www.ejthr.com

European Journal of Tourism, Hospitality and Recreation
Vol. 4, Issue 3, pp. 181-203, 2013

Printed in Portugal

ON THE MULTIVARIATE PROCESSING OF RANK-DEFICIENT TOURISM DATA

Lukáš Malec
University of Business in Prague, Czech Republic

ABSTRACT: Classical canonical correlation analysis (CCA) and variants of partial least squares (PLS) are well-known methods catalogued under the denominator of multivariate statistics. However, the CCA method often suffers from the problem of a rank-deficient data and within sets relations. In this paper, we study the behaviour of both methods using tourism data based on arrivals. Such particular application gives PLS a clear superiority over CCA as over most other regression techniques due to the stability property. Another *ad hoc* multivariate technique, a principal component analysis (PCA), was accommodated to explain the relations of different variables as well as to study linear trends in data. **Keywords**: multivariate analysis, partial least squares, canonical analysis, rank-deficiency, tourism

RESUMEN: El análisis de correlación canónica clásica (CCA) y las variantes de los mínimos cuadrados parciales (PLS) son métodos bien conocidos de la estadística multivariada. Sin embargo, la existencia de valores omisos en el rango de datos y en los conjuntos de relaciones constituyen una limitación del método CCA. En este artículo, estudiamos el comportamiento de ambos los métodos recurriendo a bases de datos de Turismo. Esta aplicación en específico muestra una clara superioridad del método PLS en relación a la CCA así como la mayoría de las otras técnicas de regresión debido a la estabilidad de la propiedad. Una otra técnica multivariada *ad hoc*, el análisis de componentes principales (PCA), fue incluida para explicar las relaciones entre variables diferentes así como para estudiar tendencias lineares en los datos. **Palabras-clave**: análisis multivariada, mínimos cuadrados parciales, análisis correlacional canónica, *rank-deficiency*, turismo.

RESUMO: A análise correlacional canónica clássica (CCA) e as variantes dos mínimos quadrados parciais (PLS) são métodos bem conhecidos de estatística multivariada. No entanto, a existência de valores omissos nas séries de dados e nos conjuntos de relações constitui uma limitação do método CCA. Neste artigo, estudamos o comportamento de ambos os métodos recorrendo a bases de dados de Turismo. Esta aplicação em específico mostra uma clara superioridade do método PLS em relação à CCA tal como a maioria das outras técnicas de regressão devido à propriedade de estabilidade. Uma outra técnica multivariada *ad hoc*, a análise de componentes principais (PCA), foi incluída para explicar as relações entre variáveis diferentes, assim como para estudar tendências lineares nos dados. **Palavras-chave**: análise multivariada, mínimos quadrados parciais, análise correlacional canónica, omissão de dados, turismo.

Lukáš Malec is the deputy head of Department of Information Technologies and Analytical Methods at University of Business in Prague. His area of research is concentrated on application of multivariate methods in economy. Author's email: Lukas.Malec@vso-praha.eu.

INTRODUCTION

To evaluate the contribution of the tourism sector in economy and other related branches, the techniques of statistical data processing are required in a high quality. PLS, CCA and the principal component analysis are the dimensionality reduction multivariate methods, similar to factor analysis and linear discriminant analysis, which are highly applicable in many research areas. Such techniques are seldom used in tourism research (Vajčerová *et al.*, 2012; Iatu and Bulai, 2011; Constantin and Grigorescu, 2009) as a great merit is given to more traditional techniques (see e.g. Bender *et al.*, 2005). Note that Herman Wold (1975) originally introduced PLS as an econometric method. However, from the present ranks of users, the majority are chemical engineers and chemometricians. One of the fundamental methods of multivariate analysis, PCA, was presented by Harold Hotelling (1933), and long before, also by Karl Pearson (1901). The generally accepted definition of classical canonical correlation analysis is attributed to Harold Hotelling (1935, 1936) in his works as a way of describing linear relationships between two sets of data.

The scope of CCA is unique, and ranges to define a new orthogonal coordinate system (for each of the two data sets) so that the new pair of coordinate systems is optimal in maximizing the correlations. Nevertheless, since Wold's original research approach, many variants of partial least squares technique have appeared in the literature. These variants exist depending on the way the original data sets are "deflated" (as PLS-W2A or PLS2). The Mode B PLS is equivalent to CCA. Although the CCA belongs to the class of PLS algorithms, for longer historical bases of canonical correlation analysis, the PLS is only assigned to Mode A class of algorithms. For a comprehensive survey of various methods, see Wegelin (2000). Note that all classical PLS variants are equivalent when just one pair of new coordinate system is computed. Among others, one important property of PLS coefficients is a measure of relative importance of individual variables in the model (Wold *et al.*, 2004).

We will discuss and apply only one variant of (two data sets) partial least squares algorithms in this study, abbreviated generally as PLS-SVD (Lorber *et al.*, 1987; Sampson *et al.*, 1989; Wegelin, 2000).[1] PLS-SVD has in recent times become a very attractive method, and is used for modelling as well as for prediction. Also some extension, e.g. sparse PLS was published (Cao *et al.*, 2008) using the Lasso penalty function. PLS-SVD, as a computationally simple Mode A algorithm, may however lead to a non-mutually orthogonal coordinate system which is one of the most significant disadvantages of this method.

Canonical correlation analysis has been widely used in fields such as dimension reduction and feature extraction. However, it is theoretically impossible to exactly compute the CCA parameters in cases when the number of variables (p and q, considering both sets) exceeds the number of observations $p > n$ and/or $q > n$, i.e. in the case of a rank-deficient data. This property is due to a computation of inverses on singular matrices and can be overcome by the l_2 regularization (penalization) technique (see below). Moreover, CCA is sensitive to collinearity or near collinearity within sets of data, which case (similarly to a variety of well-known regression methods) produces instability of algorithm and wrong interpretation of results. When testing the significance of canonical correlations, the multinormality of original variables is also an input assumption. In tourism, the sample size is often small, the number of variables mostly exceeding the number of observations, and the original variables are often collinear within sets. These computational problems of CCA are overcome by using partial least squares technique.

