

A COMPARISON STUDY OF PRINCIPLE
COMPONENT REGRESSION, PARTIAL
LEAST SQUARES REGRESSION AND RIDGE
REGRESSION WITH APPLICATION TO
FTIR DATA

Author: Ying Li

Supervisor: Dietrich von Rosen



UPPSALA
UNIVERSITET

Master thesis in statistics
Faculty of Social Sciences
Uppsala University, Sweden
June, 2010

Abstract

Least squares estimator may fail when the number of explanatory variable is relatively large in comparison to the sample or if the variables are almost collinear. In such a situation, principle component regression, partial least squares regression and ridge regression are often proposed methods and widely used in many practical data analysis, especially in chemometrics. They provide biased coefficient estimators with the relatively smaller variation than the variance of the least squares estimator. In this paper, a brief literature review of PCR, PLS and RR is made from a theoretical perspective. Moreover, a data set is used, in order to examine their performance on prediction. The conclusion is that for prediction PCR, PLS and RR provide similar results. It requires substantial verification for any claims as to the superiority of any of the three biased regression methods.

CONTENTS	3
----------	---

Contents

1 Introduction	1
2 Data Description	1
3 Methodology	3
3.1 Singular value decomposition	4
3.2 Principal component regression	4
3.3 Partial least squares regression	6
3.4 Ridge regression	8
3.5 Links and comparisons	9
4 Data Analysis	10
5 A Simulation Study	14
5.1 Performance in prediction	14
5.2 Performance in estimation	15
6 Conclusion	17
References	18

1. Introduction

Least squares multiple regression with a single dependent variable has been successfully applied to a variety of scientific fields. This can be attributed to the Gauss-Markov theorem, which states that the least squares estimator is the best linear unbiased estimator (BLUE). The best estimator is based on assumptions $E(\boldsymbol{\varepsilon}_i) = 0$, $V(\boldsymbol{\varepsilon}_i) = \sigma^2$ and $cov(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) = 0$. Even if in the cases when the assumptions are not fulfilled, the least squares estimator is "best" among the unbiased and linear estimators. It is well known that there exist estimators which are either biased or non-linear or both, which have an average performance far better than the least squares estimator.

The least squares estimator also fails when the number of explanatory variables is relatively large in comparison to the sample size. The explanatory variables are often near collinear. A typical data set is, in spectroscopy, where the intention is to predict the concentrations of the analyte, as the response variable, from the wavelength values, as the explanatory variables. Usually, the number of wavelengths, can be up to several hundred, exceeding the number of samples, and there exists collinearity among the wavelength values.

Principle component regression, partial least squares regression and ridge regression are the most popular regression methods that have been applied to collinear data. There also exists a large number of other regression methods that have been proposed for near collinear data: Latent root regression (Hawkins, 1973), various methods suggested by calibration theory or by Bayesian theory, intermediate least squares (Frank, 1987), variable selection methods, James-Stein shrinkage (James and Stein, 1961), etc.

The purpose of the present paper is to make a brief literature review of PCR, PLS and RR. Then a data set is used in comparing the three methods from a prediction point of view.

2. Data Description

Grass and mixtures of grasses and legumes were harvested for experimental purpose at the Kungsängen Research Center, Department of Animal Nutri-

tion and Management, Swedish University of Agricultural Sciences in Uppsala during 2002-2006. The researchers are interested in the concentration of lactate (La) in the grass. High-performance of liquid chromatography (HPLC) was used for analysis of the concentration of lactate in the grass. Meanwhile, the same grass samples were also analyzed by Fourier transform infrared (FTIR) analysis. By FTIR analysis, each grass sample got a corresponding spectra, as shown in Figure 1.

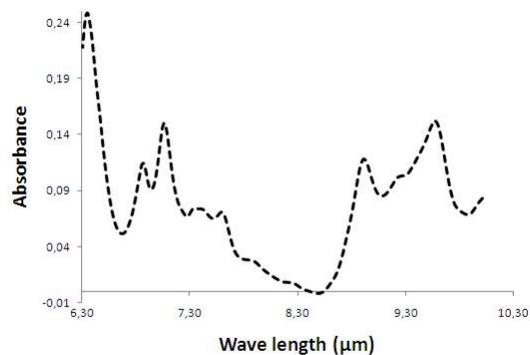


Figure 1: The spectra of a grass sample analyzed by FTIR

Our purpose is to find the relation between the concentration of lactate (\mathbf{y}) and the absorbance values of the wavelengths (\mathbf{X}). A total of 84 samples with complete absorbance values and concentrations of lactate >0.1 g/L were selected for this study. The wavelength regions of 4.71 to 4.80, 5.40 to 5.90 and 6.3 to 10 μm , comprising a total of 206 wavelengths were carefully selected by the researchers. They believe that these wavelength regions contains the information of the concentration of the analyte. Each μm of wavelength corresponded to a certain PIN where:

$$\mu m = 10000 / (3.858 \times PIN)$$

3. Methodology

The main purpose here is to find the predictive relationships of the response variable \mathbf{y} and the random explanatory variables \mathbf{X} given n observations, where \mathbf{y} is an n vector $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ and \mathbf{X} is an $n \times p$ matrix $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$ with $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})^T$.

