

STA 4107/5107
Statistical Learning:
Principle Components and Partial Least Squares Regression

March 28, 2007

1 Introduction

Principal components analysis is traditionally presented as an interpretive multivariate technique, where the loadings are chosen to maximally explain the variance in the variable. However, we will consider it here mainly as a statistical learning tool, by using the derived components in a least squares regression to predict unobserved response variables using the principal components. Principal components aims to explain as much of the variation in the data as possible by finding linear combinations that are independent of each other and in the direction of the greatest variation. Each principal component is a linear combination of all variables. The first principal component explains the most variation, the second PC the second most, and so on. There are as many principal components as there are variables, but we usually choose only the first few for both exploratory and regression analysis.

Partial least squares is a method of data dimension reduction, similar to principal components, to find the most relevant factors for both prediction and interpretation, and is derived from Herman Wold's development of iterative fitting of bilinear models (Wold, 1981, 1983). Partial least squares regression (PLSR) improves upon principal components analysis by actively using the response variables during the bilinear decomposition of the predictors. Principal components focuses on the *variance* in the predictors, while partial least squares focuses on the *covariance* between the response and the predictors. By balancing the information in both the predictors and the response, PLS reduces the impact of large, but irrelevant predictor variations. Estimation of prediction error is achieved using cross-validation.

1.1 Statistical Learning

There are many statistical data analysis techniques that fall into the category of *statistical learning*. The two regression techniques considered here, partial least squares regression (PLSR) and principal components regression, are among them. The most commonly used statistical learning method is normal linear regression. We refer to it as 'learning' because we have data that we want to use to discover the relationship between a quantity that we would like to predict and one or several other quantities, called predictors. That is, we would like to 'learn from examples.' The data we have

consists of measurements of both the response (the quantity we would like to predict), along with corresponding measurements of the predictors, and is called the ‘training set’. We then use our data to ‘train’ a statistical algorithm, which produces a mathematical relationship via parameter estimation, from the known data, which we can then use to predict the quantity of interest in a test data set where all we have measured are the predictors. In this sense we have ‘learned’ from our training data to predict future observations.

When we wish to predict Y from a high-dimensional X several issues often arise:

1. Lack of univariate predictive ability: no single X -variable adequately predicts Y .
2. Collinearity: The X variables are often closely related to one another and hence exhibit high collinearity.
3. Lack of knowledge: Our *a priori* understanding of the underlying mechanisms and the probability distribution that generated the data may be incomplete or erroneous.
4. Lack of a full-rank data matrix: We cannot use conventional classification or regression methods unless the number of predictors is less than the number of observations.

Hence we need reliable dimension-reduction techniques that offer flexibility (i.e., make no assumptions that cannot be validated, but still offer robust solutions) in the calibration of data, so that the above problems can be solved while increasing both our understanding and predictive ability. For these reasons, data compression becomes an important aspect in the calibration of complex, high-dimensional data.

1.2 Data Example

For this chapter we will come back to the US Crime Data to extract principal components for a regression and to fit a partial least squares regression, both to predict crime rate based on the other variables. Recall that this data set contains variables measured on several cities in the US. There are 47 cases and 14 variables.

- R: Crime rate: number of offenses reported to police per million population
- Age: The number of males of age 14-24 per 1000 population
- S: Indicator (or dummy) variable for Southern states (0 = No, 1 = Yes)
- Ed: Mean number of years of schooling x 10 for persons of age 25 or older
- Ex0: 1960 per capita expenditure on police by state and local government
- Ex1: 1959 per capita expenditure on police by state and local government
- LF: Labor force participation rate per 1000 civilian urban males age 14-24
- M: The number of males per 1000 females
- N: State population size in hundred thousands
- NW: The number of non-whites per 1000 population
- U1: Unemployment rate of urban males per 1000 of age 14-24
- U2: Unemployment rate of urban males per 1000 of age 35-39
- W: Median value of transferable goods and assets or family income in tens of dollars.
- X: The number of families per 1000 earning below 1/2 the median income

We will begin our discussion of bilinear modeling with principal components. Partial least squares is an easy extension after that.

2 Principal Components

Principal components is one of the most commonly used multivariate techniques. Several other multivariate techniques are derived from principal components, as we saw was true for factor analysis.

Principal component analysis was originated by Pearson (1901) and later developed by Hotelling (1933). The application of principal components is discussed by Rao (1964), Cooley and Lohnes (1971), and Gnanadesikan (1977). Excellent statistical treatments of principal components are found in Kshirsagar (1972), Morrison (1976), and Mardia, Kent, and Bibby (1979).

