



Comparison of PLSR and PCR techniques in terms of dimension reduction: an application on internal migration data in Turkey

Hatice Samkar*, Gamze Guven

Department of Statistics, Faculty of Arts and Sciences, Eskisehir Osmangazi University, Campus of Meselik, 26480 Eskisehir, Turkey

ARTICLE INFO

Article history:

Received 29 June 2016

Received in revised form

7 August 2016

Accepted 7 August 2016

Keywords:

Multicollinearity

dimension reduction

PLSR

PCR and RMSECV

ABSTRACT

Partial Least Squares Regression (PLSR) and Principle Component Regression (PCR) are dimension reduction techniques especially used in the presence of multicollinearity. In this study, these two techniques are described and their performance is compared in terms of dimension reduction. Root Mean Square Error of Cross Validation (RMSECV) is used as comparison criteria. PLSR and PCR techniques are applied on internal migration data in Turkey and it is found that PLSR technique is superior to PCR in terms of dimension reduction.

© 2016 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Multiple regression analysis is a technique to assess the functional relationship between a dependent variable and two or more independent variables. In multiple linear regression models, the method of Least Squares is widely used to estimate the regression coefficients. The least squares method produces estimators with desirable properties under certain assumptions (Chatterjee and Hadi, 2015). One of the assumptions is that independent variables are uncorrelated with each other (Gujarati, 2003). Linear or near-linear relationship between independent variables leads to multicollinearity problem. The presence of multicollinearity problem affects the signs of some regression coefficients in the model. Also, it results in large variances and covariance for the least-squares estimators of the regression coefficients, and thus the confidence intervals of coefficients and t statistics tend to be wider and smaller, respectively (Montgomery et al., 2001). While all or most of the regression coefficients are insignificant, the coefficient of determination is high and the model is significant.

The least squares estimators of the regression coefficients are the best linear unbiased estimator. Here "best" means giving the lowest variance of the estimate when compared to other unbiased, linear estimators. However, this feature is invalid in the

presence of multicollinearity. In this case, the multicollinearity should be removed. A lot of methods are proposed to remove multicollinearity. One of the methods is to use biased estimation techniques (Rawlings et al., 1998).

One of the biased estimation techniques which eliminate multicollinearity by reducing dimension is Partial Least Squares Regression (PLSR), and the other is Principle Component Regression (PCR). As these two techniques are applied to problems involving high collinearity in which the variance tends to dominate the biased, both techniques bring about similar results in many cases, and when used appropriately, the techniques produce better estimations than Least Squares (LS) technique (Fekedulegn et al., 2002; Frank and Friedman, 1993; Ziegel, 2004). In addition, these two techniques can not only be applied to data in which the number of observations is higher than that of independent variables but also be applicable to data in which the number of observations is lower than that of independent variables (Helland, 1988; Vigneau et al., 1997).

In literature, there are several studies comparing PLSR with PCR in terms of different criteria. Naes and Martens (1985) compared these two techniques in terms of Mean Squares Error (MSE) and found that PLSR used fewer latent variables than PCR. Luinge et al. (1993) showed that these two techniques were comparably similar in terms of prediction error on a real data set. Diaz et al. (1997) applied PLSR, PCR and LS on a real data set and used Root Mean Square Deviation (RMSD) and square of the correlation coefficient r^2 to evaluate the different techniques. Ni and Gong (1997) compared LS, PCR and PLSR techniques according to Relative Error of

* Corresponding Author.

Email Address: hfidan@ogu.edu.tr (H. Samkar)

