Journal of Data Science 14(2016), 739-768

The Comparison of Classical and Robust Biased Regression Methods for Determining Unemployment Rate in Turkey: Period of 1985-2012

Esra Polat[1], Semra Turkan[2]

[1,2]*Department of Statistics, Hacettepe University*

Abstract. Unemployment is one of the most important issues in macro economics. Unemployment creates many economic and social problems in the economy. The condition and qualification of labor force in a country show economical developments. In the light of these facts, a developing country should overcome the problem of unemployment. In this study, the performance of robust biased Robust Ridge Regression (RRR), Robust Principal Component Regression (RPCR) and RSIMPLS methods are compared with each other and their classical versions known as Ridge Regression (RR), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) in terms of predictive ability by using trimmed Root Mean Squared Error (TRMSE) statistic in case of both of multicollinearity and outliers existence in an unemployment data set of Turkey. Analysis results show that RRR model is chosen as the best model for determining unemployment rate in Turkey for the period of 1985-2012. Robust biased RRR method showed that the most important independent variable effecting the unemployment rate is Purchasing Power Parities (PPP). The least important variables effecting the unemployment rate are Import Growth Rate (IMP) and Export Growth Rate (EXP). Hence, any increment in PPP cause an important increment in unemployment rate, however, any increment in IMP causes an unimportant increase in unemployment rate. Any increment in EXP causes an unimportant decrease in unemployment rate.

Keywords: biased estimation, multicollinearity, outliers, partial least squares regression, principal component regression, ridge regression, robust biased estimation.

1. Introduction

In today's world, a country with unemployment that is resulted by the effects of economical and social effects comes across multidimensional problems. The condition and qualification of labor force in a country show economical developments. In the light of these facts, a developing country should overcome the problem of unemployment. According to Turkish Statistical Institute, active people in ages of between 15 and 60 that are labor force consist of non-institutionalization population. Unemployment is defined

as jobless who are looking for a job that offers the current fee level (Goktas and Isci, 2010).

In the literature, there are various factors that have effect on unemployment. Cascio (2001) investigates monetary policy and unemployment relationship for 11 OECD countries over 1979:Q1-1998:Q4 by using Vector Autoregressive (VAR) model. According to Cascio (2001), monetary shocks influence unemployment but they differ from country to country. Namely, local factor(s) is/are important how a labor market is influenced. Djivre and Ribon (2003) study monetary policy influence on unemployment, inflation and exchange over 1990-1999 for Israel and they found that tight monetary policy shocks increase unemployment. Ravn and Simonelli (2007) find that monetary policy shocks extracted influence on unemployment for the United States over 1953:Q3-2003:Q1. Karanasou and Sala (2010) investigate driving forces behind unemployment for Australia over time and find that reasons behind unemployment differ according to period investigated. For example, 1970's driving force behind unemployment is oil shock while in 1990s and 2000s; interest rate is important driving force. Yet, currently, the most influential factor is the tight foreign demand due to global crisis. Further, another study on unemployment by Valletta and Kuang (2010) shows that the recent increase in unemployment is conjectural rather than structural for the United States. In general, conjectural fluctuations, like fluctuation in exchange rate, international interest rate, and decline in foreign demand are the shocks that extract influence on unemployment (Dogan, 2012).

There are no many macro empirical studies on unemployment in Turkey. Further, in studies regarding Turkey's unemployment, structural breaks have not been considered so as they have not been introduced in VAR specification. For Turkey, Berument et al. (2006, 2009) and Berument (2008) investigates macroeconomic policy shocks on unemployment by using VAR models. The general conclusion derived from those studies is that positive income shocks reduce unemployment. Aktar and Ozturk (2009) study interaction among macroeconomic variables for Turkey and find that positive income shocks create statistically significant negative effect on unemployment. They, also, find that export is not statistically significant influence on unemployment. Dogrul and Soytas (2010) investigate relationships among unemployment, oil price and interest rate and find that interest rate shocks left long-term impact on unemployment even though initial impact on unemployment is negative and insignificant. The aim of this study is to examine the factors affecting the unemployment with Partial Least Squares (PLS) analysis in Turkey after the 2008 Global Financial Crisis. In their study macroeconomic

variables; industrial productivity index, real wage index, growth rate, consumer price index, the ratio of import to Gross National Product and the ratio of export to the Gross National Product are used in modelling the response variable; the rate of unemployment. Explanatory variables are taken for t time and eight term lags for 2005:Q1-2010:Q3. The results of the analysis show that same results are obtained for industrial productivity index and real wage index. The ratio of import to Gross National Product variable has a great contribution in modelling unemployment rate for all terms except first and second lags. Analysis results show that macroeconomic indicator: consumer price index has a significant contribution in all terms (Dogan, 2012). Goktas and Isci (2010) aimed to remove the collinearity on factors that affect the rate of unemployment and obtained the new variables from the factors via using the principal components. The new variables that are regressors are used in constructing of unemployment regression model. After they have checked the assumptions of statistical inference, they forecasted the unemployment in Turkey. Dogan (2012) investigates the response of unemployment to selective macroeconomics shocks for the period of 2000:Q1-2010:Q1. It finds that positive shocks to growth, growth in export and inflation reduce unemployment. On the other hand, shocks to exchange rate, interbank interest rate and money supply increase unemployment. The results are consistent with Phillips curve and Okun's Law suggestion. Namely, negative relationship between output and unemployment and positive relationship between unemployment and inflation are found. Umit and Bulut (2013) examine the factors affecting the unemployment with Partial Least Squares (PLS) analysis in Turkey after the 2008 Global Financial Crisis. In their study macroeconomic variables; industrial productivity index, real wage index, growth rate, consumer price index, the ratio of import to Gross National Product and the ratio of export to the Gross National Product are used in modelling the response variable; the rate of unemployment. Explanatory variables are taken for t time and eight term lags for 2005:Q1-2010:Q3. The results of the analysis show that same results are obtained for industrial productivity index and real wage index. The ratio of import to Gross National Product variable has a great contribution in modelling unemployment rate for all terms except first and second lags. Analysis results show that macroeconomic indicator: consumer price index has a significant contribution in all terms.

In several linear regression and prediction problems, the independent variables may be many and highly collinear. This phenomenon is called multicollinearity and it is known that in the case of multicollinearity the Ordinary Least Squares (OLS) estimator for the regression coefficients or predictor based on these estimates may give very poor results. Therefore, several biased estimation methods have been developed to overcome multicollinearity problem such as Ridge Regression (RR), Principal Component

Regression (PCR) and Partial Least Squares Regression (PLSR). The main goal of biased methods is to decrease the mean squared error of prediction by introducing a reasonable amount of bias into the model. In most real systems, exact collinearity of variables in **X** is rather unusual, because of the presence of random experimental noise. Nevertheless, in systems producing nearly collinear data, the solution for regression coefficient is highly unstable, such that very small interferences in the original data (for example, because of noise or experimental error) cause the method to produce madly different results. In addition, the use of highly collinear variables in Multiple Linear Regression (MLR) also increases the possibility of overfitting the model. Overfitting means that the model may fit the data very well, but fails when used to predict the properties of new samples (Martens and Naes, 1989; Naes et al., 2002, Polat and Gunay, 2015). In addition to multicollinearity, outliers are also the other problem encountered in regression models. If an observation does not follow the model that fits the majority of the data points, it is an outlier. Its occurrence can, e.g. be due to a measurement error, to a change in the experimental conditions or to the fact that the sample belongs to a population other than the one under study etc. It is also well known that the MLR method is highly sensitive to outliers. Both outliers in the space of the response variables and those in the space of the explanatory variables can unduly influence the parameter estimates (Hubert and Verboven, 2003). To solve this problem, the robust versions of regression models were proposed. In the case of the presence of both multicollinearity and outliers, the robust versions of the biased RR, PCR and PLSR methods called as Robust Ridge Regression (RRR), Robust Principal Component Regression (RPCR) and RSIMPLS are used.

In this study, for determining unemployment rate in Turkey, as the presence of simultaneously multicollinearity and outliers in the data set, performance of classical biased RR, PCR and PLSR methods and their robust versions RRR, RPCR and RSIMPLS are compared. Consequently, the method giving the best model is selected from among these six methods. The rest of the paper is organized as follows. The classical biased RR, PCR and PLSR methods are reviewed in Section 2. In Section 3, the robust versions of these three methods RRR, RPCR and RSIMPLS are described. In Section 4, a real unemployment data set of Turkey for the period of 1985-2012, in which both multicollinearity and outliers exist, analyzed by using these six methods and the performance of these methods are compared to each other by using trimmed Root Mean Squared Error (TRMSE) statistic.

2. Classical biased estimation methods

There are several classical biased estimation methods in the literature. In this study, the most commonly used three of them are used for analysis. These are RR, PCR and PLSR. In this section, first of all, MLR is briefly outlined in order to clarify the cause of using biased estimation methods in case of multicollinearity. Then, the most popular biased estimation methods RR, PCR and PLSR are presented. The emphasis here is on the algebraic derivation of the vector (or matrix) of coefficients in the linear regression models for these four methods. Throughout this paper, matrices are denoted by bold capital letters and vectors are denoted by bold lowercase letters.