Note that although the classical techniques are the best performers for describing linear relationships, those techniques fail completely in capturing nonlinear types of relationships (Alam *et al.*, 2010). In order to solve this problem, several extensions of CCA have been recently proposed to account for nonlinear

[1] This method also has synonyms as robust canonical analysis, canonical covariance or intercorrelations analysis.

relationships between two sets of data. Among them, the application of kernel canonical correlation analysis (see e.g. Huang *et al.*, 2009) is one of the most promising approaches. Kernel CCA offers a canonical correlation solution by implicitly mapping the original data onto a high-dimensional feature space using some sort of kernel transformation.

Like PLS, the canonical correlation analysis can be applied to the same problem as mentioned in Wegelin (2000). This is the case in this study, where the behaviour of proposed methods is verified through application to rank-deficient tourism data in standardized form, and the effort is on mutually comparing those techniques as well as on discussing mutual linear relations and trends. The database consists of count original variables, i.e. the number of Czech visitors and total number of non-residents (total visitors) arrived in the selected European states. Because those data are time-series in nature, we will descriptively study the similarity courses in given periods. PCA method helps to reveal the relations within as well as between sets, and serves as a tool for explaining mutual inferences given by canonical correlation and partial least squares methods. Time variable is used to reveal the linear trends in the data.

In this study, the significant multivariate relations are rarely assigned to the arrivals of Czech visitors and non-residents covering the identical state. Despite this fact, the other very important courses of similarity are identified between both sets. Contrary to the total number of non-residents, where predominantly the directions of growth are identified, in the case of Czech visitors, the situation is more difficult and we reveal multiple directions.

The rest of article is organized as follows: In Algorithms section we briefly review some essentials of CCA, PLS and the principal component analysis methods together with some computational and practical problems. Experimental section covers information about the data sets and processing methods. In section Results and discussion we introduce the most important results from the range of tourism data, and Conclusion section consists of a summary of this application study.

ALGORITHMS

Notation. Let X and Y be sample matrices measured on n observations of types (n, p) and (n, p), respectively. Furthermore, no linear dependencies among individual rows $(n \leq p, n \leq q)$ and columns $(n \geq p, n \geq q)$ of X and Y are considered within sets, i.e. $rank\ (X) = \min\ (n, p)$ and $rank\ (Y) = \min\ (n, q)$. In the following, X and Y denote the matrices standardized by columns. The sample correlation matrices $X^T X$ and $Y^T Y$ are symmetrical and positive-definite. Vectors are boldface lowercase; scalars and variables are lowercase. The sets (groups) are denoted with uppercase.

Because PLS and CCA methods have rarely been applied in tourism and also because they can be described on the same base, we will discuss their algorithms. PCA, as the representative of one set technique, is discussed too. Although canonical correlation analysis is known to be derived using a probabilistic model (e.g. Leen and Fyfe, 2006; De Bie and De Moor, 2003; Bach and Jordan, 2005), up to the present, a complete probabilistic theory of various methods of partial least squares has not been published. The PCA is described, e.g. as the maximum likelihood solution of a factor analysis (e.g. Tipping and Bishop, 1999). Despite the undeniable advantages of probabilistic algorithms, we use a more illustrative algebraic approach in the following.

First the canonical correlation analysis and its regularized variant are introduced in this study. Then the computationally less expensive methods of PLS and principal component analysis are described.

Canonical correlation analysis

The classical canonical correlation technique is explained as finding vectors $(u, v) \in R^p \times R^q$ (coefficients of linear combinations), by solving problem

$$\max_{u,v \neq 0} \frac{\boldsymbol{u}^T X^T Y \boldsymbol{v}}{\sqrt{\boldsymbol{u}^T X^T X \boldsymbol{u}} \sqrt{\boldsymbol{v}^T Y^T Y \boldsymbol{v}}}. \tag{1}$$

Because (1) is a homogenous function in $\boldsymbol{u}$ and $\boldsymbol{v}$, this task is equivalent to finding local extrema of the constrained optimization problem (Krzanowski, 2000, p. 436; De Bie *et al.*, 2005)

$$\max_{\boldsymbol{u},\boldsymbol{v}\neq 0} \boldsymbol{u}^T X^T Y \boldsymbol{v} \quad \text{s.t. } \boldsymbol{u}^T X^T X \boldsymbol{u} = 1, \boldsymbol{v}^T Y^T Y \boldsymbol{v} = 1 \tag{2}$$

where the combinations $a = X\boldsymbol{u}$ and $b = Y\boldsymbol{v}$ are called the canonical variates (latent variables – LVs). Such variates can be used as a dimension reduction tool and as graphical display.

According to the Kuhn-Tucker theorem (Kuhn and Tucker, 1951), there are numbers λ_1 and λ_2 (Lagrange multipliers) in such a way that the solution is a stationary point of the corresponding Lagrangian

$$\boldsymbol{u}^T X^T Y \boldsymbol{v} - \frac{\lambda_1}{2}\left(\boldsymbol{u}^T X^T X \boldsymbol{u} - 1\right) - \frac{\lambda_2}{2}\left(\boldsymbol{v}^T Y^T Y \boldsymbol{v} - 1\right) \tag{3}$$

Solving derivatives with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$, the maximization leads to the system

$$X^T Y \boldsymbol{v} = \lambda_1 X^T X \boldsymbol{u} \tag{4}$$

$$Y^T X \boldsymbol{u} = \lambda_2 Y^T Y \boldsymbol{v}. \tag{5}$$

Since $u^T X^T Y v$ is a scalar, then according to expression (3), the following is valid: $\lambda_1 = \lambda_2 = \lambda = \boldsymbol{u}^T X^T Y \boldsymbol{v}$ (Hardoon *et al.*, 2004). From Eq. (5) we have

$$\boldsymbol{v} = \frac{1}{\lambda}\left(Y^T Y\right)^{-1} Y^T X \boldsymbol{u} \tag{6}$$

and the substitution to Eq. (4) gives

$$\left(X^T X\right)^{-1} X^T Y \left(Y^T Y\right)^{-1} Y^T X \boldsymbol{u} = \lambda^2 \boldsymbol{u}. \tag{7}$$