<http://dx.doi.org/10.21833/ijaas.2016.08.002>

2313-626X/© 2016 The Authors. Published by IASE.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Prediction (REP) and did not observe significant differences in precisions of prediction between PLSR and PCR. Guiteras et al. (1998) applied LS, PCR and PLSR techniques on multivariate data and compared the predictions of the model in terms of the Relative Root Mean Squared Difference (RRMSD). To predict per capita gross domestic product of Turkey, Yeniay and Goktas (2002) used LS, Ridge Regression (RR), PLSR, and PCR techniques on a set of data gathered from 80 cities in Turkey and showed that PLSR and PCR techniques had the most predictive ability, respectively. Wentzell and Montoto (2003) made a simulation study on the complex chemical mixtures which contained a large number of components and reported that there was not any important difference in terms of prediction ability between PLSR and PCR techniques. Li (2010) made comparison of the prediction performances of PLSR, PCR and RR techniques according to Mean Square Error of Prediction (MSEP) and obtained similar results. Yaroshchik et al. (2012) compared PCR, PLSR, Multi-Block Partial Least Squares Regression (MB-PLSR) and Serial Partial Least Squares Regression (S-PLSR). They emphasized that PLSR and PCR models produced similar prediction accuracy, although in the case of PLSR there were notably less latent variables in use by the model. Khajehsharifi et al. (2014) compared the prediction ability of PLSR and PCR techniques with respect to Root Mean Square Error of Prediction (RMSEP) on a real data set and found that PLSR had more predictive ability than the other. Mahesh et al. (2015) compared the protein contents and hardness values predicted by PLSR and PCR models for bulk samples of Canadian wheat, and assessed prediction performances of regression models by calculating MSEP, Standard Error of Cross validation (SECV), and correlation coefficient (r).

The purpose of this study is to compare PLSR and PCR techniques with respect to Root Mean Square Error of Cross Validation (RMSECV) criterion in terms of dimension reduction on internal migration data in Turkey.

The study is organized as follows: PLSR and PCR techniques are briefly explained in Section 2; RMSECV criterion is described in Section 3; these techniques are applied on migration data in Turkey in Section 4; finally, Section 5 presents the conclusion.

2. PLSR and PCR techniques

Firstly, the matrix of X independent variables is standardized in PLSR and PCR techniques. Following this, new orthogonal variables are obtained by using the linear combination of these standardized variables. Lastly, to predict regression coefficients, LS method is applied to these variables (Vigneau et al., 1996).

The basic difference between PLSR and PCR techniques is the fact that while PLSR uses information on both dependent and independent variables, PCR uses only information on independent variables when component or latent variables are

obtained (Naes and Martens, 1985; Garthwaite, 1994).

PLSR and PCR techniques are explained in the following section.

2.1. PLSR technique

The aim of PLSR is to describe the structure between X and Y blocks and to predict Y block via X block (D'Ambra and Sarnacchiora, 2010). PLSR models the relationship between these two blocks via score vectors. PLSR decomposes zero mean variables X and zero mean variables Y as follows:

$$X = TP' + E \quad (1)$$

$$Y = UQ' + F \quad (2)$$

where T and U are matrices of score vectors (components, latent vectors); P and Q represent loading matrices; and E and F represent residual matrices (Rosipal and Krämer, 2006). This decomposition is done to maximize the covariance between T and U .

While forming 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 W^* .

$$(T = XW^*) \quad (3)$$

In PLSR, the weights are determined by maximizing the covariance between the latent variables and dependent variables (Zeng et al., 2007). In addition, T 's, the corresponding X scores, are good predictors of Y and can be given as follows:

$$(Y = TC' + F) \quad (4)$$

where C is the Y -weight matrix, and F is the Y -residual matrix and shows the deviation between the observed and modelled responses.

Finally, matrix B demonstrating the PLS-regression coefficients is obtained from following equation (Wold et al., 2001):

$$B = W'C' \quad (5)$$

PLSR technique can be used when the number of dependent variables is single or more than one (Garthwaite, 1994). However, when the dependent variable Y is single and $X'X'$ is diagonal, the PLSR arrives at the LS solution in one component, and PLSR and LS regression coefficients are equal (Wold et al., 2001).

2.2. PCR technique

PCR is a technique that deals with the multicollinearity problem by removing the nonstable structure of the model and by decreasing the variances of regression coefficients (Massy, 1965).

In PCR technique, firstly, the components are obtained with Principal Component Analysis (PCA). Following this, regression analysis is conducted by

using the principal component scores as independent variables (D'Ambra and Sarnacchiaro, 2010).

In order to demonstrate the PCR analysis, the singular value decomposition (SVD) of X is done as follows:

$$X = USP' \tag{6}$$

where U matrix represents linear transformation of X and S matrix is a diagonal matrix with elements equal to singular values. They are linked to the principal component score matrix T by $T = US$. The regression equation based on scores can be given as follows (Naes and Mevik 2001):

$$y = \alpha_0 + Ty + \varepsilon \tag{7}$$

Score vectors corresponding to small eigenvalues can be left out in order to avoid collinearity problems from influencing the solution (Geladi and Kowalski, 1986).