2.1. Multiple Linear Regression

Firstly, the regression model used for this method is defined by Equation (1). Traditionally, the most frequently used method for finding $\boldsymbol{\beta}$ is the OLS. If $\mathbf{X}$ has a full rank of p (number of independent variables) then the OLS estimate of $\boldsymbol{\beta}$ given by Equation (2). $\hat{\boldsymbol{\beta}}_{OLS}$ gives unbiased estimates for the elements of $\boldsymbol{\beta}$. The corresponding vector of fitted values obtained as in Equation (3) (Phatak and De Jong, 1997).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{2}$$

$$\hat{\mathbf{y}}_{OLS} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}_x\mathbf{y} \tag{3}$$

In order to estimate $\boldsymbol{\beta}$, by using the OLS method, requires that the $\mathbf{X}$ variables must be linearly independent and the number of independent variables, p, must be equal or smaller than the number of observations, n $\left(p \le n\right)$ (Trygg, 2002).

If there is a multicollinearity problem, the variance of the least squares (LS) estimator may be very large and subsequent predictions rather inaccurate. However, if insisting on unbiased estimators given up, biased methods can be used to overcome the problem of inaccurate predictions. For this reason, biased methods such as RR, PCR and PLSR are used with the consequent trade-off between increased bias and decreased variance. The idea behind PCR and PLSR methods is to discard the irrelevant and unstable information and to use only the most relevant part of the x-variation for regression. Hence, the collinearity problem could be solved that more stable regression equations and predictions obtained (Naes et al., 2002; Polat and Gunay, 2015).

2.2. Ridge Regression

The $\hat{\boldsymbol{\beta}}_{OLS}$ estimator is an unbiased and has a minimum variance. However, when multicollinearity exists, the matrix $\mathbf{X}'\mathbf{X}$ becomes ill-conditioned (singular). Since $\text{Var}\left(\hat{\boldsymbol{\beta}}_{OLS}\right)=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ and the diagonal elements of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ become quite large, this makes the variance of $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ to be large. This leads to an unstable estimate of $\boldsymbol{\beta}$ and some of coefficients have wrong sign. In order to prevent these difficulties of OLS, Hoerl and Kennard (1970) suggested RR as an alternative procedure to the OLS method in regression analysis, especially, multicollinearity exists. RR is an estimation method when there is multicollinearity in the data. In practical terms, it consists of adding a biasing constant k to the diagonal elements of $\mathbf{X}'\mathbf{X}$ matrix, where $\mathbf{X}=(\mathbf{x}_1,...,\mathbf{x}_p)$ is a centered and standardized matrix. The resulting estimators are more stable than the least square (LS) estimators. RR estimator of $\boldsymbol{\beta}$ can be expressed as in Equation (4) (Hoerl and Kennard, 1970; Myers, 1990; Salh, 2014; Walker and Birch, 1988).

$$\hat{\boldsymbol{\beta}}_{RR}=(\mathbf{X}'\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \tag{4}$$

Here, k is ridge parameter, $\mathbf{I}$ is $p\times p$ identity matrix and $\mathbf{X}'\mathbf{X}$ is the correlation matrix of independent variables. Values of k lie in the range 0-1. Note that if k=0, the ridge estimator ($\hat{\boldsymbol{\beta}}_{RR}$) becomes as the OLS estimator ($\hat{\boldsymbol{\beta}}_{OLS}$) (Myers, 1990; Salh, 2014).

The trick in RR is to determine the optimum value of k for developing a predictive model. There are many procedures in the literature for determining the best value. Hoerl, Kennard and Baldwin (1975) suggested using the criterion given in Equation (5) in order to choose the optimum k value (Rawlings, 1988).

$$k=ps^2/\left[\hat{\boldsymbol{\beta}}_{OLS}(0)'\hat{\boldsymbol{\beta}}_{OLS}(0)\right] \tag{5}$$

Here, p is the number of regression vectors except the constant term (β_0), s^2 is the estimated residual mean of squares in OLS method. The denominator of Equation (5) shows the sum of squares of classic OLS regression parameters $\hat{\boldsymbol{\beta}}_{OLS}(0)$, which are calculated from centered and scaled independent variables and the constant term is

excluded in the calculation. The simplest way to determine the optimum value of k is to plot the values of each $\hat{\boldsymbol{\beta}}_{RR}$ versus k (in the range 0-1), which is called as ridge trace (Myers, 1990). In ridge trace graph, a trace or a curve is formed for each of the coefficients. From the ridge trace the minimum k value that makes the $\hat{\boldsymbol{\beta}}_{RR}$ stable is could be chosen and in this chosen k value the residual sum of squares could converge its minimum value (Hoerl and Kennard, 1970). However, Van Nostrand (1980) stated that since the determination of what is the stability in ridge trace is subjective and hence the selection of k is arbitrary, there is a tendency to choose a very big value of k while choosing k based on the ridge trace. Hence, the selection of k based on Equation (5) could be better (Rawlings, 1988).

2.3. Principle Component Regression

PCR and PLSR methods assume that the p-dimensional independent x-variables and a set of q-dimensional dependent y-variables are related through a bilinear model. n is the number of observations and for i=1,…,n this bilinear model is shown as in Equation (6) and Equation (7). Here $\bar{x}$ is the mean of x variables, $\bar{y}$ the mean of the y variables, $\tilde{t}_i$ are k dimensional scores with k<<p, $P_{p,k}$ the matrix of x-loadings and $A_{k,q}$ represents the slope matrix in the regression of y_i on $\tilde{t}_i$. The error terms are denoted by f_i and g_i. In terms of the original independent variables, this bilinear model can be written as in Equation (8). The main difference between PCR and PLSR lies in the construction of the scores $\tilde{t}_i$. In PCR the scores are obtained by extracting the most relevant information present in the x-variables by performing a PCA on the independent variables, thus, using a variance criterion. In contrast, the PLSR scores are calculated by maximizing a covariance criterion between the x- and y-variables (Hubert and Verboven, 2003; Hubert and Vanden Branden, 2003).

$$x_i = \bar{x} + P_{p,k}\tilde{t}_i + g_i \tag{6}$$

$$y_i = \bar{y} + A'_{q,k}\tilde{t}_i + f_i \tag{7}$$

$$y_i = \beta_0 + B'_{q,p}x_i + e_i \tag{8}$$

PCR starts by centering the data through the mean $\bar{x}$ of the x-variables and the mean $\bar{y}$ of the y-variables. Let the centered observations be denoted by (Hubert and Verboren, 2003)

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}} \tag{9}$$

$$\tilde{\mathbf{y}}_i = \mathbf{y}_i - \bar{\mathbf{y}} \tag{10}$$

Firstly, the first k principle components of $X_{n\times p}$ are computed to handle multicollinearity. Loading vectors $\tilde{\mathbf{P}}_{p,k} = (\mathbf{p}_1,...,\mathbf{p}_k)'$ are the k eigenvectors that correspond to the k largest eigenvalues of the emprical covariance matrix, $\mathbf{S}_x = \frac{1}{n-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ (Hubert and Verboren, 2003).

Next the k-dimesional scores of each data point, $\tilde{\mathbf{t}}_i$, are computed as the co-ordinates of the projections of $\tilde{\mathbf{x}}_i$ onto this subspace, or equivalently as in Equation (11).

$$\tilde{\mathbf{t}}_i = \tilde{\mathbf{P}}'\tilde{\mathbf{x}}_i \tag{11}$$

In the final step, $\tilde{y}_i$ is regressed onto $\tilde{\mathbf{t}}_i$ via multiple linear regression. Fitted linear model can be obtained as in below.

$$\tilde{\mathbf{y}}_i = \mathbf{A}'\tilde{\mathbf{t}}_i + \tilde{\boldsymbol{\varepsilon}}_i \tag{12}$$

The parameter estimates and fitted values can be expressed as in Equation (13) and Equation (14), respectively.

$$\hat{\mathbf{A}}_{k,q} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\tilde{\mathbf{y}} \tag{13}$$

$$\hat{\mathbf{y}}_i = \hat{\mathbf{A}}'\tilde{\mathbf{t}}_i + \bar{\mathbf{y}} \tag{14}$$

The unknown regression parameter in Equation (1) without a constant term is estimated as in Equation (15) (Hubert and Verboren, 2003).

$$\hat{\boldsymbol{\beta}}_{PCR} = \tilde{\mathbf{P}}\hat{\mathbf{A}} \tag{15}$$

2.4. Partial Least Square Regression

PLSR is a linear regression technique proposed to cope with high dimensional regressors and one or several response variables. There are several ways to calculate PLSR model parameters. One of the significant algorithms for PLSR, Straightforward Implementation of a Statistically Inspired Modification of the Partial Least Squares Method (SIMPLS), was proposed by Sijmen De Jong (1993). SIMPLS algorithm is the leading PLSR algorithm because of its speed and efficiency. Therefore, in this study, the PLSR with SIMPLS algorithm is introduced due to its speed and efficiency (Sijmen De Jong, 1993; Hubert and Vanden Branden, 2003). The SIMPLS algorithm assumes that the x and y variables are related with:

$$\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{P}\tilde{\mathbf{t}}_i + \mathbf{g}_i \tag{16}$$

$$\mathbf{y}_i = \bar{\mathbf{y}} + \mathbf{A}'\tilde{\mathbf{t}}_i + \mathbf{f}_i \tag{17}$$

where $\tilde{\mathbf{t}}_i$ are called scores, $\mathbf{P}$ is matrix of x loadings, $\mathbf{A}$ is the slope matrix in the regression of $\mathbf{y}_i$ on $\tilde{\mathbf{t}}_i$, $\mathbf{g}_i$ and $\mathbf{f}_i$ are residuals (Hubert and Vanden Branden, 2003).