2.1 Examples

1. Naiman *et al.* (1994) studied the impact of beavers on aquatic biogeochemistry in boreal forests. They sampled four habitats for soil and pore water constituents. The variables measured were total nitrogen, nitrates, ammonium, phosphorus, potassium, calcium, magnesium, iron, sulfate, pH, Eh, percentage of organic carbon, bulk density, nitrogen fixation, moisture, and redox potential. Three components explained 75% of the variation. Component 1 loaded strongly on nitrogen and phosphorus, component 2 on moisture and organic matter, and component 3 on ammonium and redox.
2. Lovett *et al.* (2000) studied the chemistry of forested watersheds in the Catskill Mountains in New York. They chose 39 streams and measured the concentration of ten chemical variables and four watershed variables. These data were later used by Quinn and Keough to examine the relationships between the 39 streams.
3. A. Thomson and Randall-Maciver, R. (1905) took four measurements from five different time periods on 30 Egyptian skulls from each time period. It is of interest to test for differences across the time periods on the different variables. A principle components analysis can be used to simplify the data, and eliminate multicollinearity before performing the MANOVA.
4. A. Weber (1973) measured protein consumption in twenty-five European countries for nine food groups. A principal components analysis of the nine variables reveals four large dimensions of variability. The first principal component may be interpreted roughly as a measure of total meat consumption. The second, third, and fourth principal components may be interpreted as red meat, white meat, and fish consumption respectively.

2.2 Basic Concepts

To simplify, we'll consider the case with two variables X_1 and X_2 and then extend to the general case of p variables. As with factor analysis, we'll work with the standardized variables, i.e.: $x_1 = \frac{(X_1 - \bar{X}_1)}{S_1}$ and $x_2 = \frac{(X_2 - \bar{X}_2)}{S_2}$. This gives each of the variables expected value equal to zero with variance equal to one. This does not eliminate the correlation between the variables, however.

Principal components analysis can be performed on the covariance (rather than correlation) matrix if we simply subtract the mean but do not divide by the standard deviation. SAS uses the standardized variables and we will consider this situation from here on. The advantages of complete

standardization is that large variability can skew the analysis. Standardization helps equalize the variances and so stabilizes the analysis.

The principal components are linear functions of x_1 and x_2 and can be written as

$$\begin{aligned} C_{1i} &= a_{11}x_{1i} + a_{12}x_{2i} \\ C_{2i} &= a_{21}x_{1i} + a_{22}x_{2i} \end{aligned}$$

where i is the subscript indicating the observation number. That is, we obtain a principal component for every observation. If the sample size is given by n , then we have n values for C_1 and C_2 . We will neglect the subscript for observation number from here on.

The coefficients are chosen to satisfy the following constraints:

1. The variance of C_1 is as large as possible.
2. The n values of C_1 and C_2 are uncorrelated.
3. The sum of the squared coefficients for any principal component is equal to one. That is:

$$a_{11}^2 + a_{12}^2 = a_{21}^2 + a_{22}^2 = 1$$

It follows that:

- $\mu(C_1) = \mu(C_2) = 0$
-

$$\begin{aligned} Var(C_1) &= Var(a_{11}x_1 + a_{12}x_2) = a_{11}^2 + a_{12}^2 + 2a_{11}a_{12}\rho_{12} \\ Var(C_2) &= Var(a_{21}x_1 + a_{22}x_2) = a_{21}^2 + a_{22}^2 + 2a_{21}a_{22}\rho_{21} \end{aligned}$$

where ρ_{12} is the correlation between x_1 and x_2 .

- The total variance is simply the number of variables. In this case 2, in general p .
- The correlation between the j^{th} principal component C_j and the k^{th} variable x_k is $r_{jk} = a_{jk}(Var\{C_j\})^{1/2}$. Hence, for a given C_k we can compare the a_{jk} to quantify the relative degree of dependence of C_k on each of the standardized variables. The correlation is sometimes called the *factor loading*.

The solution to the above problem will be shown graphically on the board during lecture.

Graphical Illustration of Principal Components C_1 and C_2

Principal components analysis is essentially rotating the original axes to new ones defined by C_1 and C_2 . The angle of rotation is determined uniquely by the solution to the above optimization problem. For a given point (x_1, x_2) , the values of C_1 and C_2 are found by drawing perpendicular lines from the point to the new PC axes. The N derived values of the C_1 will have the largest variance as per constraint 1, and the N values of C_1 and C_2 will have correlation equal to zero as per constraint 2. Let's look at a two variable principal components analysis of the US crime data. Consider the variables `education` and `number of males per 1000 females`, Ed and M respectively.

