Journal of Data Science 79-92 , DOI: 10.6339/JDS.201801_16(1).0005

Determination of the Effective Economic and/or Demographic Indicators in Classification of European Union Member and Candidate Countries Using Partial Least Squares Discriminant Analysis

Esra Polat

Abstract: Partial Least Squares Discriminant Analysis (PLSDA) is a statistical method for classification and consists of a classical Partial Least Squares Regression in which the dependent variable is a categorical one expressing the class membership of each observation. The aim of this study is both analyzing the performance of PLSDA method in classifying 28 European Union (EU) member countries and 7 candidate countries (Albania, Montenegro, Serbia, Macedonia FYR, Turkey moreover including potential candidates Bosnia and Herzegovina and Kosova) correctly to their pre-defined classes (candidate or member) and determining the economic and/or demographic indicators, which are effective in classifying, by using the data set obtained from database of the World Bank.

Keywords: classification; demographic indicators; economic indicators; European Union; Partial Least Squares Discriminant Analysis

1. Introduction

Integration in the European Union (EU) is one of the primary objectives of the government's policy of the many countries in Europe. Nowadays, there are 28 member countries and 7 candidate countries (Albania, Montenegro, Serbia, Macedonia FYR, Turkey moreover including potential candidates Bosnia and Herzegovina and Kosova) for EU. Although Turkey applied for associate membership in the European Economic Community in 1963, Turkey is still an official candidate (Lorcu and Acar Bolat, 2012).

Both the new accession countries and EU will have certain advantages and costs from the enlargement. The accession countries have three major advantages: the access to new and large market, great possibility for labour migration and access to significantly high EU funds. On the other hand, joining the EU may mean great costs, since vast market may mean severe competition. Integration in the EU also means implementation of great number of EU regulations and legislative, as one aspect of major adjustment problems. Yet, new accession countries have more benefits from the enlargement than the EU (Trpkova and Tevdovski, 2010).

One advantage for the EU is securing its own values throughout the newly accepted countries. Yet, new countries may also mean significantly difficult union to govern. Also, increase in population is greater than the increase in the Gross Domestic Product (GDP). The disproportion may burden the EU economy. Another problem may be the large number of immigrant workers, yet this sometimes can be taken as advantage in terms of low–cost working force. Another financial burden is financing the necessary adjustments of the new accession countries. Also, the financial benefits that the EU will provide may mean potential loss of job and business in the "sensitive" manufacturing industries and in agriculture in the EU because of the penetration of goods from the east (Trpkova and Tevdovski, 2010).

The purpose of this study is to show and introduce how Partial Least Squares Discriminant Analysis (PLSDA) method, which is not familiar for many researchers in economics, is successful in classification. Moreover, determining the economic and/or demographic indicators, which are most important in classifying the EU member countries and candidate countries (including potential candidates Bosnia and Herzegovina and Kosova) to their pre-defined classes for the years 2014 and 2015.

2. Partial Least Squares Regression

Partial Least Squares (PLS) method was first developed by Herman Wold in the 1960s and 1970s to address problems in econometric path modeling and was subsequently adopted by his son Svante Wold (and many others) in the 1980s for regression problems in chemometric modeling (Boulesteix, 2006). PLS regression (PLSR) method in its basic form applies for one single Y-variable and this method is non-iterative. It can be modified to accommodate two or more Y-variables simultaneously (Martens and Naes, 1989). PLSR is a method which relates the variations in one or several Y variables to the variations of several X variables by using the components instead of original X variables (Phatak and De Jong, 1997). Each of these components is obtained by maximizing the covariance between y and all possible linear functions of X (Naes et al., 2002). That is why PLS is called a supervised method in contrast to principal component analysis (PCA) which does not use the y for the construction of the new components (Boulesteix, 2006). In the case of PLSR, the covariance structure of Y also influences the computations (XLSTAT, 2015).

PLSR is a method that extracts the latent variables (LVs), which serve as a new predictors and regresses the dependent variables on these new predictors. PLSR comprises of regression and classification tasks as well as dimension reduction techniques and modelling tools. Therefore, it could be applied as a discrimination tool and dimension reduction method similar to PCA (Rosipal and Krämer, 2006). This method is quick, efficient and optimal for a criterion based on covariances. It is recommended in cases where the number of variables is high, and where it is likely that the independent variables are correlated. The idea of PLSR is to create, starting from a table with n observations described by p variables, a set of h components with $h<p$. The PLS method presents the advantage of handling missing data. The determination of the number of components to keep is usually based on a criterion that involves a cross-validation (CV) (XLSTAT, 2015). CV is a method used for selecting the optimal number of components, which maximize model's predictive ability for PLSR method (Naes et al., 2002).

2.1. Partial Least Squares Regression Model

PLSR models the relationship between these two blocks via score vectors. PLSR decomposes X and Y variables as in Eq. (1) and Eq. (2), respectively. Here T and U are matrices of score vectors (components, latent vectors); P and Q represent loading matrices; and E and F represent residual matrices. This decomposition is done to maximize the covariance between T and U. Here h, is the number of components (Polat and Gunay, 2015; Rosipal and Krämer, 2006; Wold et al., 2001).