3. Robust versions of bias estimation methods

Besides multicollinearity, $\hat{\boldsymbol{\beta}}_{OLS}$ are considerably affected by only or few observations. Namely, not all data points in a data set have the same significance in determining estimates, test and other statistics. It is important that the data analyst should be aware of such kind of points known as outliers. To overcome the effects of outliers, robust regression models are proposed. In the presence of both multicollineriaty and outliers, robust versions of bias estimation methods are proposed. The robust versions of RR, PCR and PLSR are given follow.

3.1. Robust Ridge Regression

RR is based on least squares and it is sensitive to atypical observations. Hence, the approach of MM estimation, which is repeated M estimation, is proposed by Maronna (2011) to ensure both robustness and efficiency under the normal model. The vector of residuals in RR is given by Equation (18). A scale M estimator of the data vector

$\mathbf{e^{RR}} = \left(e_1^{RR},\cdots,e_n^{RR}\right)'$ is defined as the solution $\hat{\sigma} = \hat{\sigma}(\mathbf{e^{RR}})$ of Equation (19) (Maronna, 2011):

$$\mathbf{e^{RR}} = \left(e_1^{RR},\cdots,e_n^{RR}\right)' = y - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{RR}} \tag{18}$$

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i^{RR}}{\hat{\sigma}}\right) = \delta \tag{19}$$

where ρ is a bounded ρ - function and $\delta \in (0,0.5)$ determines the breakdown point of $\hat{\sigma}$. Recall that the breakdown point (BDP) of an estimator is the maximum proportion of observations that can be arbitrarily altered with the estimator remaining bounded away from the border of the parameter set (which can be at infinity). When $\rho(t) = t^2$ and $\delta = 1$, the classical mean square error: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i^{RR})^2$ is obtained. Maronna (2011) employed for ρ the bisquare ρ-function given in Equation (20) (Maronna, 2011).

$$\rho_{bis}(t) = \min\left\{1, 1-\left(1-t^2\right)^3\right\} \tag{20}$$

Let $\hat{\boldsymbol{\beta}}_{ini}$ be an initial estimator. Let $\hat{\sigma}_{ini}$ be an M-scale estimator of the $\mathbf{e^{RR}} = \mathbf{e^{RR}}(\hat{\boldsymbol{\beta}}_{ini})$, as the solution of Equation (21). Here, $\hat{\boldsymbol{\beta}}_{ini}$ and $\hat{\sigma}_{ini}$ are initial estimators, ρ_0 is a bounded ρ-function and δ is to be chosen. Then the MM estimator for RR (in this study named as Robust Ridge Regression (RRR)) is defined by Equation (22), where ρ is another bounded ρ-function such that $\rho \le \rho_0$. The factor $\hat{\sigma}_{ini}^2$ before the summation is employed to make the estimator coincide with the classical one when $\rho(t) = t^2$ (Maronna, 2011).

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{e_i^{RR}}{\hat{\sigma}_{ini}}\right) = \delta \tag{21}$$

$$L(\mathbf{X},\mathbf{y},\boldsymbol{\beta}) = \hat{\sigma}_{ini}^2 \sum_{i=1}^{n}\rho\left(\frac{e_i^{RR}(\boldsymbol{\beta})}{\hat{\sigma}_{ini}}\right) + k\|\boldsymbol{\beta}\|^2 \tag{22}$$

Therefore, Equation (23) shall be used where the constants $c_0 < c$ are chosen to control both robustness and efficiency (Maronna, 2011).

$$\rho_0(t) = \rho_{bis}\left(\frac{t}{c_0}\right), \quad \rho(t) = \rho_{bis}\left(\frac{t}{c}\right) \tag{23}$$

It is known that the classical estimator RR satisfies the "normal equations". A similar system of equations is satisfied by RRR. Define (Maronna, 2011).

$$\psi(t) = \rho'(t), \quad W(t) = \frac{\psi(t)}{t} \tag{24}$$

Let

$$t_i = \frac{e_i^{RR}}{\hat{\sigma}_{ini}}, \; w_i = \frac{W(t_i)}{2}, \; \mathbf{w} = (w_1,....,w_n)', \; \mathbf{W} = \text{diag}(\mathbf{w}) \tag{25}$$

Setting the derivatives of Equation (22) with respect to $\boldsymbol{\beta}$ to zero yields for RRR Equation (26) and Equation (27). Since for the chosen ρ, $W(t)$ is a decreasing function of $|t|$ observations with larger residuals will receive lower weights w_i (Maronna, 2011).

$$\mathbf{W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{RR}) = 0 \tag{26}$$

$$(\mathbf{X}^T\mathbf{W}\mathbf{X} + k\mathbf{I})\hat{\boldsymbol{\beta}}_{RR} = \mathbf{X}^T\mathbf{W}\mathbf{y} \tag{27}$$

As is usual in robust statistics, these "weighted normal equations" suggest an iterative produce. Starting with an initial $\hat{\boldsymbol{\beta}}_{RR}$: Compute the residual vector $\mathbf{e}^{RR}$ and the weights **w**. In order to choose k, Cross-Validation (K), K-fold cross validation process, which requires recomputing the estimate K times. $\hat{y}_{-i}$ which is the fit of y_i computed without using the i-th observation is expressed as in Equation (28). It is the RRR estimate computed without observation i. Then a first-order Taylor approximation of the estimator yields the approximate prediction errors as shown in Equation (29), where

$h_i = \mathbf{x}_i^T (\sum_{i=1}^{n} \psi'(t_i)\mathbf{x}_i\mathbf{x}_i^T + 2k\mathbf{I})^{-1}\mathbf{x}_i$. See the study of Maronna (2011) for details related to choose k and constants (c or c_0) (Maronna, 2011).

$$\hat{y}_{-i} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\mathbf{RRR}}^{\mathbf{-i}} \tag{28}$$

$$e_{-i}^{RRR} = y_i - \hat{y}_{-i} \approx e_i^{RRR}\left(1 + \frac{W(t_i)h_i}{1 - h_i\psi'(t_i)}\right) \tag{29}$$

3.2. Robust Principal Component Regression

Outliers especially bad leverage points and vertical outliers are known to be very influential for the classical least squares (LS) regression fit, because they cause the slope to be tilted in order to accommodate the outliers. PCR combines PCA on the x-variables with LS regression. However, both stages yield very unreliable results when the data set contains outlying observations. Hence, Hubert and Verboven (2003) have been suggested a robust version of PCR method called as RPCR (Hubert and Verboven, 2003).

PCR starts by centring the data as $\tilde{\mathbf{x}}_i = \mathbf{x_i} - \bar{\mathbf{x}}$ and $\tilde{\mathbf{y}}_i = \mathbf{y_i} - \bar{\mathbf{y}}$, then, performing a PCA on the x-variables in order to cope with multicollinearity. The PCA loading matrix $\tilde{\mathbf{P}}_{k,p} = (\mathbf{p}_1, \cdots, \mathbf{p}_k)^T$ then contains the first *k* dominant eigenvectors of the empirical covariance matrix, $\mathbf{S}_x = \frac{1}{n-1}\tilde{\mathbf{X}}^{\mathbf{T}}{}_{p,n}\tilde{\mathbf{X}}_{n,p}$ and the scores satisfy $\tilde{\mathbf{t}}_i = \tilde{\mathbf{P}}_{\mathbf{k,p}}^{\mathbf{T}}\tilde{\mathbf{x}}_i$. In the second step of PCR, the dependent variables $\tilde{\mathbf{y}}_i$ are regressed onto $\tilde{\mathbf{t}}_i$ as $\tilde{\mathbf{y}}_i = \mathbf{A}^{\mathbf{T}}\tilde{\mathbf{t}}_i + \tilde{\boldsymbol{\varepsilon}}_i$ using MLR. Then, the parameter estimates and fitted values are obtained as $\hat{\mathbf{A}}_{k,q} = (\mathbf{T}'\mathbf{T})^{-1}_{k,k}\mathbf{T}^{\mathbf{T}}_{\mathbf{k,n}}\tilde{\mathbf{Y}}_{n,q}$ and $\hat{\mathbf{y}}_i = \hat{\mathbf{A}}^{\mathbf{T}}_{\mathbf{q,k}}\tilde{\mathbf{t}}_i + \bar{\mathbf{y}}$, respectively. The unknown regression parameters in Equation (1) are then estimated as $\hat{\boldsymbol{\beta}}_{PCR} = \tilde{\mathbf{P}}\hat{\mathbf{A}}$ (Hubert and Verboven, 2003).