Thus, the solution of optimization task (1) is turned into a generalized eigenvalue problem (Yu *et al.*, 2007; Hardoon *et al.*, 2004) to find eigenvalue λ and corresponding eigenvector $\boldsymbol{u}$. Using Eq. (6), we find the vector $\boldsymbol{v}$. Note, the values of canonical correlations can be expressed such that $\lambda = \text{cor}\,(a, b)$.

The higher-order canonical variates and correlations are defined identically as in expression (1), but now under additional restriction so that a canonical variate of order k, with $1 < k \leq \min(p, q)$, should be uncorrelated with all the canonical variates of lower-order.

To ensure a symmetric standard eigenvalue problem, the matrix $X^T X$ is decomposed using the Cholesky decomposition (Hardoon *et al.*, 2004).[2] Because for any positive-semidefinite symmetric matrix A, the singular value decomposition of A is essentially the same as the spectral decomposition of A (Harville, 1997, p. 555), this can be solved by singular or spectral decompositions.

Note, if very small eigenvalues of matrix A are set equal to zero, the eigenvalue decomposition may provide a high-quality rank approximation to the original matrix (Harville, 1997, p. 559).

Regularized CCA

In cases of effective rank of data n lower than the corresponding number of variables, some sort of conditioning is needed in order to avoid numerical instability and degeneracy during the computation of CCA. Many numerical techniques to overcome this limitation have been presented in various works, of which the most prominent are order reduction and regularization (De Bie *et al.*, 2005; Hardoon *et al.*, 2004). In this study, the l_2 regularization (penalization) technique is used (Vinod, 1976) in the same manner as the regularization applied in ridge regression. Kuss and Graepel (2003) comprehensively review the advantages of this approach. It will be shown later that the parameters of regulari-

[2] Note that conventional CCA algorithm begins with performing singular value decomposition of $Q^T_x Q_y$, where Q_x and Q_y are matrices with orthogonal columns given by the QR decomposition of centered original data matrices. This approach is not utilized here due to the regularization step in the analysis (see below).

zed canonical correlation analysis (rCCA) interpolate smoothly between PLS and CCA.

In the regularization approach, some small perturbation (given by the parameter $k > 0$ and unit matrix *I)* is added to the diagonals of $X^T X$ and $Y^T Y$ correlation matrices in expression (2) as follows[3]

$$\boldsymbol{u}^T (X^T X + \kappa I)\boldsymbol{u} = 1, \boldsymbol{v}^T (Y^T Y + \kappa I)\boldsymbol{v} = 1. \quad (8)$$

rCCA corresponds to a modified optimization problem with linear combination of constraints. Although the manually predefined values of small perturbation decrease the stability of CCA method, the regularization algorithms are broadly used.

It should be noted that the course of minimization property within constraints of Eq. (8) remains active during the computation of canonical correlation analysis. We expect that rCCA decrease in magnitude for low-correlation directions within sets more than high-correlation directions within sets. Also, regularization has proven to be useful as a process of incorporating prior knowledge in data analysis (e.g. De Bie and De Moor, 2003; Poggio and Girosi, 1990). Particularly, in the study published by De Bie and De Moor (3003), the measurement noise is neglected in a specific way dealing with diagonal elements in constraints of canonical analysis.

In the case of $k \neq 0$, the canonical variates no longer have unit variances whose property must be taken into consideration at the computation of canonical correlations.

Partial least squares

Partial least squares method (PLS-SVD) is explained as maximizing the following term

[3] One different and frequently used regularization step of CCA (not applied here) is to use a convex combination of constraints as $\boldsymbol{u}^T ((1 - k)\, X^T X + kI)\boldsymbol{u} = 1$ and $\boldsymbol{v}^T ((1 - k)\, Y^T Y + kI)\boldsymbol{v} = 1$ for parameter boundary $\boldsymbol{k} \in \langle 0, 1 \rangle$.

$$\max_{u,v\neq 0} \frac{\boldsymbol{u}^T X^T Y \boldsymbol{v}}{\sqrt{\boldsymbol{u}^T \boldsymbol{u}}\sqrt{\boldsymbol{v}^T \boldsymbol{v}}}. \tag{9}$$

Because (9) is a homogenous function in $(u, v) \in R^p \times R^q$, this task is equivalent to finding local extrema of the optimization problem (De Bie *et al.*, 2005)

$$\max_{u,v\neq 0} \boldsymbol{u}^T X^T Y \boldsymbol{v} \tag{10}$$
$$s.t.\ \boldsymbol{u}^T \boldsymbol{u} = 1, \boldsymbol{v}^T \boldsymbol{v} = 1.$$

The combinations $a = X\boldsymbol{u}$ and $b = Y\boldsymbol{v}$ are called the latent variables (also called factors or components). The solution is analogous to the preceding one. We consider PLS a special case of CCA method if matrices $X^T X$ and $Y^T Y$ in constraints of expression (2) are both identity matrices.[4] Instead of correlations, the covariances are to be maximized in PLS.

According to Wegelin (2000) and De Bie *et al.* (2005), the PLS--SVD problem (contrary to other PLS methods and various "deflation procedures") can be solved directly by using singular value decomposition of matrix $X^T Y$, i.e. $X^T Y = UDV^T$ where U and V are orthogonal matrices of types (p, p) and (q, q) respectively, and D is a diagonal matrix of not necessarily distinct singular values. All the $r = rank\ (X^T Y)$ singular values are strictly positive (Harville, 1997, p. 551). In this case, the higher-order latent variables form the other covariances (eigenvalues) λ corresponding to different eigenvectors taking the orthogonality restriction into account as an additional constraint.