(1)

(2)

The forming of the PLSR model, a lower number of components are used instead of using all the independent variables by constructing new variables. The new variables are called X scores and denoted with T score matrix. T score matrix is formed with the linear combinations of the multiplication of original X matrix with the weight matrix as shown in Eq. (3). In addition, T's are good predictors of Y and the equation of the PLSR model can be written as in Eq. (4) (Polat and Gunay, 2015; Wold et al., 2001).

(3)

(4)

Here C is the Y-weight matrix and F is the Y-residual matrix. Using the Eq. (3), Eq. (4) can be rewritten to look as a multiple regression model as in Eq. (5) (Polat and Gunay, 2015; Wold et al., 2001).

$$\begin{aligned} Y &= T_h C_h' + F_h \\ &= XW_h^* C_h' + F_h \\ &= XW_h \left(P_h' W_h\right)^{-1} C_h' + F_h \end{aligned} \tag{5}$$

Finally, the matrix B of the PLSR coefficients of Y on X, with h components generated by the PLSR algorithm is given by Eq. (6) (Polat and Gunay, 2015; Wold et al., 2001).

$$B = W_h \left(P_h' W_h\right)^{-1} C_h' \tag{6}$$

It is well known that there are several ways to calculate PLSR model parameters. Perhaps the most intuitive method, which is also called as a classical algorithm, known as Non-Linear Iterative Partial Least Squares (NIPALS) (Wold et al., 2001; Polat and Gunay, 2015).

2.2. Non-Linear Iterative Partial Least Squares (NIPALS) Algorithm

NIPALS calculates scores, T and loadings, P and an additional set of vectors known as weights, W (with the same dimensionality as the loadings P). The addition of weights in PLSR is required to maintain orthogonal scores (Wold et al., 2001). The simple NIPALS algorithm of Wold et al. (1984) is shown as below. It starts with optionally transformed, scaled, centered data (X and Y) and proceeds as follows. If there is a single y-variable, the algorithm is non-iterative.

A. Get a starting vector of u, usually one of the Y columns. With a single y, u=y.

B. The X-weights, w: $\mathbf{w} = \mathbf{X}'\mathbf{u}/\mathbf{u}'\mathbf{u}$. Scale w to be of length one.

C. Calculate X-scores, t: $\mathbf{t} = \mathbf{X}\mathbf{w}$

D. The Y-weights, c: $\mathbf{c} = \mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$

E. Finally, an updated set of Y-scores, u: $\mathbf{u} = \mathbf{Y}\mathbf{c}/\mathbf{c}'\mathbf{c}$

F. Convergence is tested on the change in t, i.e., $\|\mathbf{t}_{\text{old}} - \mathbf{t}_{\text{new}}\| / \|\mathbf{t}_{\text{new}}\| < \varepsilon$. Where ε is very small positive number, e.g., 10^{-6} or 10^{-8}. If convergence has not been reached, return to B, otherwise continue with G and then A. If there is only one y-variable, the procedure converges in a single iteration and one proceeds directly with G.

G. Remove (deflate) the present component from X and Y and then use these deflated matrices as new X and Y, while computing the next component. Here the deflation of Y is optional, the results are equivalent whether Y is deflated or not.

$$\begin{aligned} &\text{X-loadings: } \mathbf{p} = \mathbf{X}'\mathbf{t}/(\mathbf{t}'\mathbf{t}) \\ &\text{Y-loadings: } \mathbf{q} = \mathbf{Y}'\mathbf{u}/(\mathbf{u}'\mathbf{u}) \\ &\text{Regression (u upon t): } \mathbf{b} = \mathbf{u}'\mathbf{t}/(\mathbf{t}'\mathbf{t}) \\ &\text{Residual matrices: } \mathbf{X} \rightarrow \mathbf{X} - \mathbf{t}\mathbf{p}' \text{ and } \mathbf{Y} \rightarrow \mathbf{Y} - \mathbf{b}\mathbf{t}\mathbf{c}' \end{aligned}$$

H. Continue with next component (back to step A) until CV (see below) indicates that there is no more significant information in X about Y.

The next set of iterations of algorithm starts with the new X and Y matrices as the residual matrices from the previous iteration. The iterations can continue until a stopping criteria is used or X becomes the zero matrix (Wold et al., 2001).