In Hubert and Verboven (2003) a robust PCR method is proposed by robustifying both steps of PCR. Firstly, a robust PCA method is applied on the x-variables. Secondly, a robust regression method is applied. In first step, for low-dimensional data ($p<n/2$), the the highly robust MCD estimator is used as a robust estimator of the covariance matrix of the $\mathbf{x}_i$ and for high-dimensional data the ROBPCA method. To define the MCD estimator, we consider subsets of size *h* out of the whole data set (of size *n*). The MCD

estimator then seeks that *h* subset whose classical covariance matrix has minimal determinant. The number h determines the robustness of the estimator and should be at least (n+p+1)/2. The MCD location estimate is given by the mean $\bar{\mathbf{x}}_h$ and the MCD scatter estimator by its covarince matrix $\hat{\mathbf{\Sigma}}_h$, multiplied by a consistency factor. A robust PCA method yields a tolerance ellipse which captures the covariance structure of the majority of the data points. The robust tolerance ellipse is obtained by applying the highly robust MCD estimator of location and scatter to the data, yielding $\hat{\boldsymbol{\mu}}_{MCD}$ and $\hat{\mathbf{\Sigma}}_{MCD}$ and by plotting the points x whose robust distance $D(\mathbf{x}) = D\left(\mathbf{x}, \hat{\boldsymbol{\mu}}_{MCD}, \hat{\mathbf{\Sigma}}_{MCD}\right) = \sqrt{\left(\mathbf{x} - \hat{\boldsymbol{\mu}}_{MCD}\right)^T \hat{\mathbf{\Sigma}}_{MCD}^{-1} \left(\mathbf{x} - \hat{\boldsymbol{\mu}}_{MCD}\right)}$ is equal to $\sqrt{\chi^2_{2,0.975}}$. Based on the raw MCD estimate, a reweighting step can be added which increases the finite sample efficiency considerably. In that case, each data point receives a weight one if it belongs to the robust tolerance ellipse and a weight zero otherwise. The reweighted MCD estimator then equals the classical mean and covariance matrix of the data points with weight one. Furthermore, there is no distinction any more between the raw and the reweighted MCD estimator, as it is assumed that always the reweighted one is used. The first k eigenvectors of the MCD estimator, sorted in descending order of the eigenvalues, then yield robust loadings (Engelen et al., 2004; Hubert and Verboven, 2003).

ROBPCA is a robust PCA method which combines projection pursuit ideas with MCD covariance estimation in lower dimensions. Firstly, x data are preprocessed by reducing their data space to the affine subspace spanned by the *n* observations. This can be easily performed using a singular value decomposition of $\mathbf{X}_{n,p}$. In the second step of the ROBPCA algorithm, a measure of outlyingness is computed for each data point. This is obtained by projecting the high-dimensional data points on many univariate directions. On every direction a robust centre and scale of the projected data points is computed, and for every data point its standardized distance to that centre is measured. Finally, for each data point its largest distance over all the directions is considered. The *h* data points with smallest outlyingness are then retained and from the covariance matrix of this the final *h* subset, the number of principal components (PCs) to retain, *k*, is selected. ROBPCA yields more accurate estimates at uncontaminated data sets and more robust estimates at contaminated data sets. Briefly, ROBPCA method applied to $\mathbf{X}_{n,p}$ yields robust scores can be derived as $\mathbf{t}_i = \mathbf{P}_{\mathbf{k,p}}^{\mathbf{T}}\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x\right)$. Here, $\mathbf{P}_{p,k}$ is the loading matrix with orthogonal columns and $\hat{\boldsymbol{\mu}}_x$ is a robust centre (Engelen et al., 2004; Hubert and Verboven, 2003).

MCD scatter matrix defines a metric in the PCA subspace. Let L denote the diagonal matrix which contains the eigenvalues l_j of the MCD scatter matrix, sorted from largest to smallest. Then, the score distance of a p-dimensional point x with respect to $\hat{\boldsymbol{\mu}}_x$, P and L is defined as in Equation (30) (Hubert and Verboven, 2003).

$$\begin{aligned} SD(\mathbf{x}) &= SD(\mathbf{x},\hat{\boldsymbol{\mu}},\mathbf{P},\mathbf{L}) = D\left(\mathbf{t}=\mathbf{P}^{T}(\mathbf{x}-\hat{\boldsymbol{\mu}}_x),\mathbf{0},\mathbf{L}\right) \\ &= \sqrt{(\mathbf{x}-\hat{\boldsymbol{\mu}}_x)^{T}\mathbf{P}_{p,k}\mathbf{L}_{k,k}^{-1}\mathbf{P}_{k,p}^{T}(\mathbf{x}-\hat{\boldsymbol{\mu}}_x)} \\ &= \sqrt{\mathbf{t}^{T}\mathbf{L}^{-1}\mathbf{t}} = \sqrt{\sum_{j=1}^{k}\frac{t_j^2}{l_j}} \end{aligned} \tag{30}$$

In the second stage of RPCR method $\mathbf{y}_i$ is regressed on $\mathbf{t}_i$ that if there is only one y-variable the reweighted LTS regression is preferred, else the MCD regression is performed. Here, the regression model with with intercept written as in Equation (31) with $Cov(\breve{\varepsilon}) = \boldsymbol{\Sigma}_{\breve{\varepsilon}}$. In case of one dependent variable (q=1), this model simplifies as in Equation (32) with $\sigma_{\breve{\varepsilon}}$ scale of the errors. The parameters in (6) could be estimated by using the LTS estimator. The raw LTS estimator minimizes the sum of the h smallest squared residuals as shown in Equation (33). Here, $r_{1:n}^2 \le r_{2:n}^2 \le \cdots \le r_{n:n}^2$ denote the ordered squared residuals. An initial estimate of the error dispersion is given by Equation (34). Here c_h is a consistency factor for normally distributed errors. The reweighted LTS estimator then corresponds to the LS estimator applied to the observations whose absolute standardized residual is not too large. That means, if $\left|r_i(\hat{\boldsymbol{\alpha}},\hat{\alpha}_0)_{LTS}/\hat{\sigma}_0\right| > 2.5$ it is set $w_i = 0$ and otherwise, $w_i = 1$. Then, $(\hat{\boldsymbol{\alpha}},\hat{\alpha}_0)$ final estimates are obtained as the vector which minimizes $\sum_{i=1}^{n} w_i\left(y_i - \boldsymbol{\alpha}^{T}\mathbf{t}_i - \alpha_0\right)^2$ (Hubert and Verboven, 2003).

$$\mathbf{y}_i = \alpha_0 + \mathbf{A}^{T}\mathbf{t}_i + \breve{\varepsilon}_i \tag{31}$$

$$y_i = \alpha_0 + \boldsymbol{\alpha}^{T}\mathbf{t}_i + \breve{\varepsilon} \tag{32}$$

$$(\hat{\alpha},\hat{\alpha}_0)_{LTS} = \arg\min_{\alpha,\alpha_0}\sum_{i=1}^{h}\left(r^2(\alpha,\alpha_0)\right)_{i:n} \tag{33}$$

$$\hat{\sigma}_0 = c_h \sqrt{\frac{1}{h}\sum_{i=1}^{h}\left(r^2\left(\hat{\boldsymbol{\alpha}}, \hat{\alpha}_0\right)_{LTS}\right)_{i:n}} \tag{34}$$

In case of q>1, the MCD regression estimator is used. It starts by computing the reweighted MCD estimator on the $(\mathbf{t}_i, \mathbf{y}_i)$ jointly, leading to a (k+q)-dimensional location estimate $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{\mu}}_y)$ and a scatter estimate $\hat{\boldsymbol{\Sigma}}_{k+q,k+q}$, which can be split as shown in Equation (32). Similarly with the MLR estimates which are based on the emprical covariance matrix of the joint $(\mathbf{t}_i, \mathbf{y}_i)$ variables, robust parameter estimates are then estimated as shown in Equation (33). This robust regression estimator's efficiency can also be increased by performing a reweighting step. To apply this reweighting scheme, each data point receives a zero weight if it is initial residual distance is unsually large as shown in Equation (34), with Equation (35) and Equation (36). All other observations have a weight $w_i = 1$ (Hubert and Verboven, 2003). The reweighted MCD regression parameters then correspond to the MLR estimates based on those observations with weight one. The final residual distances are obtained by filling in the reweighted estimates for **A** and $\hat{\alpha}_0$ in Equation (33), Equation (35) and Equation (36). A different notation for the final estimates and residual distances is not introduced, but it is assumed that the reweighting step is indeed applied. The fitted values are obtained as in Equation (37) and regression parameters derived as in Equation (38). Finally, $\hat{\boldsymbol{\Sigma}}_{\varepsilon} = \hat{\boldsymbol{\Sigma}}_{\bar{\varepsilon}}$ is set (Hubert and Verboven, 2003).

$$\hat{\boldsymbol{\Sigma}}_{MCD} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_t & \hat{\boldsymbol{\Sigma}}_{ty} \\ \hat{\boldsymbol{\Sigma}}_{yt} & \hat{\boldsymbol{\Sigma}}_y \end{pmatrix} \tag{35}$$

$$\hat{\mathbf{A}}_{k,q} = \hat{\boldsymbol{\Sigma}}_t^{-1}\hat{\boldsymbol{\Sigma}}_{ty} \qquad \hat{\alpha}_0 = \hat{\boldsymbol{\mu}}_y - \hat{\mathbf{A}}^{\mathbf{T}}\hat{\boldsymbol{\mu}}_t \qquad \hat{\boldsymbol{\Sigma}}_{\bar{\varepsilon}} = \hat{\boldsymbol{\Sigma}}_y - \hat{\mathbf{A}}^{\mathbf{T}}\hat{\boldsymbol{\Sigma}}_{\mathbf{t}}\hat{\mathbf{A}} \tag{36}$$

$$w_i = 0 \text{ if } RD_i > \sqrt{\chi^2_{q,0.975}} \tag{37}$$

$$\mathbf{r}_i = \mathbf{y}_i - \hat{\mathbf{A}}^{\mathbf{T}}\mathbf{t}_i - \hat{\alpha}_0 \tag{38}$$

$$RD_i = D\left(\mathbf{r}_i, \mathbf{0}, \hat{\boldsymbol{\Sigma}}_{\bar{\varepsilon}}\right) = \sqrt{\mathbf{r}_i^{\mathbf{T}}\hat{\boldsymbol{\Sigma}}_{\bar{\varepsilon}}^{-1}\mathbf{r}_i} \tag{39}$$

$$\begin{aligned}\hat{\mathbf{y}}_i &= \hat{\mathbf{A}}_{\mathbf{q,k}}^{\mathbf{T}}\mathbf{t}_i + \hat{\alpha}_0 \\ &= \hat{\mathbf{A}}_{\mathbf{q,k}}^{\mathbf{T}}\mathbf{P}_{\mathbf{k,p}}^{\mathbf{T}}\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_x\right) + \hat{\alpha}_0\end{aligned} \tag{40}$$

$$\hat{\mathbf{B}}_{p,q} = \mathbf{P}_{p,k}\hat{\mathbf{A}}_{k,q} \qquad \hat{\beta}_0 = \hat{\alpha}_0 - \hat{\mathbf{B}}_{p,q}\hat{\boldsymbol{\mu}}_x \tag{41}$$