Due to the Eckhart-Young theorem, the (p, q) matrix B of rank $r \leq n$ which minimizes the absolute-square ("Frobenius norm") error between $X^T Y$ and B is given by $B = U\widetilde{D}V^T$ where D equals to $\widetilde{D}$, except with the last $n - r$ diagonal entries set to zero. Thus,

[4] Theoretically, this very strong condition of both original data sets corresponds also to column-orthogonally centered bases of those data.

the singular value decomposition of matrix $X^T Y$ using the first few dominant singular values may form a high-quality rank approximation to the original matrix (Harville, 1997, p. 556). In PLS method, it is possible to compute just $s = min\ (p, q)$ pairs of latent variables, even in cases when $s > r$. In this case, the last $s - r$ pairs of latent variables have zero covariance.

Note that if we consider the regularized expression (2), the conditions with $k = 0\ (n \geq p, n \geq q)$ are equal to classical CCA solution while the limit as $\kappa \to \infty$ corresponds to PLS[5] (De Bie *et al.*, 2005). This is after the rescaling of eigenvalues and corresponding eigenvectors.

Principal component analysis

While PLS and CCA methods in this paper are considered as methods comparing two sets of data, the principal component analysis is constructed to study relations within one set. The algebraic model to study relations among original variables can be described as maximizing the term (Krzanowski, 2000, p. 60)

$$\max_{\boldsymbol{u}\neq\boldsymbol{0}} \frac{\boldsymbol{u}^T X^T X \boldsymbol{u}}{\boldsymbol{u}^T \boldsymbol{u}}. \tag{11}$$

Because (11) is a homogenous function in $\boldsymbol{u} \in R^p$, this task is equivalent to finding local extrema of the constrained optimization problem

$$\begin{aligned} &\max_{\boldsymbol{u}\neq\boldsymbol{0}} \boldsymbol{u}^T X^T X \boldsymbol{u} \\ &s.t.\ \ \boldsymbol{u}^T \boldsymbol{u} = 1. \end{aligned} \tag{12}$$

The combinations $a = X\boldsymbol{u}$ are called the principal components (latent variables). Equivalently to CCA and PLS methods, the analytical solution according to the Kuhn-Tucker theorem (Kuhn

[5] At the convex combination of constraints as $\boldsymbol{u}^T ((1 - k)\ X^T X + kI)\boldsymbol{u} = 1$ and $\boldsymbol{v}^T ((1 - k)\ Y^T Y + kI)\boldsymbol{v} = 1$, the case $k = 1$ is exactly equal to the PLS solution.

and Tucker, 1951) can be defined as finding stationary points of the Lagrange function by eigenvalue decomposition of $X^T X$ matrix (Krzanowski, 2000, p. 62). This problem is equivalent to computing the singular value decomposition of standardized matrix X.

There are no inverses in the derivation of this method, thus the solution is not theoretically restricted by a rank-deficiency of data.

Because we are attempting to compare the properties of rCCA and PLS methods discussed in this chapter, the coefficients given by both sets, separately, are scaled to have unit norms. This approach is utilized in PCA method too, although this is solved as one set case. However, this scaling of rCCA and PCA methods is made *a posteriori*, and thus does not theoretically influence the results of the analyses.

Some properties of CCA and PLS methods

Taking into account the information from Table 1, the PLS method has one basic interesting property on first-order latent variables, as opposed to canonical correlation analysis. This is, if one variable of given set (to which variables is collinear) is added to or removed from the analysis, this has only a negligible effect on the coefficients of the set to which the variable is assigned.

Table 1. Comparison of first-order CCA and PLS methods

	CCA	PLS
Maximization task	cor ($X\boldsymbol{u}$, $Y\boldsymbol{v}$)	cov ($X\boldsymbol{u}$, $Y\boldsymbol{v}$)
Interpretation of $\boldsymbol{u}$ and $\boldsymbol{v}$ coefficients in both sets	no straightforward readings	$u_i \propto \text{cov}(X_{.i}, Y\boldsymbol{v})$ $v_j \propto \text{cov}(Y_{.j}, X\boldsymbol{u})$
X and/or Y sets are collinear	unstable due to constraints	stable
Adding variable to X set collinear to X_i	$\boldsymbol{u}_i$ changes very	$\boldsymbol{u}_i$ changes negligible

Source: Adopted from Wegelin (2000); with modifications

The partial least squares method is to maximize the covariance of individual variables of the first data set to latent variable of the second data set. If the set consists of only one variable, the coefficients of PLS reduces almost entirely to the two-di-

mensional covariance coefficients. This is the reason to consider processing the small changes of PLS parameters of given set during incorporation, or removing variables to the same data set. On the other hand, CCA technique simultaneously maximizes the correlations of latent variables given by both considered sets. In CCA with only one variable in the set, this technique is equivalent to multiple regression using classical least squares (see e.g. De Bie *et al.*, 2005).

If we denote Φ of type *(n, s)* as a matrix of latent variables for the *X* set, and Ω of type *(n, s)* as a matrix of latent variables for the *Y* set, the properties of CCA and PLS differ as follows:

- Generally, in CCA, $\Phi^T \Phi$ and $\Omega^T \Omega$ are diagonal matrices, and $\Phi^T \Omega$ is also diagonal.
- In PLS, $\Phi^T \Phi$ and $\Omega^T \Omega$ matrices are not necessarily diagonal, but $\Phi^T \Omega$ is diagonal.

