive, 1995) is valid within the EU.

3.2 Methodology

We process the tourism data using the multivariate canonical correlation and PLS methods. Because the data sets are singular (suffering from a small-sample problem; $n = 9, r = s = 14$) and also collinear, the regularized approach (with the parameter $\delta = 0.01$) to canonical correlation analysis is used. The functional relation is expressed through the paths of both methods depending on the parameter. Only the first eigenvectors are briefly discussed, although the higher-order eigenvectors can be interpreted in quite a similar way. First-order analyses are equal in many PLS approaches. We use MATLAB7 (Mathworks, Natick, MA, USA) environment and programs written solely by the authors of this study.

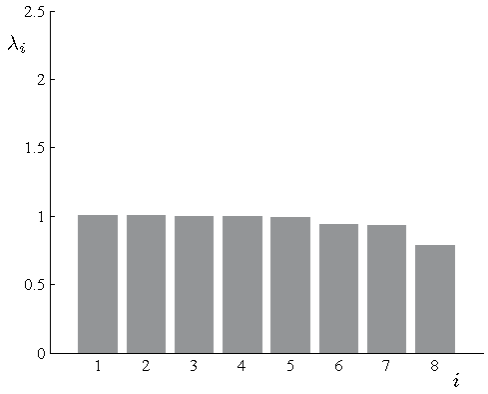
4 Results and discussion

By the preliminary inspection of data, Praha and Karlovarský regions are less influenced by economic crisis from the second half of 2008 through the end of 2009. Also, about 57% of non-residents visit only Prague, the Czech capital.

[2] The abbreviations are considered according to Czech classification, i.e. PHA – Praha, STC – Středočeský kraj, JHC – Jihočeský kraj, PLK – Plzeňský kraj, KVK – Karlovarský kraj, ULK – Ústecký kraj, LBK – Liberecký kraj, HKK – Královéhradecký kraj, PAK – Pardubický kraj, VYS – Vysočina, JHM – Jihomoravský kraj, OLK – Olomoucký kraj, ZLK – Zlínský kraj, MSK – Moravskoslezský kraj.

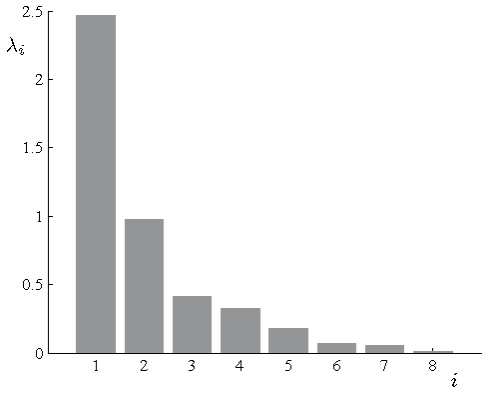
We mark arrivals of non-residents as *x*-set and of residents as *y*-set. As is mentioned earlier, the investigated data sets suffer from the small-sample problem and also from collinearity. Figure 1 shows the eigenvalues of term (9) for small δ which corresponds to rCCA. Figure 2 shows eigenvalues of (9) for $\delta = 1$ which corresponds to PLS. Figure 3 demonstrates the paths of eigenvalues at $\langle \delta_0, 1 \rangle$. The eigenvalues are of norm considering first step (rCCA).

Fig. 1: rCCA eigenvalues



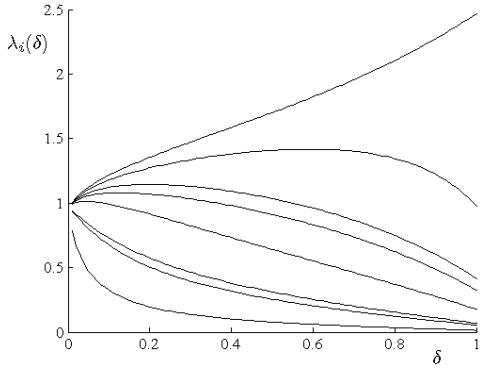
Source: author

Fig. 2: PLS eigenvalues



Source: author

Fig. 3: Path of eigenvalues



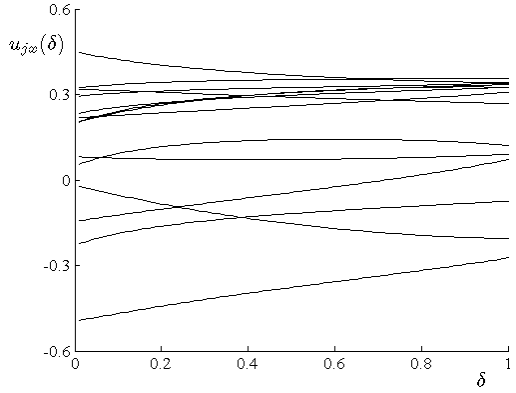
Source: author

Because of the high extent of collinearity within sets, $h < n$. The presence of collinearity is indicated by the values of determinants of within sets, both of orders 10^{-98}. If the interest is dimension reduction, as is frequently demanded in statistical methodologies, Figures 1, 2 and 3 indicate the more suitable alternative to choose δ close to 1, or just $\delta = 1$.

Figures 4 and 5 show processes of paths for individual eigenvectors expressed of unit norm. Because the eigenvalues are simple, the paths are smooth.