For the linear model, the conditional expected value of the single response, $E(\mathbf{y} | \mathbf{X})$, is

$$E(\mathbf{y} | \mathbf{X}) = \boldsymbol{\alpha} + \boldsymbol{\beta}^T \mathbf{X}. \quad (1)$$

An additive error $\boldsymbol{\varepsilon}$, assumed to be uncorrelated with the \mathbf{X} variables and having variance σ^2 , is added to the conditional expectation in equation (1). The model turns into:

$$\mathbf{y} = E(\mathbf{y} | \mathbf{X}) + \boldsymbol{\varepsilon}.$$

The expectation $E(\mathbf{y} | \mathbf{x})$ is often simply replaced by $E(\mathbf{y})$, because the model does not always refer to random explanatory variables treated conditionally. Moreover, the model can also be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}^T)^T$, and \mathbf{X} is $n \times (p+1)$ matrix including a constant column for the intercept term $\boldsymbol{\alpha}$.

Using ordinary least squares(OLS) as the method to estimate $\boldsymbol{\theta}$ is to minimize the sum of squares

$$\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2$$

to form, via the normal equation, the estimator

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \quad (3)$$

where it is assumed that \mathbf{X} is of full rank.

In most statistical chemometric analysis, it is desirable to center the variables. The $\mathbf{X}^T \mathbf{X}$ matrix contains the variance-covariance data if data are mean centered (Zeaiter and Rutledge 2009). A detailed description of different centering methods has been given by Bro and Smilde (2003). Here we use:

$$\mathbf{y} = \mathbf{y} - \bar{\mathbf{y}},$$

$$\mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}_i.$$

All of the following equations are written under the mean centering assumption. Thus, there is no intercept required in the models.

3.1 Singular value decomposition

The singular value decomposition of \mathbf{X} $n \times p$ is written as $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. $\mathbf{D} = \text{Diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$ is a $r \times r$ ($r = \text{rank}(\mathbf{X})$) diagonal matrix with non-zero singular values, which equals the square roots of the eigenvalues of both $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$. The columns of \mathbf{U} are orthogonal, which are the eigenvectors of $\mathbf{X}\mathbf{X}^T$ and the columns of \mathbf{V} are orthogonal, which represents eigenvectors of $\mathbf{X}^T\mathbf{X}$ with corresponding the eigenvalues in \mathbf{D} .

3.2 Principal component regression

Principal component regression (PCR) starts with the estimated covariance matrix $\mathbf{X}^T\mathbf{X}$ of explanatory variables. Because of collinearity among the explanatory variables, the $\mathbf{X}^T\mathbf{X}$ matrix is almost singular. Instead of using all eigenvalues of $\mathbf{X}^T\mathbf{X}$ as OLS, PCR tries to extract a suitable number (d) of the largest eigenvalues.

$$\mathbf{X}_d = \mathbf{U}_d\mathbf{D}_d\mathbf{V}_d^T$$

Hence the least squares criteria equals

$$\|\mathbf{X}_d\boldsymbol{\theta} - \mathbf{y}\|_2^2$$

The corresponding estimate of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_{PCR} = (\mathbf{X}_d^T\mathbf{X}_d)^{-1}\mathbf{X}_d^T\mathbf{y} = \mathbf{V}_d\mathbf{D}_d^{-1}\mathbf{U}_d^T\mathbf{y} \quad (4)$$

Furthermore, equation (4) can be expressed as:

$$\hat{\boldsymbol{\theta}}_{PCR} = \mathbf{V}_d\mathbf{D}_d^{-1}\mathbf{U}_d^T\mathbf{y} = \sum_{i=1}^d \left(\frac{\mathbf{u}_i^T\mathbf{y}}{\lambda_i}\right)\mathbf{v}_i$$

where $\frac{\mathbf{u}_i^T\mathbf{y}}{\lambda_i}$ denotes *loading*, a scalar value showing how much a particular eigenvector is weighted in forming the regression vector. The d orthogonal

eigenvectors \mathbf{v}_i of $\mathbf{X}^T\mathbf{X}$ provide d new canonical variables, *the principal components*, as linear combination of the original variables.

To properly use principal component regression it is natural to look at some data-dependent choices of the number of components d . Cross-validation (Stone, 1974) is one of the most popular techniques in this field, especially in chemometrics. The observed data is divided into complementary subsets. Choosing the number of extracted factors, one subset (the training set) is used to fit model, and the other subset (the test set) is offer to measure how well the model fit. In cross-validation, the observed data is divided into different complementary subset for the limits of sample size. The number of factor chosen is usually the one which minimizes the root mean of the predicted residual sum of squares (PRESS, Allen (1971)):

$$e_i = y_i - \hat{y}_i$$

$$PRESS = \sum_{i=1}^n (e_i)^2.$$

However, often the PRESS of the model with fewer components is only marginally larger than the absolute minimum. Van der Voet (1994) proposed a simple randomization t-test to compare the predicted residuals from different models. The number of components chosen, by applying Voet's test, is the fewest with residual that are insignificantly larger than the residuals of the model with absolute minimum PRESS. Several authors: Wold (1978), Eastment and Krzanowski (1983), Wallach and Goffinet (1987), Osten (1989), et al., have also proposed other approaches to compare the predicted residuals.