While the coefficients $\boldsymbol{u}$ and $\boldsymbol{v}$ of PLS method are determined as the minimum norm solution given by the constraints $\boldsymbol{u}^T \boldsymbol{u} = 1$ and $\boldsymbol{v}^T \boldsymbol{v} = 1$, canonical correlation analysis considers the process of computation a part of relations within both sets. The minimum norm solution can be generally characterized as a subspace of solutions having the smallest length (i.e. is closest to the origin). Thus, the most important property is to spread the lengths among a large number of entries of $\boldsymbol{u}$ or $\boldsymbol{v}$ instead of putting all the solution lengths into just a few entries. In addition to that, such entries do not prove significant extreme values, compared to coefficients of canonical correlation analysis.

CCA task (1) is usually considered with unit variances of canonical variates $X\boldsymbol{u}$ and $Y\boldsymbol{v}$. However, some sort of secondary information given by the matrices $X^T X$ and $Y^T Y$ in canonical correlation analysis is incorporated by constraints. Intuitively, the coefficients corresponding to variables with a significant share of correlation relations within sets are to be decreased simultaneously in magnitude.

EXPERIMENTAL

Tourism data

In this application study, annual time series data 2003 – 2008 of numbers of visitors for selected European countries are presented. The tourism data are considered as departures of Czech visitors (residents) to their most popular European states. Those time behaviours (relations) in attendance are to be compared with the total number of foreign visitors (of European and non-European origin) to the areas considered. The scope of this study covers eleven European states.[6] For the range of EU Member States, the 95/57/EC guidelines were established on the collection of statistical information in the field of tourism. Quality control and validation checks are performed by Eurostat before releasing the data.

The sample survey, or census, is the main source of information at EU Member States level. Exceptionally, the data are compiled via visitor survey or border survey. The Czech Republic is a case of sample survey of gathering data where the departure information is compiled on the basis of long trips (more than 3 overnight stays) from a randomly chosen sample of households. Because of the change in methodology in 2009, this study covers the data only till 2008, inclusive. The annually-based data of Czech Republic departure tourism are provided by the Czech Statistical Office (CSO database). To study the total number of visitors in the chosen states, the Eurostat online database (Eurostat database) is utilized. Both sources of information are free of charge. In Eurostat database, arrivals at collective tourist accommodation establishments and arrivals of non-residents were chosen.

Due to an attempt to delineate mutual relations among visitor counts in different areas, and for explaining unusual cases, other

[6] Those states are the following (abbreviated with EU codes and in sequence of the number of Czech visitors in descending order): SK – Slovakia, HR – Croatia, IT – Italy, EL – Greece, AT – Austria, UK – United Kingdom, ES – Spain, FR – France, BG – Bulgaria, HU – Hungary, DE – Germany.

data can be compiled. In this study, information about meteorology is considered, i.e. land temperature over Europe.

Methods

The time processes in arrival tourism data of Czech visitors and the total of all non-residents to selected European states are examined in this study detecting basic relations between both sets. To avoid sources of tourism seasonality in time series data, only annual data are utilized. The advantage of such approach is that tourism processes have some degree of smoothness in the annual cycle. Annual data are fundamental in economy and allow studying various trends in arrivals. The interest is in finding similar behaviours between individual states (counts of Czech visitors and the total of non-residents) using PLS and canonical correlation methods as well as to mutually compare the applicability of both approaches. The PCA *ad hoc* method applied to one data set (given by the union of both sets) is utilized to help to identify differences between results of PLS and CCA techniques. The linear trends are also studied using PCA by inserting time variable separately to both data sets.

All the programs are developed by the author using MATLAB 7.1 (Mathworks, Natick, MA, USA) software platform. Their source codes are available on request. We use a number of built-in functions, especially those from linear algebra, i.e. singular, spectral and Cholesky matrix decompositions. In rCCA, the regularization parameter κ is set equal to 0.001.

RESULTS AND DISCUSSION

We extract three latent variables in all PLS, rCCA and principal component analysis and the differences of PLS and rCCA outputs are discussed. The PCA as applied to one set (unifying total number of non-residents and Czech visitors) gives rise to an overall tendency of data and is used particularly as an explanatory method. The percentage proportion on sum of eigenvalues and the coefficients of linear combinations of LVs

as outputs of methods used are presented in Table 2. The PLS method, particularly, gives a great indication for a lower-dimensional representation. It should be noted that not many relations of total visitors and the same-state Czech visitors were identified. If we concentrate our attention to first-order analyses, in PLS, Italy, the United Kingdom and Hungary prove similar time behaviours while Croatia, Spain and Germany do not. In rCCA, only France showed similar time behaviours of numbers of visitors and Croatia together with Germany prove opposite relations. On the other hand, many other types of linear relations are more obvious.

In real situations, original variables are often collinear or nearly collinear within sets. The absolute values of Pearson correlation coefficient within sets were studied and the values higher than 0.9 are considered as an indication of collinearity (Griffith and Amrhein, 1997).

Both data sets are characterized by collinear variables, especially the data of total number of non-residents providing 28 values higher than 0.9 from 55 of possible combinations. On the other hand, the data set of Czech visitors proved only one such value. This range of within sets relations can also be descriptively measured by the singular value decomposition (eventually by spectral decomposition) separately for both sets correlation matrices. Such an experiment gives an eigenvalue average equal to 1. In agreement with the preceding results, the total numbers of non-residents in the given states prove higher indication of collinearity (eigenvalue 9.035 on first latent variable) in comparison to the numbers of Czech visitors (eigenvalue 4.407 on first latent variable). Also, according to the rule of worth interpreting the latent variables corresponding to eigenvalues higher than 1, the former case (total visitors)[7] gives two latent variables while the latter case (Czech visitors) gives four latent variables.

[7] In the following, total means the total number of non-residents arriving at given state, while the abbreviation cz means exclusively Czech visitors.