The code below creates a new variable in the crime data set called "obsno", runs the PCA and creates the plot shown above.

SAS Code and Output

```
data UScrime;
set UScrime;
obsno=_N_; run;

proc princomp data=UScrime out=UScrimePC;
var Ed M; run;
title2 'Plot of the First Two Principal Components';
```

```
%plotit(data=UScrimePC,labelvar=obsno,
        plotvars=Prin2 Prin1, color=black, colors=blue);
run;
```

The PRINCOMP Procedure

Simple Statistics

	Ed	M
Mean	105.6382979	983.0212766
StD	11.1869985	29.4673654

Correlation Matrix

		Ed	M
	Ed	1.0000	0.4369
	M	0.4369	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.43691492	0.87382983	0.7185	0.7185
2	0.56308508		0.2815	1.0000

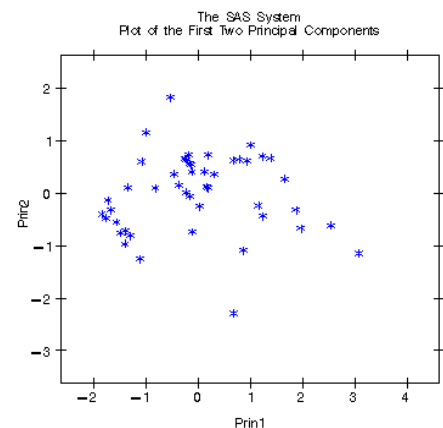
Eigenvectors

		Prin1	Prin2
	Ed	0.707107	0.707107
	M	0.707107	-.707107

The eigenvectors are the columns labeled Prin1 and Prin 2 and are the set of coefficients from the linear combination for the principal component. Here $a_{11} = 0.70107 = a_{22}$ and $a_{12} = -a_{21}$. This will always be the case for analyses with two variables, but is not generally true. Notice also that $a_{11}^2 + a_{12}^2 = a_{21}^2 + a_{22}^2 = 1$ as required by constraint 3. The eigenvectors are the loadings for the linear combination. Hence, $C_1 = 0.707107 * ed + 0.707107 * m$ and $C_2 = 0.707107 * ed - 0.707107 * m$, where ed and m represent the standardized forms of the original variables.

The first and second eigenvalues represent the variance of C_1 and C_2 , respectively. The proportions reported in the correlation matrix are the proportions of the variance explained by the corresponding PC. Because SAS is looking at the standardized variables, the sum of the variances (eigenvalues) of the two variables is 2.

It is always a good idea to take a look at plots of the data with the principal components as the axes. This is easy to do when we have only 2 PCs. If we have more we can look at pairwise plots, i.e., C_1 by C_2 , C_1 by C_3 and so on. When the number of PCs gets large, we will have to



consider other methods. These will be discussed later.

The basic concepts discussed for two principal components can be extended to the case of p variables. Each principal component is a linear combination of all of the x variables. The loadings for the linear combination are chosen to satisfy the following constraints:

1. $Var \{C_1\} \geq Var \{C_2\} \geq \dots \geq Var \{C_p\}$
2. The PCs are mutually independent, i.e. have correlation zero.
3. For each principal component the sum of the squared loadings is one.

2.3 Interpretation

One of the goals of PCA is dimension reduction. If fewer than p principal components explain a sufficient amount of the variation then we can choose the first $m < p$ principal components instead of using all p variables. In our example above, if 71.8% of the variance was deemed adequate then we could use the first principal component instead of both variables. Various rules-of-thumb have been proposed for choosing the cut-off.

- Choose the first several PCs so that 80% of the variance is explained.
- Discard all PCs that have variances less than $70/p$ percent of the total variance. Some recommend $100/p$ as the cut-off.
- Plot the PC number on the horizontal axis versus the individual eigenvalues on the vertical axis and choose the cutoff point where lines joining consecutive points are relatively steep left of the cutoff and relatively flat right of the cutoff point, i.e., the *scree plot*.
- Keep as many PCs as can be interpreted or will be useful in future analyses.

When using PCA primarily as an exploratory tool, the main idea is similar to what we saw in factor analysis – we look at the loadings to see where the high correlations are and “name” the components based on some broad categorization of the loadings.

We will primarily be using PCA as explanatory variables in a predictive model, as so we will use leave-one-out cross validation to choose the number of principal components.