Let \hat{y}_{Ai} and \hat{y}_{Bi} be predictions of y_i from two competing models, A and B, respectively. The Voet's test procedures as follows (Helland 2001):

$$e_{Ai} = y_i - \hat{y}_{Ai}, e_{Bi} = y_i - \hat{y}_{Bi}$$

1. Calculate $d_i = e_{Ai}^2 - e_{Bi}^2$, $i = 1, \dots, n$
2. Compute

$$MSEP = \frac{1}{n} PRESS$$

$$T = MSEP_A - MESP_B = \bar{d} = \frac{1}{n} \sum d_i$$

for the sample data (T_{obs})

3. Iterate steps given below m times:
 - Attach random signs to d_i , $i = 1, \dots, n$
 - Calculate $T = \bar{d}$
4. Calculate the significance level $p = \frac{k}{m+1}$, where m is the number of randomization trials, and k is the rank of T_{obs} among the randomization values of T when ranked from high to low. For practical use, a significance level 0.05 is often used for interpretation, $m = 19$ is the minimum number of randomization trials needed, which our choice here, but $m = 99$ or $m = 199$ is a more reasonable choice permitting p values down to $p = 0.01$ or 0.005 .

There are many other approaches for determining d , for example: proportion to the total variance, sizeable eigenvalues, Mallow C_p (Mallows, 1973), minimum descriptive length (Rissiden, 1983). Details of these strategies and a discussion are given in Brown (1993) and Jolliffe (1993).

3.3 Partial least squares regression

Partial least squares regression (PLS) was introduced by Wold (1975) as some kind of calibration method and has occupied an important position in the chemometric literature. The statistical property of the PLS method became clearer after contributions of several statisticians and mathematicians: Wold *et al.* (1984), Manne (1987), Næs and Marten (1987), Lorber (1987), Helland (1988), Stone and Frank (1990) *et al.*. However, there are, still many unsolved questions in this area.

Several algorithms exist for PLS, Stone and Brooks (1990)'s two-stage approach is adopted here to define PLS. The basic idea for the PLS is to find the relation between \mathbf{X} and \mathbf{y} through the orthogonal latent variables \mathbf{t} . The orthogonal latent variables are constructed by maximizing the covariance in the first stage. In the second stage, the OLS method is applied to \mathbf{y} on the latent variables \mathbf{t} . The algorithm of the first stage PLS is presented below, which is also given in Helland (1990):

1. Define the starting values for \mathbf{X} residuals (\mathbf{e}_a) and \mathbf{y} residuals (\mathbf{f}_a):

$$\mathbf{e}_0 = \mathbf{X}, \mathbf{f}_0 = \mathbf{y}.$$

For $a = 1, 2, \dots$, do the steps below:

- Construct latent variable \mathbf{t} by linear combinations of the \mathbf{X} residuals from previous step; then, in order to make the latent variables more closely related to \mathbf{y} , use covariance \mathbf{y} residuals as weights(\mathbf{w}), which are standardized to have the unity length

$$\mathbf{t}_a = \mathbf{e}_{a-1} \mathbf{w}_a,$$

$$\mathbf{w}_a = \text{cov}(\mathbf{e}_{a-1}, \mathbf{f}_{a-1}), \mathbf{w}_a^T \mathbf{w}_a = 1,$$

- Determine the \mathbf{X} loadings (\mathbf{p}_a) and the \mathbf{y} loadings (\mathbf{q}_a) by OLS:

$$\mathbf{p}_a = \frac{\text{cov}(\mathbf{e}_{a-1}, \mathbf{t}_a)}{\text{var}(\mathbf{t}_a)},$$

$$\mathbf{q}_a = \frac{\text{cov}(\mathbf{f}_{a-1}, \mathbf{t}_a)}{\text{var}(\mathbf{t}_a)},$$

- Form new residuals

$$\mathbf{e}_a = \mathbf{e}_{a-1} - \mathbf{t}_a \mathbf{p}_a^T,$$

$$\mathbf{f}_a = \mathbf{f}_{a-1} - \mathbf{q}_a \mathbf{t}_a$$

Then, at each step A , the data matrices \mathbf{X} and \mathbf{y} can be described in two linear representations:

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \cdots + \mathbf{t}_A \mathbf{p}_A^T + \mathbf{e}_A$$

$$\mathbf{y} = \mathbf{q}_1 \mathbf{t}_1 + \mathbf{q}_2 \mathbf{t}_2 + \cdots + \mathbf{q}_A \mathbf{t}_A + \mathbf{f}_A$$

It can be shown that the latent variable \mathbf{t}_i are orthogonal to each other, the residual \mathbf{e}_a will be uncorrelated with \mathbf{e}_a . It can also be proved that the weight vector \mathbf{w}_i are orthogonal. While, the number of steps needed, which is equivalent to the number of latent variables, can be determined by cross-validation as with PCR. Generally, the first stage could also be viewed as:

- $\max(\mathbf{w}^T (\mathbf{X}^T \mathbf{y}))^2$ subject to $\mathbf{w}^T \mathbf{w} = 1$, then $\mathbf{w}_1 = \mathbf{X}^T \mathbf{y} / \|\mathbf{X}^T \mathbf{y}\|$.
- $\max(\mathbf{w}^T (\mathbf{X}^T \mathbf{y}))^2$ subject to $\mathbf{w}^T \mathbf{w} = 1$ and $\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}_1 = 1$, to fix \mathbf{w}_2 (see Brown 1993) using Lagrange method.