Table 2. Coefficients of rCCA, PLS and PCA

	rCCA			PLS			PCA		
	LVs#1	LVs#2	LVs#3	LVs#1	LVs#2	LVs#3	LVs#1	LVs#2	LVs#3
Eigenvalues ratio (%)	20.17	20.14	20.12	71.65	12.50	10.80	59.88	14.25	12.46
Total number of visitors									
SK	0.379	0.148	0.142	0.337	-0.075	0.195	0.330	-0.182	0.044
HR	0.250	0.196	-0.005	0.323	0.194	0.121	0.325	0.120	0.163
IT	0.233	-0.280	0.004	0.315	-0.135	-0.230	0.323	-0.020	-0.302
EL	0.267	0.080	0.751	0.309	-0.336	0.394	0.312	-0.415	0.072
AT	0.424	0.453	0.049	0.336	0.115	0.408	0.322	-0.114	0.328
UK	0.134	-0.602	-0.398	0.276	-0.036	-0.696	0.290	0.238	-0.619
ES	0.320	-0.336	0.109	0.308	-0.359	-0.185	0.309	-0.260	-0.409
FR	-0.414	0.052	-0.033	-0.093	0.541	-0.133	-0.067	0.623	0.243
BG	0.248	0.000	-0.180	0.323	0.159	-0.110	0.326	0.168	-0.045
HU	0.191	0.413	-0.407	0.282	0.601	0.084	0.282	0.470	0.374
DE	0.316	-0.044	-0.211	0.333	0.059	-0.130	0.332	0.069	-0.125
Czech visitors									
SK	-0.004	0.338	-0.041	0.030	-0.107	0.427	0.034	-0.481	0.261
HR	-0.437	0.138	0.063	-0.463	0.007	0.200	-0.460	-0.096	0.146
IT	0.092	0.366	0.189	0.229	0.054	0.455	0.245	-0.254	0.386
EL	0.187	-0.212	0.403	0.194	-0.484	0.155	0.189	-0.517	-0.122
AT	0.034	0.331	-0.251	0.186	0.571	-0.053	0.205	0.496	0.224
UK	0.561	0.069	-0.255	0.439	-0.107	0.103	0.424	-0.152	0.126
ES	-0.171	-0.010	0.108	-0.316	-0.285	0.251	-0.328	-0.332	0.105
FR	-0.220	0.216	0.025	-0.315	-0.009	0.342	-0.319	-0.120	0.327
BG	0.074	0.676	-0.350	0.163	0.492	0.407	0.179	0.122	0.595
HU	0.163	0.016	0.727	0.214	-0.066	0.413	0.223	0.039	0.427
DE	-0.583	0.265	0.069	-0.444	0.293	0.129	-0.423	0.133	0.161

Source: Author

We consider the coefficients of the first LVs in explaining the relationships of PLS and rCCA methods. In both mentioned techniques, the number of visitors in Austria (total) has positive relation to the United Kingdom (cz), and the opposite to Croatia and Germany (cz). The opposite relation of French visitors (total) to the United Kingdom (cz), and so positive to Croatia and Germany (cz) on first latent variables of rCCA, are also evident. This is not the case in PLS where in particular Croatia, Italy, Greece, Bulgaria and Germany (total), and Spain and France (cz) acquire higher values of coefficients, relatively, to rCCA. If some of the PLS coefficients rather significant in relations between sets are to be lowered

in magnitude in rCCA, then those coefficients are significant in one set method, PCA (having also a high degree of within set relations). On the other hand, France (total), which has a high magnitude of coefficient on the first latent variable of rCCA (in comparison to PLS), proves low coefficient in principal component analysis.

Because of the existence of collinearity within data sets, the PLS method should be generally preferred to canonical correlation analysis. For that reason, we discuss PLS results in greater detail. First LVs are characterized mainly by similar time behaviours of total visitors to Slovakia, Austria and Germany, and Czech visitors to the United Kingdom, together with opposite relations to other numbers of Czech visitors to Croatia and Germany. Second LVs (which have a significantly lower percent proportion on the sum of eigenvalues, 12.50% compared to 71.65% on the first latent variables) reveal positive time behaviours of visitors to France and Hungary (total), and Austria and Bulgaria (cz) opposite to Greece (cz). Third LVs (with the eigenvalue percent proportion 10.80%) prove positive relations of Greece and Austria visitors (total) to Slovakia, Italy, Bulgaria and Hungary (cz). Also, negative relation of United Kingdom (total) to all mentioned states attended by Czech visitors is evident from those coefficients (of third latent variables).

It should be noted that basic features of mutual relations are still generally preserved in both rCCA and PLS methods, particularly considering first-order analyses. However, in some well-defined situations (given by collinearity within sets), the results of those methods differ.

The linear trends of standardized data are also studied in the following. Those trends are examined separately for both sets using PCA method and by considering the time variable (see Table 3). The first LVs are extracted which explain relatively high proportions of variances, i.e. 83.44% (total), and 44.73% (cz). All the states in total number of non-residents reveal a positive trend (growing numbers of arrivals over time) with the exceptions of France and only weak growth for Hun-

gary. In the case of Czech visitors, the situation is more complicated as the direction of growth is indicated to the United Kingdom. On the other hand, a decrease can be seen in Croatia, Spain and Germany (cz), and the weaker decrease is also indicated in France (cz). Because of the great extent of linear trends compared to all sources of linear relations in time series data, the comparison of linear trends of both data sets roughly corresponds to the results of first LVs of PLS, and also to rCCA (see above). But the analysis of trends reveals the directions of those data behaviours.