Let the matrix U_A be defined as the columns of U corresponding to the A largest eigenvalues of $X'X$. The PCR is then defined as the regression of y onto U_A .

$$y = \alpha_0 + U_A \alpha_A + f \tag{8}$$

where f is generally different from the error term ε above. The estimates of the α 's in α_A are found by LS. The PCR predictor \hat{y}_{PCR} is obtained as:

$$\hat{y}_{PCR} = \bar{y} + u'_A \hat{\alpha}_A \tag{9}$$

The value of u_A for a new sample is found by projecting x onto the A first principal components and by dividing the score/projection, t , by the square root of the eigenvalues. Note that for $A = p$, the PCR predictor becomes identical to the LS predictor \hat{y} . In practice, the best choice of A is usually determined by cross-validation (Naes and Mevik, 2001).

3. RMSECV

In this study, the RMSECV criterion is used to compare the PLSR and PCR techniques in terms of the dimension reduction ability on a real data set. Therefore, we focused on certain concepts associated with RMSECV and calculation of RMSECV.

Cross Validation is a generally applicable way to predict the performance of a model. In this study, the Leave-One-Out Cross Validation (LOOCV) technique is used. Leave-One-Out is the most classical Cross Validation procedure. In this procedure, each data point is successively "left out" from the sample and used for validation (Arlot and Celisse, 2010). In other words, the Leave-One-Out Cross Validation (LOOCV) uses a single observation from the whole sample as the validation data, and the remaining observations as the training data. This process is repeated until each observation in the entire sample is used once as the validation data (He et al., 2010).

Predicted Residual Sum of Squares (PRESS) statistics is calculated using the LOOCV method. The sum of squares of the difference between y_i and $\hat{y}_{(i)}$ is called PRESS (Allen, 1974):

$$PRESS = \sum_{i=1}^n \varepsilon_{(i)}^2 = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2 \tag{10}$$

where, $\hat{y}_{(i)}$ is the fitted value of the i th response based on all the observations except the i th one. PRESS is a generally regarded as a measure of how well a regression model will perform in predicting new data. A model with a small value of PRESS is desired (Montgomery et al., 2001).

RMSECV is a concept which is based on PRESS statistics and which is calculated in relation to the number of principle components or latent variables. RMSECV can be given as follows:

$$RMSECV = \sqrt{\frac{PRESS}{n}} \tag{11}$$

When different models are compared, the one with the smallest RMSECV is considered the best predictive model (Bodzioch et al., 2009).

4. The application of PLSR and PCR techniques on internal migration data in Turkey

In this section, internal migration data of Turkey from 2011 compiled by Turkish Statistical Institute (TSI) are used to compare PLSR and PCR techniques in terms of dimension reduction ability (TSI, 2011). The analysis is conducted with Matlab PLS_Toolbox. The data set contained 81 observations and 13 independent variables. The dependent and independent variables are shown in Table 1.

Table 1: The dependent and independent variables

Y	In-migration
X_1	Population
X_2	Unemployment ratio (%)
X_3	Number of Beds in Hospitals (100000 per person)
X_4	Number of Doctors
X_5	Number of Nurses
X_6	Electricity consumption per capita
X_7	Percentage of housing with piped water system
X_8	Infant Mortality Rate
X_9	Total Fertility Rate
X_{10}	Number of Motor Vehicles
X_{11}	Number of Students per teacher
X_{12}	Number of students per school
X_{13}	Annual income per capita(\$)

The data has been standardized after the logarithmic transformation has been applied, and then multiple regression analysis has been conducted. The results of regression analysis have revealed that about 92% of the variance in the dependent variable has been explained by the independent variables, and the model is significant at the level of 5%. Regression coefficients, Tolerance Values (TV), variance inflation factor (VIF) and t statistics can be seen in Table 2.

As can be seen in the table, the VIF of X_1, X_4, X_5 and X_{10} variables are greater than 10 and hence the TV of these variables are less than 0.1. Although the model is significant, all the regression coefficients except for X_1 are insignificant. Also, the maximum eigenvalue (λ_{max}) is 13.776; the minimum eigenvalue (λ_{min}) is 0.000029; and the condition number is about 6475034. In addition, the sum of

inverse of eigenvalues is greater than the number of independent variables.