SAS Ouput for US Crime Data: the full analysis

		Correlation Matrix						
		R	Age	So	Ed	Ex0	Ex1	LF
R	R	1.0000	-.0895	-.0906	0.3228	0.6876	0.6667	0.1889
Age	Age	-.0895	1.0000	0.5844	-.5302	-.5057	-.5132	-.1609
So	So	-.0906	0.5844	1.0000	-.7027	-.3726	-.3762	-.5055
Ed	Ed	0.3228	-.5302	-.7027	1.0000	0.4830	0.4994	0.5612
Ex0	Ex0	0.6876	-.5057	-.3726	0.4830	1.0000	0.9936	0.1215
Ex1	Ex1	0.6667	-.5132	-.3762	0.4994	0.9936	1.0000	0.1063
LF	LF	0.1889	-.1609	-.5055	0.5612	0.1215	0.1063	1.0000
M	M	0.2139	-.0287	-.3147	0.4369	0.0338	0.0228	0.5136
N	N	0.3375	-.2806	-.0499	-.0172	0.5263	0.5138	-.1237

NW	NW	0.0326	0.5932	0.7671	-.6649	-.2137	-.2188	-.3412
U1	U1	-.0505	-.2244	-.1724	0.0181	-.0437	-.0517	-.2294
U2	U2	0.1773	-.2448	0.0717	-.2157	0.1851	0.1692	-.4208
W	W	0.4413	-.6701	-.6369	0.7360	0.7872	0.7943	0.2946
X	X	-.1790	0.6392	0.7372	-.7687	-.6305	-.6482	-.2699

	M	N	NW	U1	U2	W	X
R	0.2139	0.3375	0.0326	-.0505	0.1773	0.4413	-.1790
Age	-.0287	-.2806	0.5932	-.2244	-.2448	-.6701	0.6392
So	-.3147	-.0499	0.7671	-.1724	0.0717	-.6369	0.7372
Ed	0.4369	-.0172	-.6649	0.0181	-.2157	0.7360	-.7687
Ex0	0.0338	0.5263	-.2137	-.0437	0.1851	0.7872	-.6305
Ex1	0.0228	0.5138	-.2188	-.0517	0.1692	0.7943	-.6482
LF	0.5136	-.1237	-.3412	-.2294	-.4208	0.2946	-.2699
M	1.0000	-.4106	-.3273	0.3519	-.0187	0.1796	-.1671
N	-.4106	1.0000	0.0952	-.0381	0.2704	0.3083	-.1263
NW	-.3273	0.0952	1.0000	-.1565	0.0809	-.5901	0.6773
U1	0.3519	-.0381	-.1565	1.0000	0.7459	0.0449	-.0638
U2	-.0187	0.2704	0.0809	0.7459	1.0000	0.0921	0.0157
W	0.1796	0.3083	-.5901	0.0449	0.0921	1.0000	-.8840
X	-.1671	-.1263	0.6773	-.0638	0.0157	-.8840	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.83821023	3.19805466	0.4170	0.4170
2	2.64015557	0.68668964	0.1886	0.6056
3	1.95346594	0.56783130	0.1395	0.7451
4	1.38563464	0.75103497	0.0990	0.8441
5	0.63459967	0.28138293	0.0453	0.8894
6	0.35321674	0.04316423	0.0252	0.9147
7	0.31005251	0.05728970	0.0221	0.9368
8	0.25276280	0.02455954	0.0181	0.9549
9	0.22820326	0.03886209	0.0163	0.9712
10	0.18934117	0.09703968	0.0135	0.9847
11	0.09230149	0.02326696	0.0066	0.9913
12	0.06903453	0.02106454	0.0049	0.9962
13	0.04796999	0.04291851	0.0034	0.9996
14	0.00505147		0.0004	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
R	0.195406	0.258391	0.193888	0.534144	-.017243	0.085481	-.557090
Age	-.303443	-.048076	0.198955	0.305854	-.182525	0.747824	0.184035
So	-.322315	0.226472	0.109461	0.146650	-.245934	-.394639	-.036166
Ed	0.346603	-.220979	0.054885	0.059558	-.082809	0.008493	-.010315
Ex0	0.329105	0.307847	0.149102	0.118965	-.133669	-.053776	0.137157
Ex1	0.330720	0.302508	0.151496	0.097174	-.171296	-.058292	0.162132
LF	0.177258	-.359197	0.230322	0.250198	0.566783	-.230435	0.214260
M	0.124218	-.326638	-.187086	0.575464	-.016305	-.046333	0.123007
N	0.118995	0.440153	0.114267	-.152021	0.630370	0.321258	0.062383
NW	-.282229	0.281385	0.176728	0.238623	0.017361	-.227110	0.617694
U1	0.039383	0.057652	-.666912	0.183434	0.077115	0.159778	0.187261
U2	0.024753	0.352858	-.533100	0.137922	0.085386	-.100172	-.039114
W	0.388120	0.066180	0.021890	-.054517	-.166165	-.003452	0.104487

X X -0.366836 0.060676 0.029364 0.207467 0.303002 -0.160441 -0.339527

In interpretive PCA, we would now follow a similar procedure to what we did in factor analysis. We would group highly loaded variables in the first few PCs and call each PC by some grouping name for the variables that are highly loaded in that PC. For example, the first PC in the crime data has high loadings for `age`, `so`, `ed`, `ex0`, `ex1` and `x`. If we can think of a “name” for this group of variables, then we would say that PC1 is the “name” PC. etc.