Continuing in this way, the remaining \mathbf{w}_i can be derived. And the linear space of $\mathbf{w}_1, \dots, \mathbf{w}_d$ is spanned by the Krylov sequence

$$\{\mathbf{X}^T \mathbf{y}, (\mathbf{X}^T \mathbf{X}) \mathbf{X}^T \mathbf{y}, \dots, (\mathbf{X}^T \mathbf{X})^{d-1} \mathbf{X}^T \mathbf{y}\}.$$

The second stage of *PLS* is to form the predictor by regressing \mathbf{y} on d orthogonal latent variables, where d is carefully chosen, usually by the cross-validation:

$$\hat{\mathbf{y}} = \mathbf{b}_1 \mathbf{t}_1 + \mathbf{b}_2 \mathbf{t}_2 + \dots + \mathbf{b}_d \mathbf{t}_d. \quad (5)$$

3.4 Ridge regression

Ridge regression (RR) has been promoted by Hoerl and Kennard (1970). Their success was due to theoretical and practical insights into the benefits, together with their use of the ridge trace graphic (Brown, 1990). The basic idea of ridge regression is adding a constant k to the diagonal elements of the $\mathbf{X}^T \mathbf{X}$ matrix and we have:

$$\hat{\boldsymbol{\theta}}_{RR} = (\mathbf{X}^T \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}, \quad (6)$$

The constant k can also be viewed as a Lagrange multiplier, which amounts to minimization of:

$$\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + k \|\boldsymbol{\theta}\|_2^2$$

where k is fixed beforehand. The computation becomes much more convenient after an orthogonal transformation of the equation (2):

$$\mathbf{h} = \mathbf{U}^T \mathbf{y}$$

where \mathbf{U} is extracted from the singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Letting

$$\mathbf{V}^T \boldsymbol{\theta} = \boldsymbol{\alpha}$$

the model (2) becomes

$$\mathbf{h} = \mathbf{D}\boldsymbol{\alpha}.$$

The estimator is

$$\hat{\boldsymbol{\alpha}} = (\mathbf{D}^T \mathbf{D} + k\mathbf{I})^{-1} \mathbf{D}^T \mathbf{h}, \quad (7)$$

which then is converted back using:

$$\hat{\boldsymbol{\theta}}_{RR} = \mathbf{V} \hat{\boldsymbol{\alpha}}.$$

At the beginning, a choice of \hat{k} is based on the trace graphic, which stabilized the coefficient trace and at the same time does not penalize the sum of squares too much. This does not provide an unambiguous estimator. There are several unambiguous choices for k : maximum integrated likelihood estimator (Andersson and Bloomfield (1974)), Hoerl *et al.* (1975)'s estimator, Lawless and Wang (1976)'s estimator, Minimum unbiased risk estimator, Khalaf and Shukur (2005)'s estimator, *et al.*. For its attractive computational property, the estimate of k by the generalized cross-validation (Golub *et al.* 1979), is adopted here. In the cases k is chosen to minimize

$$\frac{\|\mathbf{I} - \mathbf{A}(k)\mathbf{y}\|_2^2}{[\text{tr}(\mathbf{I} - \mathbf{A}(k))]^2} \quad (8)$$

where $\mathbf{A}(k) = \mathbf{X}(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T$.

3.5 Links and comparisons

Principle component regression, partial least squares regression and ridge regression are sometimes called regularized regression methods, which emanates from the regularization in approximation theory. It is well known that the OLS estimator is unbiased, but one can always achieve a lower mean square estimation error with a biased estimator. The estimators of PCR, PLS and RR are all such estimators and often referred to as shrinkage estimators, due to the principal goal of these three methods is to shrink the solution toward directions in the explanatory variable space of larger sample spread (Frank and Friedman 1993).