Table 2: Regression coefficients, tolerance values (TV), variance inflation factors (VIF) and t statistics

Model	Unstandardized coefficients		Standardized coefficients	t statistics	p value	Multicollinearity statistics	
	Beta	Standard error	Beta			VIF	TV
Constant	0.417	1.535	—	0.272	0.787		
X ₁	0.770	0.309	0.887	2.495	0.015	107.731	0.009
X ₂	-0.249	0.134	-0.093	-1.853	0.068	2.135	0.468
X ₃	-0.016	0.157	-0.007	-0.104	0.918	4.083	0.245
X ₄	-0.067	0.228	-0.087	-0.292	0.771	74.773	0.013
X ₅	0.278	0.267	0.323	1.041	0.302	82.302	0.012
X ₆	0.008	0.088	0.006	0.088	0.930	3.573	0.280
X ₇	-0.264	0.521	-0.026	-0.508	0.613	2.171	0.461
X ₈	-0.104	0.135	-0.035	-0.769	0.445	1.760	0.568
X ₉	-0.362	0.277	-0.130	-1.309	0.195	8.388	0.119
X ₁₀	-0.181	0.105	-0.265	-1.729	0.088	19.955	0.050
X ₁₁	-0.066	0.341	-0.016	-0.194	0.847	5.503	0.182
X ₁₂	0.218	0.133	0.110	1.634	0.107	3.838	0.261
X ₁₃	0.060	0.104	0.040	0.582	0.562	3.929	0.255

Consequently, it could be stated that there is high multicollinearity between the independent variables. For this reason, PLSR and PCR from biased estimation techniques are applied to the data. Below are the RMSECV graphics associated with PLSR and PCR and the tables showing the percentage of the variance in the dependent and independent variables explained by latent variables obtained with these two techniques. As can be seen in Fig. 1, the optimal number of latent variables, which is determined by the minimum RMSECV value, is two.

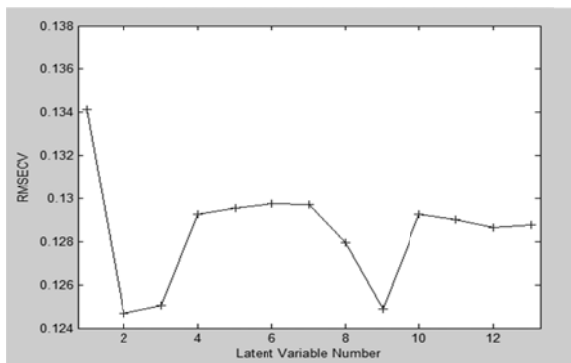


Fig. 1: RMSECV graphic for PLSR technique

When Table 3, which presents the percentage variance captured by PLSR, is examined, it is seen that the minimum RMSECV value is reached with two latent variables and that RMSECV value goes up when the number of latent variables increases. Thus, it is appropriate to use two latent variables. These two latent variables in PLSR could explain 89.41% of the variation in the dependent variable.

When PCR technique is used, in order to decide the number of components which will remain in the model, the RMSECV graphic corresponding to the number of component and the table presenting the percentage variance captured as in the PLSR technique should be examined. Fig. 2 and Table 4 show RMSECV graphic corresponding to the number of component and percentage variance captured by PCR technique, respectively.

As can be seen in Fig. 2 and Table 4, the optimal number of components, which is determined by the minimum RMSECV value, is seven. According to the table, while the number of components increases, related RMSECV values go up. Hence, it is appropriate to use seven components. These seven components in PCR could explain 90.75% of the variation in the dependent variable.

Table 3: Percentage variance captured using PLSR

The number of latent variable	X BLOCK		Y BLOCK		RMSECV
	Variance explanation rate	Total	Variance explanation rate	Total	
1	39.63	39.63	87.22	87.22	0.13412
2	30.69	70.32	2.19	89.42	0.12469
3	7.89	78.21	1.21	90.62	0.12502
4	3.71	81.92	0.49	91.11	0.12925
5	4.83	86.75	0.22	91.33	0.12956
6	2.45	89.21	0.41	91.74	0.12975
7	3.82	93.03	0.12	91.86	0.12972
8	2.17	95.20	0.04	91.91	0.12794
9	1.53	96.73	0.06	91.97	0.12489
10	1.19	97.92	0.06	92.03	0.12926
11	0.49	98.40	0.10	92.13	0.12903
12	1.53	99.93	0.00	92.13	0.12865
13	0.07	100.00	0.00	92.14	0.12879

According to the results of the analysis, while the minimum RMSECV value obtained with PLSR

technique is 0.12469, RMSECV value obtained with PCR is 0.12498. At the end of the analysis, although

almost the same RMSECV values are obtained in two techniques, PCR and PLSR reduce thirteen independent variables to seven components and two latent variables, respectively. Therefore, PLSR is superior to PCR in terms of dimension reduction.