As for the MCD estimator, the robustness of the RPCR algorithm depends on the value of h, which is chosen in the ROBPCA algorithm, LTS regression and MCD regression. In MATLAB implementation the user can either choose h from the start or choose a value of α with $0.5 \le \alpha \le 1$, such that $1-\alpha$ corresponds with the percentage of outliers that the algorithm should be able to resist. Then, h is set as the maximum of $h_1 = [\alpha n]$ and $h_2 = \left[\frac{n+k+q+1}{2}\right]$ where h_2 is the required minimal value for the MCD regression estimator to have a positive breakdown value. When a large proportion of contamination is presumed, α should be chosen close to 0.5. Otherwise an intermediate value for α, such as 0.75, is recommended because it increases the finite sample efficiency. Therefore, the default choice is $\alpha = 0.75$ (Engelen et al., 2004; Hubert and Verboven, 2003).

3.3. Robust Partial Least Squares Regression

SIMPLS algorithm is the leading PLSR method because of its speed and efficiency. It assumes that the x and y variables are related through a bilinear model as given in (1) and (2). This bilinear structure implies a two-step algorithm. $\tilde{\mathbf{X}} = \{(\mathbf{x}_i - \bar{\mathbf{x}})\}_{i=1}^{n}$ and $\tilde{\mathbf{Y}} = \{(\mathbf{y}_i - \bar{\mathbf{y}})\}_{i=1}^{n}$ be the centered data matrices. After mean centering the data, SIMPLS will first construct k latent variables (LVs) $\tilde{\mathbf{T}}_{\mathbf{n,k}}^{T} = \left(\tilde{\mathbf{t}}_1, \cdots, \tilde{\mathbf{t}}_n\right)'$ and then the dependent variables will be regressed on these k variables. k components, the columns of $\tilde{\mathbf{T}}_{n,k}$, are obtained as a linear combination of the x-variables which has maximum covariance with a certain linear combination of the y-variables. In order to obtain k components, firstly, it is required to calculate weight vectors. Hence, the first normalized PLSR weight vectors $\mathbf{r}_1$ and $\mathbf{q}_1$ are obtained as linear combinations of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ that maximizes $\text{cov}\left(\tilde{\mathbf{Y}}_{n,q}\mathbf{q}_1, \tilde{\mathbf{X}}_{n,p}\mathbf{r}_1\right)$. The solution of this maximization problem is found by taking $\mathbf{r}_1$ and $\mathbf{q}_1$ are obtained as the first left and right singular eigenvectors of

$\mathbf{S}_{\mathbf{yx}}^{\mathbf{T}} = \mathbf{S}_{xy} = \tilde{\mathbf{X}}_{\mathbf{p,n}}^{\mathbf{T}} \tilde{\mathbf{Y}}_{n,q} / (n-1)$, the cross-covariance matrix of the x- and y-variables. For each observation the first coordinate of the score $\tilde{\mathbf{t}}_{\mathbf{i}}$ is computed as $\tilde{\mathbf{t}}_{i1} = \tilde{\mathbf{x}}_{\mathbf{i}}^{\mathrm{T}} \mathbf{r}_1$. To obtain the other PLSR weight vectors r_a and q_a for $a = 2, \cdots, k$, the components $\tilde{\mathbf{X}}\mathbf{r}_j$ are required to be orthogonal. Hence, if we require that $\sum_{i=1}^{n} \mathbf{t}_{\mathbf{ia}} \mathbf{t}_{\mathbf{ib}} = 0$ and a≠b, a deflation of the cross-covariance matrix $\mathbf{S}_{xy}$ provides the solutions for the other PLSR weight vectors. This deflation is carried out by first calculating the x-loading $\mathbf{p}_j = \mathbf{S}_x \mathbf{r}_j / (\mathbf{r}_\mathbf{j}^{\mathrm{T}} \mathbf{S}_\mathbf{x} \mathbf{r}_\mathbf{j})$ with $\mathbf{S}_x$ empirical covariance matrix of the *x*-variables. Next an orthonormal base $\{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_a\}$ of $\{\mathbf{p}_1, \ldots, \mathbf{p}_a\}$ is constructed and $\mathbf{S}_{xy}$ is deflated as $\mathbf{S}_{xy}^{a} = \mathbf{S}_{xy}^{a-1} - \boldsymbol{\nu}_a (\boldsymbol{\nu}_\mathbf{a}^{\mathbf{T}} \mathbf{S}_{xy}^{a-1})$ with $\mathbf{S}_{xy}^{1} = \mathbf{S}_{xy}$. In general, PLSR weight vectors $\mathbf{r}_a$ and $\mathbf{q}_a$ are obtained as the left and right singular vectors of $\mathbf{S}_{xy}^{a}$. The elements of the scores $\tilde{\mathbf{t}}_i$ are then defined as linear combinations of the mean-centered data, $\tilde{\mathbf{t}}_{ia} = \tilde{\mathbf{x}}_\mathbf{i}^{\mathrm{T}} \mathbf{r}_\mathbf{a}$ or equivalently $\tilde{\mathbf{T}}_{n,k} = \tilde{\mathbf{X}}_{n,p} \mathbf{R}_{p,k}$ with $\mathbf{R}_{p,k} = (\mathbf{r}_1, \cdots, \mathbf{r}_k)$. Finally, where the scores are k-dimensional, MLR is performed of the dependent variables $\mathbf{y}_i$ on these scores $\tilde{\mathbf{t}}_i$. Thus, the formal regression model under consideration in Equation (42), where $E(f_i) = 0$ and $Cov(f_i) = \boldsymbol{\Sigma}_f$. MLR provides estimates as in Equation (43), Equation (44) and Equation (45). By inserting $\tilde{\mathbf{t}}_i = \mathbf{R}_{\mathbf{k,p}}^{\mathbf{T}} (\mathbf{x}_i - \bar{\mathbf{x}})$ in Equation (42), estimates for the parameters in the original model in Equation (46) are obtained as in Equation (47). Finally, also an estimate of $\boldsymbol{\Sigma}_e$ is provided by rewriting $\mathbf{S}_f$ in terms of the original parameters: $\mathbf{S}_e = \mathbf{S}_y - \hat{\mathbf{B}}^{\mathbf{T}} \mathbf{S}_\mathbf{x} \hat{\mathbf{B}}$ (Engelen et al., 2004; Hubert and Vanden Branden, 2003).

$$\mathbf{y}_i = \alpha_0 + \mathbf{A}'_{q,k} \tilde{\mathbf{t}}_i + f_i \tag{42}$$

$$\hat{\mathbf{A}}_{k,q} = (\mathbf{S}_t)^{-1} \mathbf{S}_{ty} = (\mathbf{R}'_{\mathbf{k,p}} \mathbf{S}_\mathbf{x} \mathbf{R}_{p,k})^{-1} \mathbf{R}'_{\mathbf{k,p}} \mathbf{S}_{\mathbf{xy}} \tag{43}$$

$$\hat{\alpha}_0 = \bar{\mathbf{y}} - \hat{\mathbf{A}}'_{q,k} \bar{\bar{\mathbf{t}}} \tag{44}$$

$$\mathbf{S}_f = \mathbf{S}_y - \hat{\mathbf{A}}'_{\mathbf{q,k}} \mathbf{S}_\mathbf{t} \hat{\mathbf{A}}_{\mathbf{k,q}} = \mathbf{Y}'_{\mathbf{q,n}} \mathbf{Y}_{n,q} - \hat{\mathbf{A}}'_{\mathbf{q,k}} \mathbf{T}'_{\mathbf{k,n}} \mathbf{T}_{\mathbf{n,k}} \hat{\mathbf{A}}_{\mathbf{k,q}} \tag{45}$$

$$\mathbf{y}_i = \beta_0 + \mathbf{B}_{\mathbf{q,p}}^{\mathbf{T}} \mathbf{x}_i + e_i \tag{46}$$

$$\hat{\mathbf{B}}_{p,q} = \mathbf{R}_{p,k}\hat{\mathbf{A}}_{k,q} \text{ and } \hat{\boldsymbol{\beta}}_0 = \bar{\mathbf{y}} - \hat{\mathbf{B}}_{\mathbf{q,p}}^{\mathbf{T}}\bar{\mathbf{x}} \quad (47)$$