2.3. Determination of the Ideal Number of Components Retaining in PLSR Models

First of all, the optimal number of components for the model must be chosen. Hence, quality indexes could be used for this purpose. The Q^2_{cum} index measures the global contribution of the h first components to the predictive quality of the model (and of the sub-models if there are several dependent variables). The $Q^2_{cum(h)}$ index can be given as in Eq. (7):

$$Q^2_{cum(h)} = 1 - \prod_{j=1}^{h} \frac{\sum_{k=1}^{q} PRESS_{kj}}{\sum_{k=1}^{q} SSE_{k(j-1)}} \tag{7}$$

Prediction sum of squares (PRESS) statistics is a measure, which assesses model's validation and predictive ability. In general, the smaller the PRESS value, the better the model's predictive ability (Naes et al., 2002). The index involves the PRESS statistic (that requires a cross-validation) is the predicted sum of squares of a model containing h components and the Sum of Squares of Errors (SSE) is the residual sum of squares of a model containing *h*–1 components. *PRESS* is computed by cross-validation as shown in Eq. (8). Here $\hat{y}_{j-1,-i}$ represents the predicted y-value for observation *i* based on *j–1* components when observation *i* was left out of the estimation of the regression parameters (Pérez-Enciso and Tenenhaus, 2003).

$$PRESS_j = \sum_{i=1}^{n} \left(y_{j-1,i} - \hat{y}_{j-1,-i} \right)^2 \tag{8}$$

The quality of classification model was visualized as graphic which contained cumulative value of Q^2, R^2X and R^2Y. The cumulative value of Q^2 is the indicator of global goodness of fit and predictive capability of model using certain amount of components. It is similar to R^2 from cross validation process (Barker and Rayens, 2003). The correlation between X and Y to the related components is performed by R^2X and R^2Y value. The search for the maximum of the Q^2cum index is equivalent to finding the most stable model. The R^2Ycum index is the sum of the coefficients of determination between the dependent variables and the h first components. It is therefore a measure of the explanatory power of the h first components for the dependent variables of the model. The R^2Xcum index is the sum of the coefficients of determination between the independent variables and the h first components. It is therefore a measure of the explanatory power of the h first components for the independent variables of the model (Ibrahim, 2009; Rohman et al, 2016).

3. Partial Least Squares Discriminant Analysis

PLS was not originally designed as a tool for statistical discrimination. In spite of this, applied scientists routinely used PLS for classification and there is substantial empirical evidence to suggest that it performs well in that role. Using PLS in this manner (PLS-LDA) had heuristic support owing to the relationship between PLS and canonical correlations analysis (CCA) and the relationship, in turn, between CCA and linear discriminant analysis (LDA). Barker and Rayens (2003) handled PLS as a penalized canonical correlation analysis. PLS is surely to be preferred over PCA when discrimination is the goal and dimension reduction is required, since at least with PLSDA information involving group separation is directly involved in the structure extraction (Barker and Rayens, 2003; Liu and Rayens,

2007, Polat et al., 2009). The uniqueness of PLSDA is the capability to construct the classification function. Objects are directed to certain class or group if the required passing grade of the group is achieved. Thus, PLSDA is belonged to supervised pattern recognition rather than PCA which classify the objects based on the similarity on Principal Component (PC) and lead to unsupervised classification (Rohman et al, 2016).

PLSR can be adapted to fit discriminant analysis. The PLSDA uses the PLS algorithm to explain and predict the membership of observations to several classes using quantitative or qualitative independent variables. PLSDA is a PLSR of a set Y of binary variables describing the categories of a categorical variable on a set X of predictor variables. It is a compromise between the usual discriminant analysis and a discriminant analysis on the significant principal components of the predictor variables (Pérez-Enciso and Tenenhaus, 2003). NIPALS algorithm is called as "PLS1" for the case where there is only one dependent variable (q=1) and "PLS2" for the case where there are several dependent variables (Hubert and Vanden Branden, 2003). NIPALS method is a method presented by Wold (1973) allowing PCA with missing values (Wold, 1973; XLSTAT-Missing Data Imputation using NIPALS in Excel Tutorial, 2017). XLSTAT-PLS uses the PLS2 algorithm applied on the full disjunctive table obtained from the qualitative dependent variable. PLSDA can be applied in many cases when classical discriminant analysis cannot be applied. For example, when the number of observations is low and when the number of independent variables is high. When there are missing values, PLSDA can be applied on the data that is available. Finally, as PLSR, it is adapted when multicolinearity between independent variables is high. As many models as categories of the dependent variable are obtained. An observation is associated to the category that has an equation with the highest value. Let k be the number of categories of the dependent variable Y. For each category an equation of the model is obtained as in Eq. (9) (XLSTAT, 2015).

(9)

With being a category of the dependent qualitative variable, being the intercept of the model associated to , p being the number of independent variables and being the coefficients of the same model. Observation i is associated to class k if (XLSTAT, 2015):

(10)

3.1. PLSDA Specific Results

PLSDA offers an interesting alternative to classical linear discriminant analysis. The output mixes the outputs of the PLSR with classical discriminant analysis outputs such as confusion matrix.

Classification functions: The classification functions can be used to determine which class an observation is to be assigned to using values taken for the various independent variables. These functions are linear. An observation is assigned to the class with the highest classification function F() as in Eq. (9) (XLSTAT, 2015).

Prior and posterior classification and scores: This table shows for each observation its membership class defined by the dependent variable, the membership class as deduced by the membership probabilities and the classification function score for each category of the dependent variable (XLSTAT, 2015).