5. Conclusion

PLSR and PCR techniques are applied in case of multicollinearity, and these biased estimation techniques remove multicollinearity problem by reducing data dimension. In this study, an application on internal migration data of Turkey was conducted to compare the ability of data dimension reduction in terms of RMSECV. At the end of the analysis, it was seen that the number of latent variables in PLSR is less than the number of components in PCR for almost the same RMSECV value.

Consequently, according to this study, PLSR technique is superior to PCR in terms of dimension reduction.

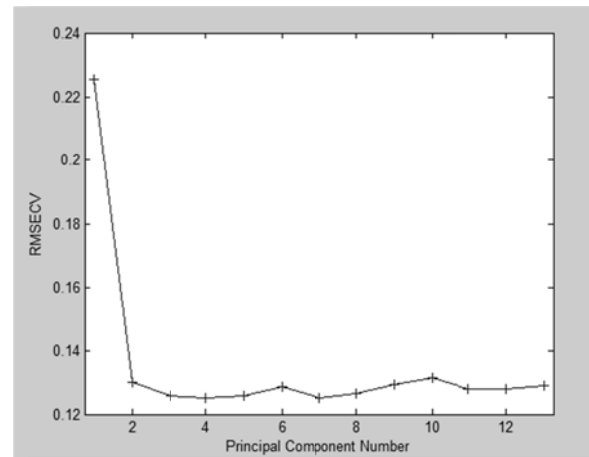


Fig. 2: RMSECV graphic for PCR technique

Table 4: Percentage Variance Captured Using PCR

The number of Component	X BLOCK		Y BLOCK		RMSECV
	Variance explanation rate	Total	Variance explanation rate	Total	
1	41.46	41.46	64.92	64.92	0.22554
2	29.95	71.40	23.06	87.98	0.13016
3	7.77	79.18	1.26	89.24	0.12595
4	6.47	85.65	0.39	89.63	0.12506
5	4.63	90.27	0.17	89.80	0.12587
6	3.01	93.28	0.07	89.87	0.12854
7	2.45	95.73	0.87	90.75	0.12498
8	1.66	97.39	0.01	90.75	0.12664
9	1.50	98.89	0.00	90.76	0.12914
10	0.67	99.56	0.08	90.83	0.13141
11	0.32	99.88	0.89	91.73	0.12789
12	0.07	99.95	0.17	91.90	0.12774
13	0.05	100.00	0.24	92.14	0.12879

Funding

The present study was supported with the project coded 2015-908 within the scope of Scientific Research Projects of Osmangazi University (ESOGU-BAP).

References

Allen DM (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1): 125-127.

Arlot S and Celisse A (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4: 40-79.

Bodzioch K, Bączek T, Kaliszan R and Vander HY (2009). The molecular descriptor logSum AA and its alternatives in QSRR models to predict the retention of peptides. *Journal of Pharmaceutical and Biomedical Analysis*, 50(4): 563-569.

Chatterjee S and Hadi AS (2015). *Regression analysis by example*. John Wiley & Sons. New Jersey, USA: 389.

D'Ambra A and Sarnacchiaro P (2010). Some data reduction methods to analyze the dependence with highly collinear variables: A simulation study. *Asian Journal Math Statistics*, 3(2): 69-81.

Diaz TG, Guiberteau A, Burguillos JO and Salinas F (1997). Comparison of chemometric methods: derivative ratio spectra and multivariate methods (CLS, PCR and PLS) for the resolution of ternary mixtures of the pesticides carbofuran carbaryl and phenamifos after their extraction into chloroform. *Analyst*, 122(6): 513-517.

Fekedulegn BD, Colbert JJ, RR Jr J and Schuckers ME (2002). *Coping with multicollinearity: An example on application of principal components regression in dendroecology*. Newton Square, PA: US. Department of Agriculture, Forest Service, Northeastern Research Station: 43.

Frank LE and Friedman JH (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2): 109-135.

Garthwaite PH (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425): 122-127.