OLS, PCR and PLS have been tied together by Stone and Brooks (1990) in *continuum regression*. It is a stepwise procedure, where a generalized criteria is maximized in each step. This criteria depends on a parameter α , where $0 \leq \alpha \leq 1$ and $\alpha = 0$ gives OLS, $\alpha = 1/2$ gives PLS and $\alpha = 1$ gives PCR (Helland 2001). The weak point of this method is that they use cross validation to estimate both α and the number of latent variables d . Table 1 lists the different minimization criteria for OLS, PCR, PLS and RR. A simulation study of comparison of OLS, PLS, PCR together with RR and VSS (forward stepwise selection) can be found in Frank and Friedman (1993). According to Frank and Friedman's study: all of the biased methods (RR, PCR, PLS and VSS) provide substantial improvement over OLS. VSS provided distinctly inferior performance to the other biased methods except in

Table 1: **Minimization Criteria**

Method	Minimization Criteria
OLS	$\ \mathbf{X}\boldsymbol{\theta} - \mathbf{y}\ _2^2$
PCR	$\ \mathbf{X}_d\boldsymbol{\theta} - \mathbf{y}\ _2^2$
PLS	$\ (\mathbf{w}^T\mathbf{X})\boldsymbol{\theta} - \mathbf{y}\ _2^2$
RR	$\ \mathbf{X}\boldsymbol{\theta} - \mathbf{y}\ _2^2 + k\ \boldsymbol{\theta}\ _2^2$

the well-conditioned case in which all methods gave nearly the same performance. About PCR, PLS and RR, their results are quite similar and RR gives slightly better result. The claims as to the distinct superiority of any of these three techniques would require substantial verification (Frank and Friedman, 1993).

4. Data Analysis

Considering the data set, the aim to investigate the relationship between concentration of lactate (La) and the absorbance values of wavelengths. Thus the concentration of lactate is the response variable, and the absorbance value of a wavelength is the explanatory variable. There are 206 explanatory variables and 84 observations. Then \mathbf{X} consists of a 84×206 matrix, and apparently, the matrix $\mathbf{X}^T\mathbf{X}$ is singular. In this situation, the least squares estimator can not be applied. We apply PCR, PLS and RR separately and compare their performance.

In order to form the non-singular matrix $\mathbf{X}_d = \mathbf{U}_d\mathbf{D}_d\mathbf{V}_d^T$, a suitable number d of the largest eigenvalues can be extracted by PCR. Then the corresponding PCR-estimates of $\boldsymbol{\theta}$ is calculated by equation (4). These estimated regression coefficients are plotted for selected values of d in Figure 2(a).

An appropriate value of the number of components d is critical to successful use of PCR. If d is too small, an underfitted model results as 4 factors coefficients curve shown in Figure 2(a). The use of only 4 factors shows a very suppressed set of estimates, which does not provide enough information. If d is too large, an overfitted model is obtained. It is often observed that some small eigenvalues represent essentially noise and can be ignored with

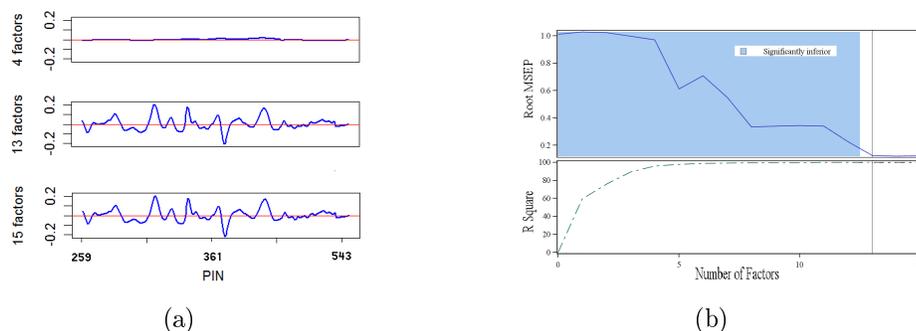


Figure 2: PCR (a): The principal component regression coefficients for 4 , 13 and 15 factors fitted, by PIN (b):Cross-validated root mean squared error of predicted by number of factors fitted

impunity. As Figure 2(a) shows, the coefficients curves of 13 and 15 components are almost same. We should consider it to exclude the 14th and 15th component when conducting model.

Table 2: Percent Variation Accounted for by PCR Factors

No. F	1	2	3	4	5	6	7
Current	59.38	16.34	13.48	6.91	1.85	1.02	0.29
Total	59.38	75.72	89.20	96.11	97.96	98.99	99.27
No. F	8	9	10	11	12	13	
Current	0.24	0.15	0.10	0.07	0.06	0.04	
Total	99.52	99.67	99.77	99.84	99.90	99.93	

Furthermore, leave-one-out cross validation method is applied to select d . The absolute minimum root MSE is 0.1209 with 14 components (figure 2(b)). But according to the Voet's test, the 13 components model's root MSE 0.1257 is only marginally larger than the absolute minimum with $p = 0.433$. So the model with 13 components is our choice in this case. The cumulative amount of variation by these components is 99 percent and the amount variation of each component is displayed in Table 2.

When applying the PLS to the data set, the estimate of θ across 206 explanatory variables are given in Figure 3(a) progressively with 4, 6, 15 latent

factors fitted. By 4 factors the final shape has emerged, and perhaps shows its most satisfactory form at 6 factors. After 15 factors, high frequency noise begins to appear. Indeed, by the leave-one-out cross validation results, the absolute minimum of root MSE is 0.1177 at 10 factors. And the smallest number of factors with $p > 0.05$ of the Voet'test is 9 with root MSPE of 0.1456. The amount of variation by each component is displayed in Table 3.