Confusion matrix for the estimation sample: The confusion matrix is deduced from prior and posterior classifications together with the overall percentage of well-classified observations. The confusion matrix summarizes the reclassification of the observations, and allows to quickly seeing the

% of well classified observations, which is the ratio of the number of observations that have been well classified over the total number of observations (XLSTAT, 2015; XLSTAT- Partial Least Squares Discriminant Analysis PLSDA Tutorial, 2017).

4. Application and Results

The aim of this application study is both analyzing the performance of PLSDA method in classifying the 28 EU member countries and 7 candidate countries (Albania, Montenegro, Serbia, Macedonia FYR, Turkey moreover including potential candidates Bosnia and Herzegovina and Kosova) correctly to their pre-defined classes (candidate or member) and determining the most effective economic and/or demographic indicators in classification by using the variables obtained from database of the World Bank for the years 2014 and 2015. Leaving the political issues aside, the analysis is only concerned with the economic and demographic variables that have potential influence on country's eligibility for EU entrance. These economic and demographic variables, determined considering the study of Altas and Turgan (2008), are given in Table 1.

Table 1: The names of the economic and demographical variables in the analysis.

Economic Variables	**DIR:** Deposit interest rate (%)
	EGS: Exports of goods and services (% of GDP)
	EBGS: External balance on goods and services (% of GDP)
	GDP: GDP (current US$)
	GDP growth: GDP growth (annual %)
	GDPP_PCG: GDP per capita growth (annual %)
	GDP, PPP: GDP, PPP (current international $)
	GNI: GNI (current US$)
	GNI_PC, PPP: GNI per capita, PPP (current international $)
	GNI, PPP: GNI, PPP (current international $)
	GDS: Gross domestic savings (% of GDP)
	GNE: Gross national expenditure (% of GDP)
	IGS: Imports of goods and services (% of GDP)
	INF: Inflation, consumer prices (annual %)
	GE: Goods exports (BoP, current US$)
	GI: Goods imports (BoP, current US$)
	PPP_CF: PPP conversion factor, GDP (LCU per international $)
	TR: Total reserves (includes gold, current US$)
	Trade: Trade (% of GDP)
Demographic Variables	**MRI:** Mortality rate, infant (per 1,000 live births)
	PA_0-14: Population ages 0-14 (% of total)
	PA_65+: Population ages 65 and above (% of total)
	PG: Population growth (annual %)

Since there are missing values in the data set (for both the training sample of 2014 and prediction sample of 2015), PLSDA analysis is the best method for classification. The NIPALS algorithm is applied on the dataset and the obtained PCA model is used to predict the missing values. Then, the optimal number of components for the model must be chosen. Hence, quality indexes could be used for this purpose. Table 2 displays the model quality indexes. The quality corresponds here to the cumulated contribution of the components to the indexes. PLS has selected four components (Comp1,…,Comp4) automatically. The values of Q^2cum, the R^2Ycum and R^2Xcum with the four components are 0.533, 0.878 and 0.774, respectively. This indicates that the four components generated by the PLSR summarize well both the Xs and the Y.

Table 2: Model quality indexes for selected 4 components.

Statistic	Comp1	Comp2	Comp3	Comp4
Q^2 cum	0.599	0.533	0.484	0.533
R^2Y cum	0.681	0.786	0.836	0.878
R^2X cum	0.255	0.426	0.649	0.774

A bar chart is also displayed as showed in Figure 1 to allow the visualization of the evolution of the three indexes when the number of components increases. From Figure 1 it is clear that while the R^2Ycum and R^2Xcum indexes necessarily increase with the number of components, this is not the case with Q^2cum. It is seen that Q^2 remains low even with 4 components. This suggests that the quality of the fit varies a lot depending on the EU membership.

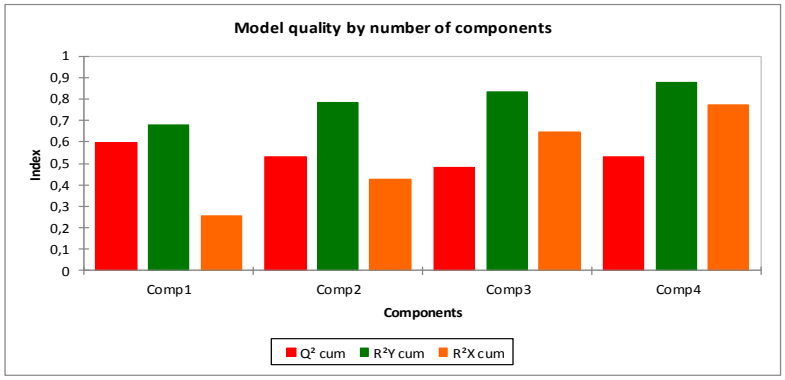

Figure 1: The chart of Q^2cum, R^2Ycum and R^2Xcum indexes for increasing number of components (automatically selected 4 components).