Since SIMPLS is based on the empirical cross-covariance matrix between the y-variables and the x-variables and on linear LS regression, the results are also affected by outliers in the data set. Hence, Hubert and Vanden Branden (2003) have been suggested a robust version of SIMPLS method called as RSIMPLS (Hubert and Vanden Branden, 2003). A robust method RSIMPLS starts by applying ROBPCA on the *x*- and *y*-variables in order to replace $\mathbf{S}_{xy}$ and $\mathbf{S}_x$, which are used to calculate $\tilde{\mathbf{t}}_i$, by robust estimates and then proceeds analogously to the SIMPLS algorithm. Similar to RPCR instead of MLR a robust regression method (ROBPCA regression) is performed in the second stage (Engelen et al., 2004; Hubert and Vanden Branden, 2003). To obtain robust scores, firstly, ROBPCA is applied on $\mathbf{Z}_{n,m} = \left(\mathbf{X}_{n,p}, \mathbf{Y}_{n,q}\right)$. ROBPCA is robust covariance estimator for high-dimensional data sets (m>n). ROBPCA combines the two approaches. Using projection pursuit ideas, it computes the outlyingness of every data point and then considers the empirical covariance matrix of the h data points with smallest outlyingness. The data are then projected onto the subspace K_0 spanned by the $k_0 << m$ dominant eigenvectors of this covariance matrix. Next the MCD method is applied to estimate the center and scatter of the data in this low dimensional subspace. Finally these estimates are back transformed to the original space and a robust estimate of the center $\hat{\boldsymbol{\mu}}_z$ of $\mathbf{Z}_{n,m}$ and of its scatter $\hat{\boldsymbol{\Sigma}}_z$ are obtained. This scatter matrix can be decomposed as $\hat{\boldsymbol{\Sigma}}_z = \mathbf{P}^z\mathbf{L}^z\left(\mathbf{P}^z\right)^T$ with robust Z-eigenvectors $\mathbf{P}^z_{m,k_0}$ and Z-eigenvalues $\text{diag}\left(\mathbf{L}_{k_0,k_0}\right)$. Note that the diagonal matrix $\mathbf{L}^z$ contains the k_0 largest eigenvalues of $\hat{\boldsymbol{\Sigma}}_z$ in decreasing order. Then **Z**-scores $\mathbf{T}^z$ can be obtained from $\mathbf{T}^z = \left(\mathbf{Z} - \mathbf{1}_n\hat{\boldsymbol{\mu}}_Z^T\right)\mathbf{P}^z$. After application of ROBPCA on $\mathbf{Z}_{n,m}$, this yields robust estimates $\hat{\boldsymbol{\mu}}_z = \left(\hat{\boldsymbol{\mu}}_x^T, \hat{\boldsymbol{\mu}}_y^T\right)^T$ and $\hat{\boldsymbol{\Sigma}}_z$. $\hat{\boldsymbol{\Sigma}}_z$ can be split as in Equation (48). The cross-covariance matrix $\boldsymbol{\Sigma}_{xy}$ is estimated by $\hat{\boldsymbol{\Sigma}}_{xy}$ and the PLS weight vectors $\mathbf{r}_a$ are computed as in the SIMPLS algorithm, but now starting with $\hat{\boldsymbol{\Sigma}}_{xy}$ instead of $\mathbf{S}_{xy}$. The x-loadings are defined as $\mathbf{p}_j = \left(\mathbf{r}_j^T\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}\mathbf{r}_j\right)^{-1}\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}\mathbf{r}_j$. Then the deflation of the scatter matrix $\hat{\boldsymbol{\Sigma}}_{xy}^a$ is performed as in SIMPLS. In each step the robust scores are calculated as in Equation (49), where the $\breve{\mathbf{x}}_i$ are the robustly centered observations (Engelen et al., 2004).

$$\hat{\Sigma}_z = \begin{pmatrix} \hat{\Sigma}_x & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_y \end{pmatrix} \tag{48}$$

$$t_{ia} = \breve{x}_i^T r_a = \left(x_i - \hat{\mu}_x\right)^T r_a \tag{49}$$

Once the robust scores are derived, a robust linear regression is performed. The regression model based on robust scores is written as in Equation (50). In order to estimate parameters in this model a robust regression method called ROBPCA regression is used. This method uses additional information from the previous ROBPCA step (Engelen et al., 2004).

$$\mathbf{y}_i = \alpha_0 + \mathbf{A}_{\mathbf{q,k}}^{\mathbf{T}} \mathbf{t}_i + \breve{\mathbf{f}}_i \tag{50}$$

Firstly, to obtain the robust scores t_i ROBPCA has been applied to the (x,y)-variables and a k_0-dimensional subspace K_0, which represented these (x,y)-variables well, has been obtained. Because the scores were then constructed to summarize the most important information given in the x-variables, it is expected that outliers with respect to this k_0-dimensional subspace are often also outlying in the (t, y)-space. Hence, the center $\boldsymbol{\mu}$ and scatter $\boldsymbol{\Sigma}$ of the (t,y)-variables is estimated as the weighted mean and covariance matrix of those $(\mathbf{t}_i, \mathbf{y}_i)$ whose corresponding $(\mathbf{x}_i, \mathbf{y}_i)$ are not outlying to K_0 as shown in Equation (51). If observation *i* is not identified as an outlier by applying ROBPCA on (x, y) then $w_i = 1$ and otherwise $w_i = 0$.

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\boldsymbol{\mu}}_t \\ \hat{\boldsymbol{\mu}}_y \end{pmatrix} = \sum_{i=1}^{n} w_i \begin{pmatrix} \mathbf{t}_i \\ \mathbf{y}_i \end{pmatrix} / \sum_{i=1}^{n} w_i \qquad \hat{\Sigma} = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_t & \hat{\boldsymbol{\Sigma}}_{ty} \\ \hat{\boldsymbol{\Sigma}}_{yt} & \hat{\boldsymbol{\Sigma}}_y \end{pmatrix} = \sum_{i=1}^{n} w_i \begin{pmatrix} \mathbf{t}_i \\ \mathbf{y}_i \end{pmatrix} \begin{pmatrix} t_i^T & \mathbf{y}_i^T \end{pmatrix} / \left(\sum_{i=1}^{n} w_i - 1 \right) \tag{51}$$

Two types of outliers: those which are outlying within K_0 and those which are lying far from K_0 are could be identified by applying ROBPCA The first type of outliers can be easily identified as those observations whose robust distance $D_{i(k_0)} = \sqrt{\left(\mathbf{t}_i^{\mathbf{z}}\right)^T \left(L^z\right)^{-1} \mathbf{t}_i^z}$ exceeds $\sqrt{\chi^2_{k_0,0.975}}$. To determine the second type of outliers, for each data point its orthogonal distance $OD_i > \sqrt{\hat{\mu}_{od^2} + \hat{\sigma}_{od^2} z_{0.975}}$ to the subspace K_0 is considered. The distribution of these orthogonal distances is difficult to determine exactly, however, motivated by the central limit theorem, it appears that the squared

orthogonal distances are roughly normally distributed. Hence, their center and variance is estimated with the univariate MCD, yielding $\hat{\mu}_{od^2}$ and $\hat{\sigma}^2_{od^2}$. Then if $OD_i > \sqrt{\hat{\boldsymbol{\mu}}_{od^2} + \hat{\boldsymbol{\sigma}}_{od^2} z_{0.975}}$, w_i is set to zero. Having identified the observations with weight one, thus, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are computed from (51). Then these estimates are inserted in Equations (43)-(45) and a reweighted MLR is performed. It is recommended to reweight these initial regression estimates in order to improve the finite sample efficiency. Let $r_{i(k)}$ be the residual of the i_{th} observation based on the initial estimates that were calculated with k components. If $\hat{\boldsymbol{\Sigma}}_f$ is the initial estimate for the covariance matrix of the errors, then the robust distance of the residuals is defined as in Equation (52). The weights $c_{i(k)}$ are computed as in Equation (53) with I the indicator function. The final regression estimates are then calculated as in classical MLR, but only based on those observations with weight $c_{i(k)}$ equal to one. This reweighting step has the advantage that it might again include observations with $w_i = 0$ which are not regression outliers. The robust residual distances $RD_{i(k)}$ are recomputed as in Equation (52) and also the weights $c_{i(k)}$ are adapted. Robust parameters for the original model in Equation (46) are then given by Equation (54).

$$RD_{i(k)} = \left(r_{i(k)}^T \hat{\boldsymbol{\Sigma}}_{\bar{f}}^{-1} r_{i(k)}\right)^{1/2} \tag{52}$$

$$c_{i(k)} = I\left(RD_{i(k)}^2 \le \chi^2_{q,0.975}\right) \tag{53}$$

$$\hat{\beta}_0 = \hat{\alpha}_0 - \hat{\mathbf{B}}\hat{\boldsymbol{\mu}}_x \qquad \hat{\mathbf{B}}_{p,q} = \mathbf{R}_{\mathbf{p,k}}\hat{\mathbf{A}}_{k,q} \qquad \hat{\boldsymbol{\Sigma}}_e = \hat{\boldsymbol{\Sigma}}_{\bar{f}} \tag{54}$$

4. Application on a real unemployment data set of Turkey

In this study, the annual data set of Turkey (1985-2012), including the eleven factors affecting the unemployment rate, is analyzed by using RR, PCR, PLSR, RRR, RPCR, RSIMPLS methods. The unemployment rate (UR) is considered as dependent variable. The variables that explain the unemployment are determined by following the previous studies on this issue. These independent variables are Gross Domestic Product Growth Rate with Expenditure Approach (GDP), Import Growth Rate (IMP), Export Growth Rate (EXP), Population Growth Rate (P), Exchange Rate (E), Consumer Price Index (CPI), Relative Consumer Price Indices (RCPI), Civilization of Employment

Agriculture (CEA), Unit Labor Cost (ULC), Purchasing Power Parities (PPP), Trend (T). The data set is taken from OECD data base.