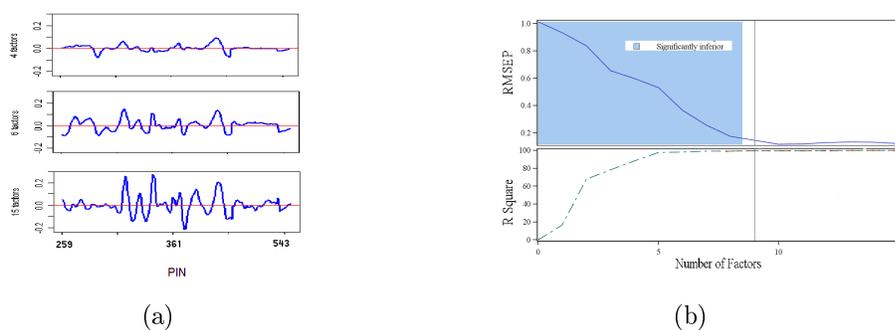


Figure 3: PLS (a): Coefficient for various numbers of factors fitted (b): Cross-validated root mean squared error of predicted by number of factors fitted

Table 3: Percent Variation Accounted for by PLS Factors

No. F	1	2	3	4	5	6	7	8	9
Current	16.65	51.28	10.56	9.66	9.72	0.44	0.94	0.13	0.20
Total	16.65	67.93	78.48	88.14	97.86	98.30	99.24	99.37	99.57

The first component accounts most variation 59.38 percent in PCR, because the maximization criteria used when specified the components. However, the first components of PLS does not accounts most. The optimal number of components of PLS is 9, which is smaller than PCR's 14. Such behavior has often been observed and is one of PLS's superior properties. Moreover PLS gives better result in prediction than PCR, according to the relatively smaller RMEP. In many cases, it is not necessary to compare the explanatory variables' loading to the factor of PCR and PLS from the prediction point of view. Instead, the solution coefficients on the original variables as a measure

of strength of the predictive relationship between the response variables and the explanatory variables will be compared later.

Considering to use of ridge regression, the cross validation ridge estimator k , based on equation (8), is 0.02. Then it is easy to obtain an estimator θ by applying equation (7). The coefficients are shown in Figure 4. The coefficients pattern of RR, exhibited many peaks in common with the patterns of PCR and PLS. The common peaks are at about PIN315, PIN328, PIN340, PIN363, PIN379 and PIN399.

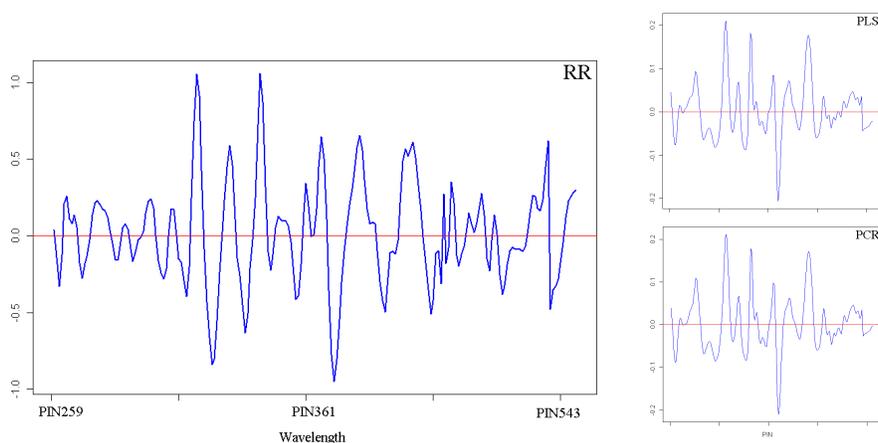


Figure 4: The estimate coefficients of the wavelength by RR, PCR and PLS

The predicted result of PCR, PLS and RR are quite similar. According to the root of the mean squares error (RMSE), ridge regression performs slightly better with $RMSE = 0.07975679$, closely followed by PCR with $RMSE = 0.2200242$ and PLS with $RMSE = 0.2252097$.

Furthermore, the absorbance values at these 206 wavelength are also used to predict the concentrations of the other 6 analyte in the grass sample. These are ethanol (EtOH), butanediol (Butdiol), succinic acids (Succi), total volatile fatty acids (VFA), water soluble carbohydrates (WSC) and ammonia nitrogen (NH₃). PCR, PLS and RR perform good and give similar result. All of the RMSEPs (Table 4) indicate that the ridge regression is slightly superior than the other two.