The table of the standardized coefficients (also named beta coefficients) in Table 3 allows comparing the relative weight of the variables in the model. To compute the confidence intervals, in the case of PLSR, the classical formulae based on the normality hypotheses used in Ordinary Least Squares (OLS) regression do not apply. A bootstrap method allows estimating the confidence intervals. The greater the absolute value of a coefficient, the greater the weight of the variable in the model. When the confidence interval around the standardized coefficients includes 0, which can easily be observed on the chart given in Figure 2, the weight of the variable in the model is not significant.

Table 3: Standardized coefficients of the model.

Variable	Coefficient	Std. deviation	Lower bound (95%)	Upper bound (95%)
DIR	0.040	0.111	-0.186	0.266
EGS	0.047	0.136	-0.230	0.324
EBGS	**-0.310**	**0.122**	**-0.558**	**-0.062**
GDP	0.015	0.086	-0.159	0.189
GDP growth	0.006	0.147	-0.292	0.304
GDP_PCG	-0.069	0.142	-0.358	0.219
GDP, PPP	-0.007	0.078	-0.166	0.152
GNI	0.018	0.087	-0.159	0.195
GNI_PC, PPP	-0.102	0.064	-0.233	0.029
GNI, PPP	-0.002	0.080	-0.164	0.160
GDS	**-0.248**	**0.114**	**-0.480**	**-0.016**
GNE	**0.310**	**0.122**	**0.062**	**0.558**
IGS	0.176	0.189	-0.207	0.560
INF	0.114	0.124	-0.137	0.365
GE	0.094	0.127	-0.164	0.352
GI	0.029	0.097	-0.169	0.227
PPP_CF	0.092	0.128	-0.169	0.352
TR	-0.048	0.079	-0.207	0.112
Trade	0.106	0.160	-0.220	0.431
MRI	-0.025	0.133	-0.295	0.245

PA_0-14	0.116	0.103	-0.094	0.327
PA_65+	-0.241	0.165	-0.575	0.094
PG	0.060	0.136	-0.216	0.335

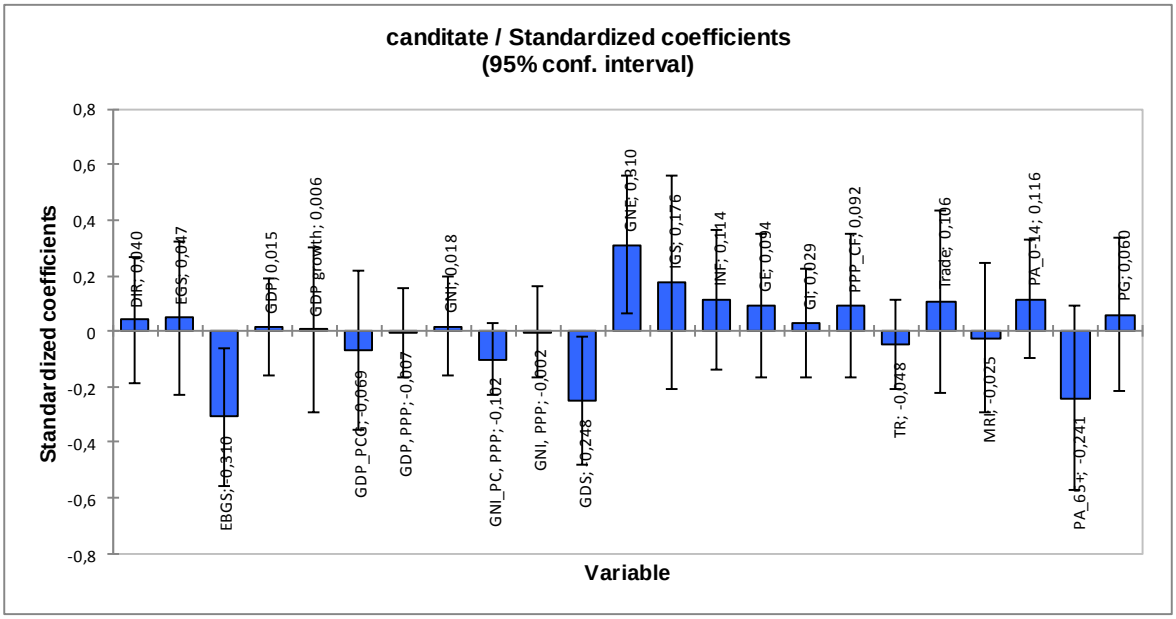

Figure 2: 95% Bootstrap confidence intervals of the standardized coefficients of the PLSR model.

Both of the results in Table 3 and Figure 2 indicate that the most important variables that determine the statuses of member and candidate countries in terms of economic indicators are identified as “external balance on goods and services (% GDP)”, “gross domestic savings (% GDP)” and “gross national expenditure (% GDP)” that means economical structure of countries is the most important determinant of EU membership.

The performance of PLSDA in discriminating the countries can be evaluated by confusion matrix as shown in Table 4. It is clear from Table 4 that the value of percentage correctness of 100 % indicate that overall of the 35 countries are classified correctly for the training sample of year 2014. This result can be easily also verified by examining the Pred(EU) column given in Table 5. It is seen that all countries are correctly classified to their pre-defined classes (member or candidate).