Firstly, MLR analysis applied on the data set and it is found that the model obtained by using OLS method is significant with a probability of 95% (F=4.95; p=0.002). Even though the MLR model fits the data well, multicollinearity may severely prohibit quality of the prediction. Table 1 shows that all independent variables are not significant as an indicator of multicollinearity problem. Since $\lambda_1,\cdots,\lambda_p$ are the eigenvalues of correlation matrix of $X'X$, the existence of collinearity problem also could be seen by examining the condition number that calculated as $\lambda_{max}/\lambda_{min}$ =9.112/0.0003=30373. The condition number greater than 30 means that there is multicollinearity. The other multicollinearity measure is Variance Inflation factor (VIF). The larger the VIF value, the more serious the collinearity problem. In practice, if any of the VIF values is equal or larger than 10, there is a near-collinearity. In this case, the regression coefficients are not reliable. The VIF values of GDP, IMP, EXP, P, E, CPI, RCPI, CEA, ULC, PPP and T are found as 3.32, 9.79, 6.05, 6.79, 39.25, 32.99, 22.25, 67.86, 2.98, 157.42 and 51.57, respectively. Since the VIF values of E, CPI, RCPI, CEA, PPP and T are found greater than 10, there is a near-collinearity problem for this data set.

Table 1: The estimated regression coefficients for the MLR model

Model	Coefficients	Standart Error of Coefficients	T	P
Constant	24.866	9.464	2.63	0.018
Trend	-0.3588	0.1838	-1.95	0.069
GDP	-0.11589	0.08178	-1.42	0.176
EXP	-0.01865	0.01123	-1.66	0.116
IMP	0.00542	0.01735	0.31	0.759
P	1.120	1.787	0.63	0.540
E	-2.001	1.867	-1.07	0.300
PPP	12.982	6.365	2.04	0.058
CPI	0.02521	0.03897	0.65	0.527
RCPI	-0.11798	0.06699	-1.76	0.097
CEA	-0.2399	0.1841	-1.30	0.211
ULC	0.03096	0.01681	1.84	0.084

Secondly, whether outliers exist or not is examined using normal Q-Q plot of the MLR residuals given in Figure 1. As seen from Figure 1, there are two outliers in the data.

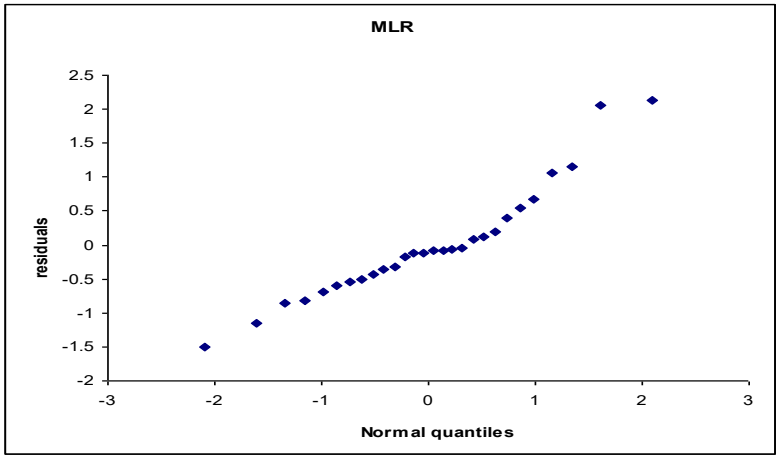
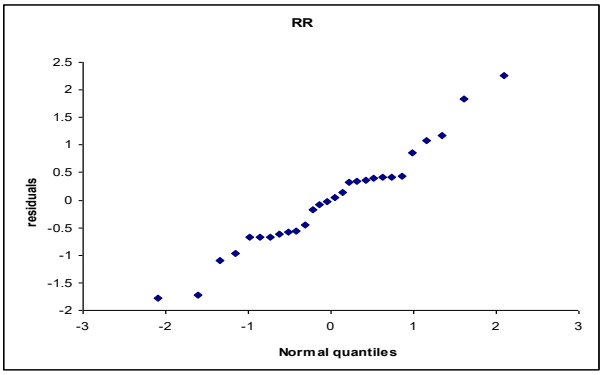
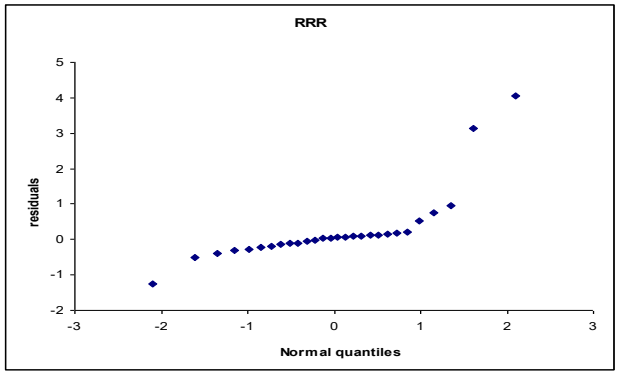
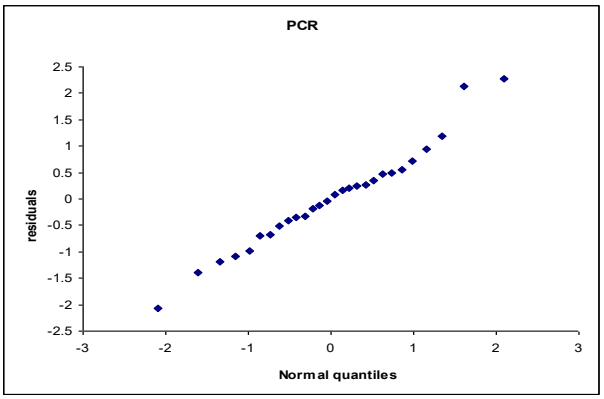


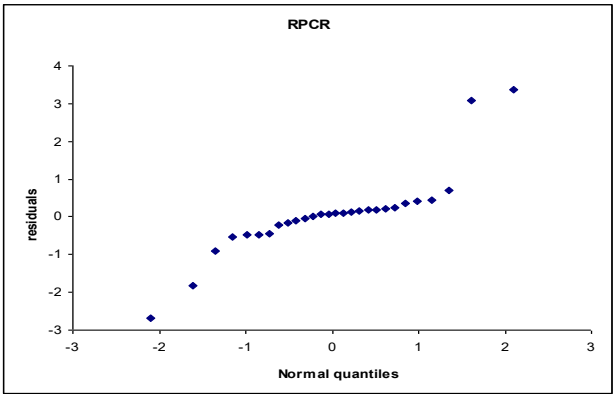
Figure 1: Q-Q plot for MLR residuals

It is found that the data set simultaneously includes multicollinearity and outliers. Hence, RR, PCR, PLSR, RRR, RPCR and RSIMPLS methods are applied on the data set. Firstly, the Q-Q plots for residuals of methods are obtained in Figure 2. As seen from Figure 2, there are two outliers as to all methods. In addition, classical RR, PCR and PLSR methods are relatively much more dispersed than the ones corresponding to the RRR, RPCR and RSIMPLS.


RR
residuals
Normal quantiles
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-3
-2
-1
0
1
2
3


RRR
residuals
Normal quantiles
5
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3


PCR
residuals
Normal quantiles
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-2
-1
0
1
2
3


RPCR
residuals
Normal quantiles
4
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

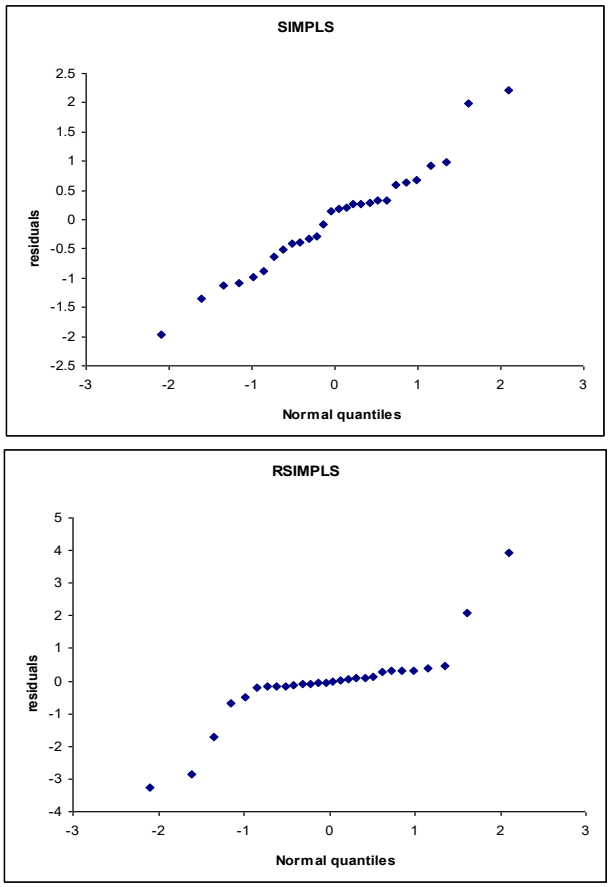


Figure 2: Residuals Q-Q Plots of the Classical and Robust Biased Regression Methods

The predictive performance of the methods are evaluated by using the Root Mean Square Error (RMSE), $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$) with upper %10 trimming which is considered to be safer in the presence of outliers. The exclusion of a certain percentage of unusually large residuals leads to an acceptable robust performance criterion. The regression coefficients and TRMSE values of methods are given in Table 2.