Table 4: RMSE by RR, PCR and PLS for each analyte

Concentration	RR	PCR	PLS
EtOH	0.1146	0.3090	0.3354
Eutdiol	0.0842	0.2482	0.1328
Succi	0.0705	0.2131	0.1981
VFA	0.1176	0.2423	0.2419
WSC	1.8175	2.7159	2.6734
NH3-N	0.1975	0.4293	0.2717

5. A Simulation Study

This section presents a simulation study of the relative performance of RR, PCR and PLS, both in prediction and estimation. This simulation is conducted in the following steps and the procedure is repeated 100 times:

1. Generate data \mathbf{X} according to the multivariate normal distribution, with random covariance matrix and mean vector equal to $\mathbf{0}$. The sample size is fixed, always equal to 200 and the number of variables changes, varies from 10, 100 to 300.
2. Generate the response variable by equation: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \in N(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\beta}$ is specified.
3. The generated data is divided into two subsets with equal sample size. One (the training set) is to build the model, the other (the test set) is to check the performance of the model.
4. Apply OLS, RR, PLS and PCR to the training set. The models are selected by cross-validation in all of the biased methods.
5. Compute the root mean of the predicted residual sum of squares over the test set by applying the model selected of each method.

5.1 Performance in prediction

In each repetition, the RMSEP (root of mean squares error of prediction) is computed, which is the criteria of whether the model's prediction is good

or not. More smaller, more better. The number of variables is set to 10, 100 and 300, which is smaller, equal, or larger than the sample size $n = 100$. Figure 5 presents the prediction result over 100 repetition. When the number of variables is equal to 10, which is quite small comparing with sample size, both the mean and the variance of RMSEP of OLS are smaller than the other three methods. It indicates that the prediction of OLS is better than the biased methods in the non-singular situation. Under the singular situation $p = 300$ and OLS fails, PCR gives inferior prediction results out of the three. Meanwhile, the number of variables has been set to other values and the simulation result is shown in Table 6. It seems that PLS performs better and more stable than RR in this case, which is different from Frank and Friedman (1993)'s study.

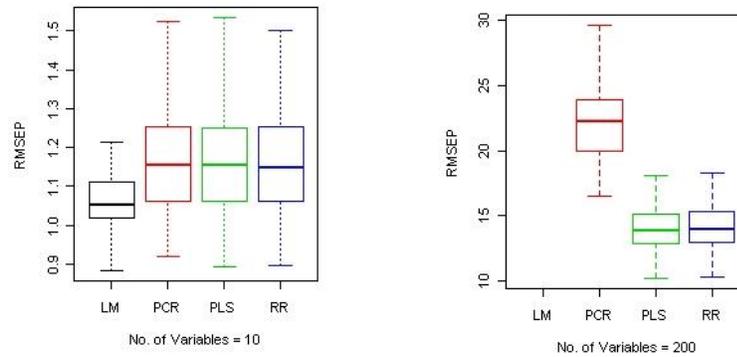


Figure 5: RMSEP of OLS, PCR, PLS and RR

5.2 Performance in estimation

In this section, the properties of the estimators has been discussed. We follow the same procedure in above simulation except that: the coefficient vector β is specified separately at step 2 in every repetition of the last simulation. While here β is fixed before simulation starts. Then the coefficient vector do not change.

The first variable's coefficient is equal to 0.2276367. Table 6 presents the summary of estimates by OLS, RR, PCR and PLS over 100 repetition when

Table 5: Mean of RMSE by OLS, RR, PCR and PLS over 100 repetitions

No. of variables	OLS	RR	PCR	PLS
10	1.062374	1.183330	1.180290	1.180106
80	2.320359	2.859169	10.18753	2.738541
90	3.231508	3.694958	11.44897	3.120313
100	2.17927	12.31211	12.09775	3.82858
110	NA	39.47694	13.49854	4.829504
200	NA	14.16937	22.11257	14.01878
300	NA	23.33061	29.96588	23.07086

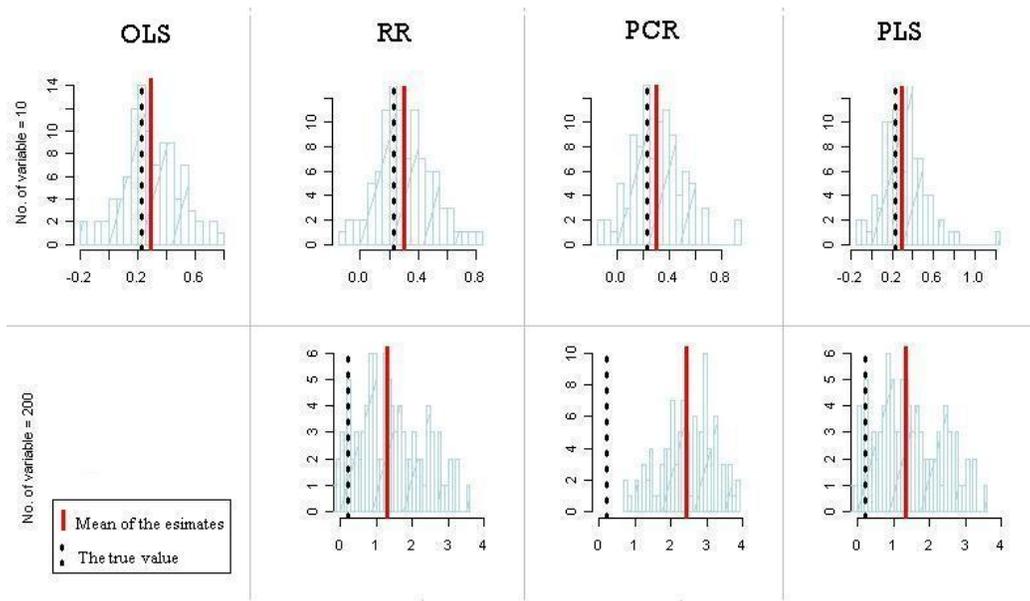


Figure 6: Histogram of the distributions of the estimators over 100 repetitions

the number of variable is set to 10 and 200. In the non-singular case, the mean of OLS estimates is relatively more close to the true value and the variance is slightly smaller as expected. RR, PCR and PLS still give similar results. In the non-singular case, the estimates by the biased methods is far away from the true value. It is difficult to find out which one is better. Moreover it seems that the distributions of their estimators also tend to the Normal distribution as shown in the Figure 6.