Table 4: Confusion matrix of PLSDA for discrimination of EU and candidate countries (for the training sample of 2014).

from \ to	**canditate**	**member**	**Total**	**% correct**
canditate	7	0	7	100.00%
member	0	28	28	100.00%
Total	7	28	35	100.00%

Table 5: Prior and posterior classification and scores (for the training sample of 2014).

Observation	EU	Pred(EU)	F(canditate)	F(member)	P(canditate)	P(member)
Austria	member	member	-0.043	1.043	0.253	0.747
Belgium	member	member	0.207	0.793	0.358	0.642
Bulgaria	member	member	0.015	0.985	0.275	0.725
Croatia	member	member	0.002	0.998	0.270	0.730
Cyprus	member	member	0.335	0.665	0.418	0.582
Czech Republic	member	member	-0.041	1.041	0.253	0.747
Denmark	member	member	-0.116	1.116	0.226	0.774
Estonia	member	member	-0.016	1.016	0.263	0.737
Finland	member	member	0.060	0.940	0.293	0.707
France	member	member	0.172	0.828	0.342	0.658
Germany	member	member	-0.169	1.169	0.208	0.792
Greece	member	member	0.015	0.985	0.275	0.725
Hungary	member	member	0.126	0.874	0.321	0.679
Ireland	member	member	-0.094	1.094	0.234	0.766
Italy	member	member	-0.108	1.108	0.229	0.771
Latvia	member	member	0.060	0.940	0.293	0.707
Lithuania	member	member	0.060	0.940	0.293	0.707
Luxembourg	member	member	-0.236	1.236	0.187	0.813
Malta	member	member	0.204	0.796	0.356	0.644
Netherlands	member	member	-0.040	1.040	0.253	0.747
Poland	member	member	0.118	0.882	0.318	0.682
Portugal	member	member	-0.046	1.046	0.251	0.749
Romania	member	member	0.037	0.963	0.284	0.716
Slovak Republic	member	member	0.240	0.760	0.373	0.627
Slovenia	member	member	-0.080	1.080	0.239	0.761
Spain	member	member	-0.059	1.059	0.246	0.754
Sweden	member	member	-0.149	1.149	0.214	0.786
United Kingdom	member	member	0.229	0.771	0.368	0.632
Albania	canditate	canditate	0.948	0.052	0.710	0.290
Montenegro	canditate	canditate	0.920	0.080	0.698	0.302
Serbia	canditate	canditate	0.698	0.302	0.598	0.402
Macedonia. FYR	canditate	canditate	0.813	0.187	0.652	0.348
Turkey	canditate	canditate	1.015	-0.015	0.737	0.263
Kosovo	canditate	canditate	1.021	-0.021	0.739	0.261
Bosnia and Herzegovina	canditate	canditate	0.903	0.097	0.691	0.309

Then, the model validated to prove the predictive ability by using the data set for the year 2015. For prediction sample it is seen from Table 6 that %97.14 of the countries are correctly classified. From Table 7 it is clear that both the membership probability and the classification function score of F(member)=0.831 and P(member)=0.660 are greater than F(canditate)=0.169 and P(canditate)=0.340 for Bosnia and Herzegovina. Hence, Bosnia and Herzegovina is predicted as a member of EU wrongly as it is still a potential candidate for EU.

Table 6: Confusion matrix of PLSDA for discrimination of EU and candidate countries (for the prediction sample of 2015).

from \ to	**canditate**	**member**	**Total**	**% correct**
canditate	6	1	7	85.71%
member	0	28	28	100.00%
Total	6	29	35	97.14%

Table 7: Prior and posterior classification and scores (for the prediction sample of 2015).

Observation	**Pred(EU)**	**F(canditate)**	**F(member)**	**P(canditate)**	**P(member)**
Austria	member	-0.099	1.099	0.232	0.768
Belgium	member	0.173	0.827	0.342	0.658
Bulgaria	member	-0.036	1.036	0.255	0.745
Croatia	member	-0.060	1.060	0.246	0.754
Cyprus	member	0.275	0.725	0.389	0.611
Czech Republic	member	-0.087	1.087	0.236	0.764
Denmark	member	-0.138	1.138	0.218	0.782
Estonia	member	-0.019	1.019	0.262	0.738
Finland	member	-0.040	1.040	0.253	0.747
France	member	0.085	0.915	0.304	0.696
Germany	member	-0.152	1.152	0.213	0.787
Greece	member	-0.084	1.084	0.237	0.763
Hungary	member	0.092	0.908	0.307	0.693
Ireland	member	-0.810	1.810	0.068	0.932
Italy	member	-0.199	1.199	0.198	0.802
Latvia	member	0.023	0.977	0.278	0.722
Lithuania	member	0.107	0.893	0.313	0.687
Luxembourg	member	-0.348	1.348	0.155	0.845
Malta	member	0.234	0.766	0.370	0.630
Netherlands	member	-0.094	1.094	0.234	0.766
Poland	member	0.007	0.993	0.272	0.728
Portugal	member	-0.068	1.068	0.243	0.757
Romania	member	-0.044	1.044	0.252	0.748
Slovak Republic	member	0.228	0.772	0.367	0.633
Slovenia	member	-0.154	1.154	0.213	0.787
Spain	member	-0.122	1.122	0.224	0.776
Sweden	member	-0.200	1.200	0.198	0.802
United Kingdom	member	0.154	0.846	0.333	0.667
Albania	canditate	0.848	0.152	0.668	0.332
Montenegro	canditate	0.917	0.083	0.697	0.303
Serbia	canditate	0.567	0.433	0.533	0.467
Macedonia. FYR	canditate	0.752	0.248	0.623	0.377
Turkey	canditate	0.914	0.086	0.696	0.304
Kosovo	canditate	1.410	-0.410	0.860	0.140
Bosnia and Herzegovina	**member**	**0.169**	**0.831**	**0.340**	**0.660**