- Geladi P and Kowalski BR (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185: 1-17.
- Guiteras J, Beltran JL and Ferrer R (1998). Quantitative multicomponent analysis of polycyclic aromatic hydrocarbons in water samples. *Analytica Chimica Acta*, 361(3): 233-240.
- Gujarati DN (2003). *Basic econometrics*. McGraw Hill, New York, USA.
- He G, Sentell T and Schillinger D (2010). A new public health tool for risk assessment of abnormal glucose levels. *Preventing Chronic Disease*, 7(2): 1-9.
- Helland IS (1988). On the structure of partial least squares regression. *Communications in Statistics-Simulation and Computation*, 17(2): 581-607.
- Khajehsharifi H, Pournasheer E, Tavallali H, Sarvi S and Sadeghi M (2014). The comparison of partial least squares and principal component regression in simultaneous spectrophotometric determination of ascorbic acid, dopamine and uric acid in real samples. *Arabian Journal of Chemistry*, DOI:10.1016/j.arabj.2014.02.006
- Li Y (2010). A Comparison Study of Principle Component Regression, Partial Least Square Regression and Ridge Regression with Application to FTIR Data. M.Sc Thesis in statistics, Faculty of Social Sciences, Uppsala University, Sweden.
- Luinge HJ, Hop E, Lutz ETG, Van Hemert JA and De Jong EAM (1993). Determination of the fat, protein and lactose content of milk using Fourier transform infrared spectrometry. *Analytica Chimica Acta*, 284(2): 419-433.
- Mahesh S, Jayas DS, Paliwal J and White NDG (2015). Comparison of partial least squares regression (PLSR) and principal components regression (PCR) methods for protein and hardness predictions using the near-infrared (NIR) hyperspectral images of bulk samples of Canadian wheat. *Food and Bioprocess Technology*, 8(1): 31-40.
- Massy WF (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309): 234-256.
- Montgomery DC, Peck EA and Vining GG (2001). *Introduction to linear regression analysis*. John Wiley & Sons. USA: 641.
- Naes T and Martens H (1985). Comparison of prediction methods for multicollinear data. *Communications in Statistics-Simulation and Computation*, 14(3): 545-576.
- Naes T and Mevik BH (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15(4): 413-426.
- Ni Y and Gong X (1997). Simultaneous spectrophotometric determination of mixtures of food colorants. *Analytica Chimica Acta*, 354(1): 163-171.
- Rawlings JO, Pantula SG and Dickey DA (1998). *Applied regression analysis: a research tool*. Springer Science & Business Media. New York, USA: 650
- Rosipal R and Krämer N (2006). Overview and recent advances in partial least squares. In C Saunders, M Grobelnik, S Gunn, and J Shawe-Taylor (Eds.), *Subspace, latent structure and feature selection*. Volume 3940 of the series *Lecture Notes in Computer Science*: 34-51. Springer Berlin Heidelberg.
- TSI (2011). İç göç istatistikleri, İllerin Aldığı Göç, Verdiği Göç, Net Göç ve Net Göç Hızı. Retrieved June 27, 2016 online available <http://www.tuik.gov.tr>.
- Vigneau E, Bertrand D and Qannari EM (1996). Application of latent root regression for calibration in near-infrared spectroscopy. Comparison with principal component regression and partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 35(2): 231-238.
- Vigneau E, Devaux MF, Qannari EM and Robert P (1997). Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of Chemometrics*, 11(3): 239-249.
- Wentzell PD and Montoto LV (2003). Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, 65(2): 257-279.
- Wold S, Sjöström M and Eriksson L (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2): 109-130.
- Yaroshchuk P, Death DL and Spencer SJ (2012). Comparison of principal components regression, partial least squares regression, multi-block partial least squares regression, and serial partial least squares regression algorithms for the analysis of Fe in iron ore using LIBS. *Journal of Analytical Atomic Spectrometry*, 27(1): 92-98.
- Yeniay O and Goktas A (2002). A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, 31(99): 99-101.
- Zeng XQ, Li GZ, Wu GF and Zou HX (2007). On the number of partial least squares components in dimension reduction for tumor classification. In T. Washio et al. (Eds.), *Emerging Technologies in Knowledge Discovery and Data Mining*. Volume

4819 of the series Lecture Notes in Computer Science: 206-217. Springer Berlin Heidelberg.

Ziegel ER (2004). A User-Friendly Guide to Multivariate Calibration and Classification. *Technometrics*, 46(1):108-110.