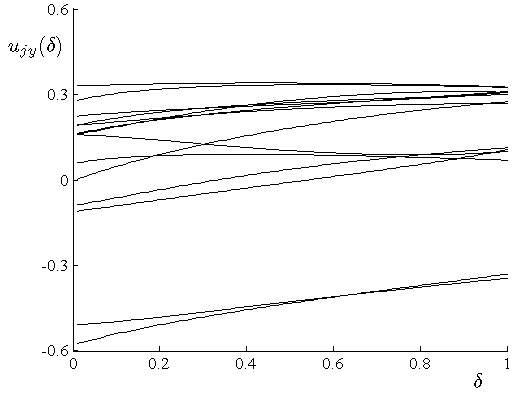
For the interpretation purposes we present first eigenvectors for both sets. In the similar way can be interpreted the remaining eigenvectors. The initial and final points of the paths in Figures 4 and 5 agree with values given in Table 1. The coefficients of both of analyses close to or exceeding 0.3 (in absolute values) are considered as significant.

Fig. 4: Eigenvectors path of *x*-set



Source: author

Fig. 5: Eigenvectors path of *y*-set



Source: author

Tab. 1: Coefficients of two-set methods

PHA	STC	JHC	PLK	KVK	ULK	LBK	HKK	PAK	VYS	JHM	OLK	ZLK	MSK
rCCA – non-residents													
-0.49	0.29	-0.14	0.08	-0.02	0.45	0.32	0.23	0.32	0.22	-0.22	0.20	0.20	0.05
residents													
-0.51	0.19	0.16	-0.09	-0.58	-0.11	0.00	0.33	0.06	0.19	0.16	0.16	0.28	0.22
PLS – non-residents													
-0.27	0.34	0.07	0.09	-0.21	0.34	0.36	0.33	0.27	0.31	-0.07	0.33	0.33	0.12
residents													
-0.34	0.27	0.31	0.11	-0.33	0.11	0.27	0.33	0.07	0.30	0.10	0.31	0.33	0.31

Source: author

We reveal basic relations (similarity of profiles) between both sets in the usual way dealing with magnitudes and signs of individual elements of eigenvectors.

For rCCA analysis we can see that only Praha region and Královéhradecký region (with a lower value of coefficient for non-residents and opposite profile to Prague) have significantly similar profiles for arrivals of non-residents and residents on first latent variables. On the other hand, Ústecký, Liberecký and Pardubický regions (with opposite profile to Prague) from non-residents, and Karlovarský region (with similar profile to Prague) from residents have dissimilar profiles to their counterparts from the second set.

The results of PLS method reveals that Praha, Středočeský, Karlovarský, Liberecký, Královéhradecký, Vysočina, Olomoucký and Zlínský regions show similar profiles for

arrivals of non-residents and residents, although some of regions prove opposite profiles to the others and also demonstrate lower values of coefficients. On the other hand, Ústecký and Pardubický regions from non-residents, and Jihočeský and Moravskoslezský regions from residents did not prove similar profiles to their counterparts from the second set.

Conclusion

Although both methods (rCCA and PLS) reveal similar results in first-order eigenvectors, the PLS method is preferred in the present case because of singularity of data (small-sample problem). Also the collinearity within sets unambiguously disrupts the assumptions for applying canonical correlation-based methodologies. The processes of individual eigenvalues and eigenvectors reveal their smooth paths which result is in accordance with the presented theory. We reveal many similar profiles in arrival tourism data divided into non-residents and residents mostly for the same regions of the Czech Republic. We can conclude similar behaviour of those two sets of tourists for Praha, Středočeský, Karlovarský, Liberecký, Královéhradecký, Vysočina, Olomoucký and Zlínský regions. Praha and Karlovarský regions prove different profiles from the rest, reading from the PLS analysis. The interpretation of eigenvalues and eigenvectors for dimension reduction is very useful in this particular tourism data processing approach.

Acknowledgment

The authors wish to acknowledge the financial support of the University of Business in Prague internal grant FRV No. 3/2013.

References

Bunse-Gerstner, A., Byers, R., Mehrmann, V., & Nichols, N. K. (1991). Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numerische Mathematik*, *60*, 1-39.

Council Directive 95/57/EC of 23 November 1995 on the collection of statistical information in the field of tourism.

CSO Database. (2013, 4 9). Retrieved from http://www.czso.cz/csu/redakce.nsf/i/cru40_cr.

Ewerbring, L. M., & Luk, F. T. (1989). Canonical correlation and generalized svd: Applications and new algorithms. *Journal of Computational and Applied Mathematics*, *27*, 37-52.

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations*. (3rd ed.). Baltimore: The Johns Hopkins University Press.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*, 2639-2664.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer-Verlag.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, *28*, 321-377.

Krzanowski, W. J. (2000). *Principles of multivariate analysis: A user's perspective*. (2nd ed.). Oxford: University Press.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, *4*, 147-166.

Wegelin, J. A. University of Washington, (2000). *A survey of partial least squares (pls) methods, with emphasis on two-block case*. Seattle: University Press.

Wold, H. (1975). Path models with latent variables: The nipals approach. In H. Blalock (Ed.), *Quantitative sociology: International perspectives on mathematical and statistical modelling* (pp. 307-357). New York: Academic Press.

Yamamoto, H., Yamaji, H., Fukusaki, E., Ohno, H., & Fukuda, H. (2008). Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting. *Biochemical Engineering Journal*, *40*, 199-204.

Contact

Lukáš Malec

Department of Mathematics and Statistics, University of Business in Prague

Spálená 76/14, 110 00, Prague 1, Czech Republic

malec@vso-praha.eu

Miloslav Malec

Department of Economy and Economics, Institute of Hospitality and Management in Prague

Svídnická 506, 181 00, Prague 8, Czech Republic

malec@vsh.cz