Table 6: The summary of the estimates: $\beta_{true} = 0.2276367$

Statistics	OLS	RR	PCR	PLS
mean($p = 10$)	0.2881599	0.3065444	0.3032562	0.3007144
var($p = 10$)	0.03573084	0.03614587	0.04042382	0.04315854
mean($p = 200$)	NA	1.323866	2.433365	1.326888
var($p = 200$)	NA	0.8876413	0.5461385	0.8911809

6. Conclusion

In this paper, we have given a description of the algorithms of the three most popular biased regression methods: PCR, PLS and RR. Then we discussed the connections and differences among them. PCR, PLS and RR are the proposed regression methods under collinear situation when OLS fails. They make a compromise between accuracy and precision. Their estimators are biased but with a lower mean squared estimation error. Furthermore, PCR, PLS and RR are applied to a spectral data set. The solution and performance of PCR, PLS, and RR tend to be quite similar in those practical data sets we applied. PLS is preferable to PCR with a smaller number of factors. According to the root mean of the predicted residual sum of squares, RR performs slightly better. But in the simulation study, PLS's prediction ability is more precise and more stable. So it requires substantial verification for any claims as to the superiority of any of the three biased regression methods.

References

- [1] Zeaiter, M. & Rutledge, D., 2009. Preprocessing Methods. *Comprehensive Chemometrics*, vol.3, p.139.
- [2] Kalivas, J.H., 2009. Calibration Methodologies *Comprehensive Chemometrics*, vol.3, p.10.
- [3] Jolliffe, I.T., 1986. *Principal component analysis*, Spring-Vlg, New York.
- [4] Brown, P.J., 1993. *Measurement, regression, and calibration*, Clarendon Press, Oxford.
- [5] Hoerl, A.E. & Kennard, R.W., 1970. Ridge regression. Biased estimation for nonorthogonal problems, *Techometrics*, vol.12, p.55-67.
- [6] Hoerl, A.E. & Kennard, R.W., 1970. Ridge regression. Applications to nonorthogonal problems, *Techometrics*, vol.12, p.69-82.
- [7] Helland, I.S., 1988. On the structure of partial least squares regression, *Communications in statistics- simulation and computation*, vol.17, p.581-607.
- [8] Helland, I.S., 1990. Partial least squares regression and statistical models, *Scandinavian Journal of statistics*, vol.17, p.97-114.
- [9] Helland, I.S., 2001. Some theoretical aspects of partial least squares regression, *Chemometrics and intelligent laboratory systems*, vol.58, p.97-107.
- [10] Stone, M. & Brooks, R.J., Cross-validated sequentially constructed prediction embracing ordinary least squares and principal components regression, *Journal of the royal statistical society. Series B(Methodological)*, vol.52, p.237-269.
- [11] Van der Voet, H., 1994. Comparing the predictive accuracy of the models using a simple randomization test, *Chemometrics and intelligent laboratory systems*, vol.25, p.313-323.
- [12] Frank, I.E. & Friedman, J.H., 1993. A statistical view of some chemometrics regression tools, *Technometrics*, vol.35, p.109-135.

- [13] Vigneau, E. & Bertrand, D., & Qannari, E.M., 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 system*, vol.35, p.231-238.
- [14] Farkas, O. & Hberger, K., 2005. Comparison of ridge regression, partial least-squares, pairwise correlation, forward-and best subset selection methods for prediction of retention indices for aliphatic alcohols, *Journal of chemical information and modeling*, vol.45, p.339-346.
- [15] Andersson, R. & Hedlund. B., 1983, HPLC analysis of organic acids in lactic acid fermented vegetables, *Z. Lebensm. Unters. Forsch.*, vol.176, p440-443.
- [16] Broderick, G.A. & Kang, J.H., 1980, Automated simultaneous determination of ammonia and total amino acids in ruminal fluid and in vitro media, *Journal of Dairy Science*, vol.63, p64-75.
- [17] Peter. U, 2006, In vitro studies on microbial efficiency from two cuts of ryegrass (*Lolium perenne*, cv. Aberdart) with different proportions of sugars and protein, *Animal Feed Science Technology*, vol.126, p145-156.
- [18] Broderick, G.A., 1987, Determination of protein degradation rates using a rumen in vitro system containing inhibitors of microbial nitrogen metabolism, *British Journal Nutrition*, vol.58, p463-476.