The following chart in Figure 3 represents the observations on the t axes. It allows confirming that the EU member and candidate countries are very well discriminated on the factor axes extracted from the original independent variables.

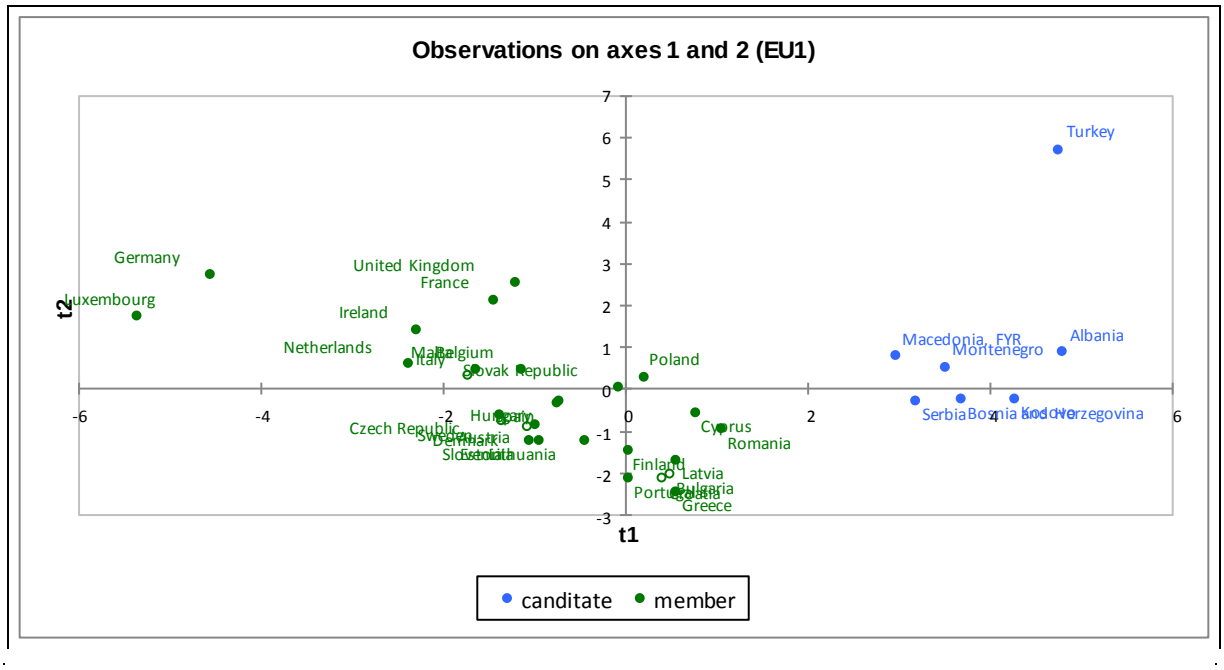

Figure 3: The chart of the observations on the first two components.

5. Conclusion

As a result of the PLSDA, the value of percentage correctness of 100 % indicate that overall of the 35 countries are classified correctly. Moreover, the most important variables that determine the statuses of member and candidate countries in terms of economic indicators are identified as “external balance on goods and services (% GDP)”, “gross domestic savings (% GDP)” and “gross national expenditure (% GDP)” that means for the 2014 economical structure of countries is the most important determinant of EU membership. Subsequently, the model validated to prove the predictive ability by using the data set for 2015. For prediction sample, %97.14 of the countries correctly classified. An interesting result is obtained for only Bosnia and Herzegovina, which is still a potential candidate for EU, predicted as a member of EU by using the indicators data set for 2015 as a prediction sample. However, as mentioned in Agir and Gursoy (2016) “Although Bosnia and Herzegovina has made significant transformation from a war torn country to a semi-functional state, ethnic tensions, nationalistic rhetoric and political disagreements are still evident which inhibit Bosnian progress towards the EU.”

References

[1] Agir, B.S. and Gursoy, B. (2016). The European Union's State-Building Efforts in the case of Bosnia and Herzegovina, *Ankara Avrupa Çalışmaları Dergisi* **15(1)**, 1-27.