Table 2: The estimated regression coefficients and TRMSE values of methods

	Regression Coefficients					
Variables	RR	PCR	PLSR	RRR	RPCR	RSIMPLS
Constant	17.9265	25.6056	28.2426	23.2473	27.1417	13.4399
T	-.0976	-.1611	-.2322	-.4540	-.3524	-.4302
GDP	-.0819	-.0929	-.0895	-.0430	-.1003	-.0937

IMP	-.0007	.0004	.0003	.0048	.0162	.0246
EXP	-.0076	-.0094	-.0136	-.0073	-.0054	.0017
P	.1890	-.0903	1.2939	-1.1617	-.2164	-.1439
E	-.2168	.6063	1.0358	-1.0185	2.3267	3.2841
PPP	3.1610	.1285	1.1919	11.9718	.5134	.9214
CPI	-.0153	-.0204	.0001	.0189	-.0167	-.0077
RCPI	-.0526	-.0644	-.0913	-.0829	-.0680	.0482
CEA	-.1119	-.2333	-.3283	-.1669	-.2870	-.1663
RULC	.0130	.0142	.0212	.0272	.0256	.0090
Biasing Parameters	$k_{RR} = 0.5$			$k_{RR} = 11$		
TRMSE(0.9)	0.6935	0.6640	0.6575	0.3240	0.5059	0.5984

Since there is both multicollinearity and outliers in the data set, it is expected that robust biased regression methods give more accurate results than their classical versions. As seen from Table 2, the smallest TRMSE values are obtained using robust RRR and RPCR methods, respectively. According to TRMSE values, robust methods RRR, RPCR and RSIMPLS are clearly better than their classical counterparts RR, PCR and PLSR.

As mentioned in Maronna (2011) this version of robust ridge regression which we called here briefly RRR estimator is also robust when p/n is large, and is resistant to both bad leverage points and vertical outliers. In Maronna (2011) a simulation study for the case n>p in which the proposed estimator RRR is seen to outperform its competitors (other robust ridge estimators proposed before in literature). Moreover, in his study he described a simulation for p>n in which the RRR estimator is seen to outperform RSIMPLS method. In many applications in literature, as seen from this study too, PCR and PLSR methods (and their robust counterparts) give close results. This comparative study and their study show that RRR is a good alternative to the most popular robust biased methods, namely RPCR and RSIMPLS methods. All of the three robust methods RRR, PCR and PLSR could be used for both n>p or p>n situations and resist to different types of outliers (bad leverage points and vertical outliers). Hence, generally, we could mention that RRR is a good alternative but we could not declare that it always give better results than RPCR and RSIMPLS as in many real data applications the situation could be differ. However, in this study by applying on a real data set we showed that RPCR and RSIMPLS are not the only methods to be preferred in case of multicollinearity and outlier existence. Thus, especially for fields such as chemometrics, in which these kinds of data sets (including multicollinearity and outliers) are seen usually with the case of p>>n, RRR could be used as a good alternative to RPCR and RSIMPLS methods and the best model for the purpose (fitting to data set or prediction) could be selected.

5. Conclusion

In this study, for determining unemployment rate in Turkey during 1985-2012 period, it is aimed to choose the best model by comparing classical biased RR, PCR and PLSR methods and robust biased RRR, RPCR and RSIMPLS methods in case of both multicollinearity and outliers existence. For the unemployment data set, RRR model is chosen as the best model according to TRMSE(0.9) criteria. In addition, robust methods RRR, RPCR and RSIMPLS outperform classical RR, PCR and PLSR methods in terms of predictive ability. The results obtained from RRR robust biased regression method showed that the most important independent variable effecting the unemployment rate is Purchasing Power Parities (PPP). The least important variables effecting the unemployment rate are Import Growth Rate (IMP) and Export Growth Rate (EXP), respectively. Hence, any increment in PPP cause an important increment in unemployment rate, however, any increment in IMP causes an unimportant increase in unemployment rate. Any increment in EXP causes an unimportant decrease in unemployment rate.

References

[1] Aktar, I. and Ozturk, L. (2009). Can Unemployment be Cured by Economic Growth and Foreign Direct Investment in Turkey. *International Research Journal of Finance and Economics* **27**, 203-211.

[2] Aktas, C. (2007). Çoklu Bağıntı ve Liu Kestiricisiyle Enflasyon Modeli için Bir Uygulama. *ZKÜ Sosyal Bilimler Dergisi*, **Cilt 3**, **Sayı 6**, 67–79.

[3] Aqil, M., Qureshi, M. A., Ahmed, R. R. and Qadeer, S. (2014). Determinants of Unemployment in Pakistan. *International Journal of Physical and Social Sciences* **4:4**.

[4] Berument, M. H., Dogan, N. and Tansel, A. (2006). Economic Performance and Unemployment: Evidence from and Emerging Economy. *International Journal of Manpower* **27(7),** 604-623.

[5] Berument, M. H., Dogan, N. and Tansel, A. (2009). Macroeconomic Policy and Unemployment by Economic Activity: Evidence from Turkey. *Emerging Markets Finance and Trade* **45(3)**, 21-34.

[6] Bilgin, H. (2004). Döviz Kuru ve İşsizlik İlişkisi: Türkiye Üzerine Bir İnceleme. *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* **8(2)**, 1-15.

[7] Cascio, I. L. (2001). *Do Labour Markets Really Matter? Monetary Shocks and Asymetric Effects across Europe.* Unpublished, Department of Economics, University of Essex, Colchester.

[8] De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression, *Chemometrics Intell. Lab. Syst*. **18**, 251–263.

[9] Djıvre, J and Ribon, S. (2003). Inflation, Unemployment. The Exhange Rate, and Monetary Policy In Israel, 1990-1999: A SVAR Approach. *Israel Economic Review* **2**, 1-29.

[10] Dogan, T. T. (2012). Macroeconomic Variables and Unemployment: The Case of Turkey. *International Journal of Economics and Financial Issues* **2(1)**, 71-78.

[11] Dogrul, H. G. and Soytas, U. (2010). Relationship between Oil Prices, Interest Rate, and Unemployment: Evidence from an Emerging Market. *Energy Economics* **32**, 1523-1528.

[12] Engelen, S., Hubert, M., Vanden Branden, K. and Verboven, S. (2004). Robust PCR and Robust PLSR: a comparative study. *Theory and Applications of Recent Robust Methods, Statistics for Industry and Technology,* 105-117.

[13] Goktas, A. and Isci, O. (2010). Turkiye'de İssizlik Oraninin Temel Bilesenli Regresyon Analizi ile Belirlenmesi. *Sosyal ve Ekonomik Arastirmalar Dergisi-Selcuk Universitesi* **14:20**, 279-294.

[14] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12:1**, 55-67.

[15] Hubert, M. and Verboven, S. (2003). A robust PCR method for high-dimensional regressors. *Journal of Chemometrics* **17**, 438–452.

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

[17] Karanassou, M. and Sala, H. (2010). Labour Market Dynamics in Australia: What Drives Unemployment? *The Economic Record, The Economic Society of Australia*, **vol. 86 (273)**, 185-209.

[18] Maronna, R. A. (2011). Robust ridge regression for high-dimensional data, *Technometrics* **53(1)**, 44-53.

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

[20] Mayers, R. H. (1990). *Classical and modern regression with applications*, 2nd edition, Duxbury Press.

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

[22] Phatak, A. and De Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics* **11**, 311–338.

[23] 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 PM_{10} Concentration Level Based on Meteorological Parameters. *Journal of Data Science* **13**, 663-692.

[24] Ravn, M. O. and Simonelli, S. (2007). *Labor Market Dynamics and the Business Cycle: Structural Evidence for the United States*. Working Paper Series, no 182, Centre for Studies in Economics and Finance.

[25] Rawlings, J. O. (1988). *Applied Regression Analysis: A Research Tool*, Pacific Grove, California: Wadsworth & Brooks/Cole Advanced Books & Software.

[26] Salh, S. M. (2014). Using Ridge Regression model to solving multicollinearity problem. *International Journal of Scientific & Engineering Research* **5:10**, 992-998.

[27] Trygg, J. (2002). Have you ever wondered why PLS sometimes needs more than one component for a single-y vector? *Chemometrics Homepage*, February 2002, http://www.chemometrics.se/editorial/feb2002.html.

[28] Umit, A. O. and Bulut, E. (2013). Türkiye'de İşsizliği Etkileyen Faktörlerin Kısmi En Küçük Kareler Regresyon Yöntemi İle Analizi: 2005-2010 Dönemi, *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi* **37**, 131-142.

[29] Valletta, R. and Kuang, K. (2010). Is the Structural Unemployment on the Rise? *The Economic Report* **86(273)**, 185-209.

[30] Verboven, S. and Hubert, M. (2005). LIBRA: a MATLAB Library for Robust Analysis. *Chemometrics and Intelligent Laboratory Systems* **75**, 127–136.

[31] Walker, E., and Birch, J. B. (1988). Influence Measures in Ridge Regression. *Technometrics* **30**, 221-227.

[32] Yılmaz, Ö. (2005). Türkiye Ekonomisinde Büyüme ile İşsizlik Oranları Arasındaki Nedensellik İlişkisi, *İstanbul Üniversitesi İktisat Fakültesi Ekonometri ve İstatistik Dergisi* **S.2**: 11-29.

[33] Data sources: http://www.mod.gov.tr/en/SitePages/mod_easi.aspx and http://www.yaklasim.com/BookList.aspx?AnnouncementId=2780&AnnouncementCategoryId=8&_

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

Semra Turkan
Department of Statistics
Hacettepe University, Faculty of Science
Beytepe, 06800, Ankara, Turkey.
sturkan@hacettepe.edu.tr