Table 3. PCA coefficients

TIME	SK	HR	IT	EL	AT	UK	ES	FR	BG	HU	DE
Total number of visitors											
0.313	0.311	0.309	0.310	0.298	0.301	0.281	0.294	-0.055	0.311	0.267	0.315
Czech visitors											
0.425	0.046	-0.400	0.262	0.152	0.229	0.356	-0.312	-0.277	0.213	0.236	-0.334

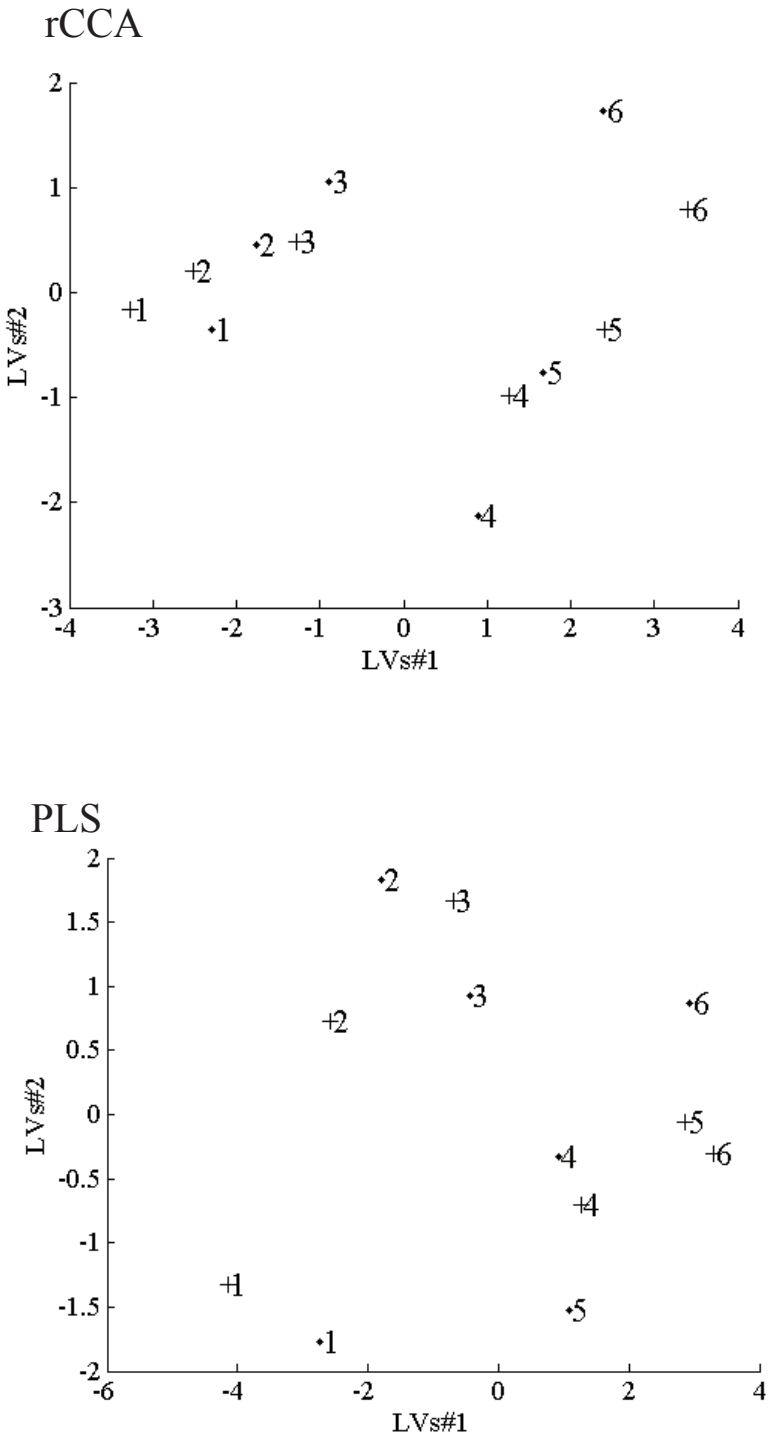
Source: Author

The scores of the first two LVs for rCCA and PLS methods are examined. In rCCA, there is one well-bounded cluster in data (Fig. 1) separating the first three years considered and the rest of the years. This is generally in accordance with the PLS score (Fig. 1) where the exception of the other outlier points occurs as cluster of the year 2003. In the scope of relations between sets, using an inspection of original annual 2005 and 2006 data, great changes were found in numbers of arrivals due to Croatia and Germany (total), and Italy, Spain and Germany (cz). This indicates the former separation, together with similarity of the other points in clusters rather close to corresponding boundary points. The reasons for the latter separation (outliers of year 2003) are more difficult.

Additionally, a number of other variables are not discussed in this study, in particular, dealing with the economic situations of individual states, the occurrence of global infectious diseases or various parameters of meteorological conditions. Although some variables, e.g. tourism accommodation capacity, prove a higher influence than the meteorological conditions

in specific situations (see e.g. Surugiu *et al.*, 2011), the mutual relations and trends in our study seem to also be influenced by a higher average temperature in Europe considering years 2006, 2007 and 2008 (EEA database) and the accession of new EU members.

Figure 1. rCCA and PLS score plots



(+ means total number of visitors, • means Czech visitors)
Source: Author

CONCLUSION

The main contribution of this study is to discuss the application of PLS and canonical correlation analysis methods, and aims to explain mutual relations and trends in tourism rank-deficient data. Although both methods reveal the basic features of data, it was proven by the results of this study (and also previously published by different authors), that the PLS method should be preferred in the case of collinearity within sets. This is crucial information because PLS method up till now has only rarely been applied in practice except for chemical disciplines such as chemical engineering and chemometrics. The approach of principal component analysis is an efficient algorithm for studying the overall tendency of data as well as revealing linear trends.

On the base of multivariate results investigating the selected European states, behaviour of the same-state arrivals of non-residents and Czech visitors is rather different. In particular, the following results are identified:

- Using first LVs of PLS, similar time behaviours are identified in the total number of non-residents of Slovakia, Austria and Germany, and Czech visitors in the United Kingdom, together with opposite relations to other numbers of Czech visitors, notably in Croatia and Germany.
- Particularly, total numbers of visitors to Slovakia, Croatia, Italy, Austria, Bulgaria and Germany growth, similarly to the United Kingdom numbers of Czech visitors, considering standardized data. On the other hand, Croatia, Spain, France and Germany (cz) show a decrease in arrivals.
- In rCCA and also in PLS, the first three years are separated from the rest, having a different decisive influence to between-set relations. The year 2003 is characterized by the other outliers in PLS.