[2] Altas. D. and Turgan. S.G. (2008). Avrupa Birliği (AB) ve OECD'ye Üyelikte Etkili olan Ekonomik ve Demografik Değişkenlerin İncelenmesi, *Marmara Üniveristesi İ.İ.B.F. Dergisi* **Cilt XXIV. Sayı 1**, 285-298.

[3] Barker, M. and Rayens, W. (2003). Partial Least Squares for Discrimination, *Journal of Chemometrics* **17**, 166-173.

[4] Boulesteix, AL and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, *Briefings in Bioinformatics* **8 (1)**, 32-44.

[5] Hubert, M. and Vanden Branden, K. (2003). Robust Methods for Partial Least Squares Regression, *Journal of Chemometrics* **17**, 537-549.

[6] Ibrahim, M.A.M. (2009). Comparison Between Different Procedures to Determine the Relative Importance of the Lifetime Performance Traits in Predicting Breeding Values of Holstein Cows, *Egyptian Journal of Animal Production* **46(2)**, 93-102.

[7] Liu, Y. and Rayens, W. (2007). PLS and Dimension Reduction for Classification, *Computational Statistics* **22**, 189-208.

[8] Lorcu, F. and Acar Bolat, B. (2012). Comparison Member and Candidate Countries to the European Union by Means of Main Health Indicators, *China-USA Business Review*. *ISSN 1537-1514* **Vol. 11. No. 4**., 556-563.

[9] Martens, H. and Naes, T. (1989). *Multivariate Calibration.* New York, Brisbane, Toronto, Singapore: John Wiley & Sons.

[10] Naes, T., Isaksson, T., Fearn, T. and Davies, T. (2002). *A User-Friendly Guide to Multivariate Calibration and Classification*. UK: NIR Publications Chichester.

[11] Pérez-Enciso M. and Tenenhaus, M. (2003). Prediction of Clinical Outcome with Microarray Data: a Partial Least Squares Discriminant Analysis (PLS-DA) Approach, *Human Genetics* **112**, 581–592.

[12] Phatak, A. and De Jong, S. (1997). The Geometry of Partial Least Squares, *Journal of Chemometrics* **11**, 311–338.

[13] Polat, E., Türkan, S. and Günay, S. (2009). Classification of the Banks in Turkey According to their Financial Performances Using Linear Discriminant Analysis, SIMCA and PLS Discriminant Analysis, in *Pls '09: Proceedings of the 6th International Conference On Partial Least Squares and Related Methods*, 156-163.

[14] Polat, E. and Gunay, S. (2015). The Comparison of Partial Least Squares Regression, Principal Component Regression and Ridge Regression with Multiple Linear Regression for Predicting PM10 Concentration Level Based on Meteorological Parameters, *Journal of Data Science* **13(2)**, 663-692.

[15] Rohman, A., Lumakso, F. A. and Riyanto, S. (2016). Use of Partial Least Square-Discriminant Analysis Combined with Mid Infrared Spectroscopy for Avocado Oil Authentication, *Research Journal of Medicinal Plants* **10(2)**, 175-180.

[16] Rosipal, R. and Krämer, N. (2006). *Overview and Recent Advances in Partial Least Squares*, In Saunders C., Grobelnik M., Gunn S., Shawe-Taylor J. (Eds.), Subspace, Latent Structure and Feature Selection Techniques Springer, 34-51.

[17] Trpkova, M. and Tevdovski, D. (2010). Applied Discriminant Analysis in Estimation of Potential EU Members, *Revista Tinerilor Economisti (The Young Economists Journal)* **1(15)**, 135-147.

[18] Wold, H. (1973). *Non-linear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments*. In Krishnaiah, P.R. (Ed.), Multivariate Analysis, Vol. III. Academic Press, New York, 383–407.

[19] Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses, *SIAM Journal on Scientific and Statistical Computing* **5(3)**, 735–743.

[20] Wold, S. et al. (2001). PLS-regression: a Basic Tool of Chemometrics, *Chemometrics and Intelligent Laboratory Systems* **58**, 109-130.

[21] XLSTAT, (2015). Copyright © 2015, Addinsoft, Paris, FRANCE. Available at http://drjackson.ca/applied_research_methods/xlstat_user_manual.pdf

[22] XLSTAT, (2017). Missing Data Imputation using NIPALS in Excel Tutorial. Available at https://help.xlstat.com/customer/en/portal/articles/2062415-missing-data-imputation-using-nipals-in-excel?b_id=9283 (accessed 01 March 2017).

[23] XLSTAT, (2017). Partial Least Squares Discriminant Analysis PLSDA Tutorial. Available at https://help.xlstat.com/customer/en/portal/articles/2062368-partial-least-squares-discriminant-analysis-plsda-tutorial?b_id=9283 (accessed 01 March 2017).

Esra Polat
Department of Statistics
Hacettepe University, Faculty of Science
Beytepe, 06800, Ankara, Turkey. espolat@hacettepe.edu.tr