ACKNOWLEDGEMENT

The author wishes to thank the unnamed statistician for his insightful comments which have led to the substantial improvement of the manuscript. Also, the financial support of the University of Business in Prague internal grant FRV No. 3/2013 is acknowledged.

REFERENCES

Alam, A.M., Nasser, M. and Fukumizu, K. (2010). A comparative study of kernel and robust canonical correlation analysis. *Journal of Multimedia*, *5*(*1*), 3-11.

Bach, F.R. and Jordan, M.I. (2005). *A probabilistic interpretation of canonical correlation analysis*. Technical report, University of California, California, US.

Bender, O., Schumacher, K.P. and Stein D. (2005). *Measuring seasonality in Central Europe's tourism – how and for what*? In: Schrenk, M. (Ed.): CORP 2005. *Proceedings of 10th Symposium on Information Technology in Urban- and Spatial Planning*, Vienna University of Technology. Vienna, pp. 303-9.

Cao, K.A., Rossouw, D., Robert-Granie, C. and Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, *7*(*35*), 1-23.

Constantin, B. and Grigorescu, A. (2009). *Interrelations between development factors and tourism factors. A quantitative point of view*. MPRA paper No. 25076. Munich University Library, Munich, Germany.

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

CSO database. URL: http://www.czso.cz/csu/redakce.nsf/i/cru40_cr, downloaded 7.6.2012.

De Bie, T., Cristianini, N. and Rosipal, R. (2005). *Eigenproblems in pattern recognition*. In: Bayro-Corrochano, E. (Ed.). *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neural Computing, and Robotics*, Springer-Verlag, Heidelberg, pp. 1-39.

De Bie, T. and De Moor, B. (2003). *On the regularization of canonical correlation analysis*. In: Amari, S.I., Cichocki, A., Makino, S., Murata, N. (Eds.). *Proceedings of the International Conference on Independent Component Analysis and Blind Source Separation (ICA2003)*, Nara, Japan, pp. 785-90.

EEA database. URL: http://www.eea.europa.eu/data-and-maps/figures/average-temperatures-in-europe-and-relative-heating-degree-days-in-eu27/, downloaded 3.7.2012.

Eurostat database. URL: http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database, downloaded 7.6.2012.

Griffith, D.A. and Amrhein, C.G. (1997). *Multivariate statistical analysis for geographers*. New Jersey: Prentice Hall.

Hardoon, D.R., Szedmak, S. and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16, 2639-64.

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

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(*6*), 417-41 and 498-520.

Hotelling, H. (1935). The most predictable criterion. *Journal of Educational Psychology*, *26*(*2*), 139-42.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika*, *28*(*3/4*), 321-77.

Huang, S.Y., Lee, M.H. and Hsiao, C.K. (2009). Nonlinear measures of association with kernel canonical correlation analysis and applications. *Journal of Statistical Planning and Inference*, *139*(*7*), 2162-74.

Iatu, C. and Bulai, M. (2011). New approach in evaluating tourism attractiveness in the region of Moldavia (Romania). *International Journal of Energy and Environment*, 2(5), 165-74.

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

Kuhn, H.W. and Tucker, A.W. (1951). *Nonlinear programming*. In: Neyman, J. (Ed.). *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, California, pp. 481-92.

Kuss, M. and Graepel, T. (2003). *The geometry of kernel canonical correlation analysis*. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Leen, G. and Fyfe, C. (2006). *A Gaussian process latent variable model formulation of canonical correlation analysis*. In: Gnoss, E. (Ed.). *European Symposium on Artificial Neural Network (ESANN)*, Bruges, Belgium, pp. 413-18.

Lorber, A., Wangen, L. and Kowalski, B. (1987). A theoretical foundation for the PLS algorithm. *Journal of Chemometrics*, *1*(*9*), 19-31.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine*, *2*(*6*), 559-72.

Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*(*4945*), 978-82.

Sampson, P.D., Streissguth, A.P., Barr, H.M. and Bookstein, F.L. (1989). Neurobehavioral effects of prenatal alcohol: Part II. Partial least squares analysis. *Neurotoxicology and Teratology*, *11*(*5*), 477-91.

Surugiu, C., Surugiu, M.-R., Frent, C. and Breda, Z. (2011). Effects of climate change on Romanian mountain tourism: Are they positive or mostly negative. *European Journal of Tourism, Hospitality and Recreation*, *2*(*1*), 42-71.

Tipping, M.E. and Bishop, C.M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, *61*(*3*), 611-22.

Vajčerová, I., Šácha, J. and Ryglová, K. (2012). Using principal component analysis for evaluating the quality of a tourist destination. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 60(2), 449-58.

Vinod, H.D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, *4*(*2*), 147-66.

Wegelin, J.A. (2000). *A survey of partial least squares (PLS) methods, with emphasis on two-block case*. Technical report, University of Washington, Seattle, US.

Wold, H. (1975). *Path models with latent variables: The NIPALS approach*. In: Blalock H.M. *et al.* (Eds.). *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, Academic, NY, pp. 307-57.

Wold, S., Eriksson, L., Trygg, J. and Kettaneh, N. (2004). *The PLS method – partial least squares projections to latent structures – and its applications in industrial RDP (research, development, and production)*. Technical report, Umeå University, Umeå, Sweden.

Yu, S., De Moor, B., Moreau, Y. (2007). *Learning with heterogeneous data sets by weighted multiple kernel canonical correlation analysis*. In: Diamantaras, K. *et al.* (Eds.). *Proceedings of the Machine Learning for Signal Processing XVII*, IEEE, Thessaloniki, Greece, pp. 81-6.

Submitted: 11th October, 2012 *Accepted: 29th September, 2013*
Final version: 12th May, 2013 *Refereed anonymously*