We do not plan to stop here, however. Our next step will be to perform a PC regression, where we remove the variable we’d like to predict `crime rate` and build PCs with the remaining variables. Finally, we will perform a *principal components regression* on `crime rate` using the first few PCs.

2.4 Principal Components Regression

We will use a different SAS procedure `Proc PLS`. This procedure will perform both principal components regression and *partial least squares* regression. (Actually, it performs a couple other kinds of bilinear modeling as well.) First we will look at principal components regression.

The basic code for principal components regression and the output is shown below.

SAS Code and Output for PCR

```
proc pls data=UScrime method=pcr;
model r = Age So Ed Ex0 Ex1 LF M N NW U1 U2 W X;
run;
```

The `method=pcr` statements specifies that we want a principal components regression and not one of the other methods available in `Proc PLS`.

The PLS Procedure

Percent Variation Accounted for by Principal Components

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	43.4751	43.4751	15.4707	15.4707
2	19.3863	62.8614	7.7896	23.2603
3	14.6609	77.5223	2.9128	26.1731
4	7.5052	85.0276	35.9542	62.1274
5	4.8794	89.9069	0.0422	62.1696
6	2.7103	92.6173	0.0290	62.1986
7	2.0080	94.6253	0.5503	62.7488
8	1.8222	96.4476	0.4889	63.2377
9	1.4704	97.9180	0.2639	63.5016
10	1.0257	98.9436	9.1612	72.6628
11	0.5417	99.4853	0.6104	73.2732
12	0.4745	99.9598	2.8914	76.1646
13	0.0402	100.0000	0.7590	76.9236

At this point we could choose the number of components we want to include in the model by specifying the percent of the variance we want explained, either in the predictors or in the model effects. However, there is a more rigorous way to choose the number of factors.

2.4.1 Cross Validation and Choosing the Number of Factors

None of the regression methods implemented in the PLS procedure fit the observed data any better than ordinary least squares (OLS) regression; in fact, all of the methods approach OLS as more factors are extracted. However, when there are a large number of predictors, OLS can over-fit the observed data leading to large prediction errors. Regression methods with fewer extracted factors can provide better predictability of future observations. We consider here several types of cross validation procedures that are available in `Proc PLS`.

Test set validation consists of fitting the model to only part of the available data (the training set) and then measuring how well model predicts the rest of the data (the test set). However, it is rare that we have enough data to make both sets large enough for test set validation to be useful.

Another method is to hold out successive blocks of observations as test sets, for example, observations 1 through 7, then observations 8 through 14, and so on; this is known as blocked cross validation. A similar method is split-sample cross validation, in which successive groups of widely separated observations are held out as the test set, for example, observations 1, 11, 21, ..., then observations 2, 12, 22, ..., and so on. Finally, test sets can be selected from the observed data randomly; this is known as random sample cross validation.

We have already discussed leave-one-out cross validation. This cross validation method is generally preferred because it uses as much of the data as possible during model fitting, but still allows for cross validation on the left-out points.

Which validation you should use depends on your data. Test set validation is preferred when you have enough data to make a division into a sizable training set and test set that represent the predictive population well. You can specify that the number of extracted factors be selected by test set validation by using the `CV=TESTSET(data set)` option, where `data set` is the name of the data set containing the test set. If you do not have enough data for test set validation, you can use one of the cross validation techniques.

The most common technique is leave-one-out cross validation (which you can specify with the `CV=ONE` option or just the `CV` option). We can specify the number of test sets in blocked or split-sample validation with a number in parentheses after the `CV=` option (i.e. `CV=split(10)`). Note that `CV=ONE` is the most computationally intensive of the cross validation methods, since it requires a re-computation of the PLS model for every input observation. Also, note that using random subset selection with `CV=RANDOM` may lead two different researchers producing different PLS models on the same data (unless they use the same seed).

Whichever validation method you use, the number of factors chosen is usually the one that minimizes the predicted residual sum of squares (PRESS); this is the default choice if you specify any of the CV methods with `PROC PLS`. However, often models with fewer factors have PRESS statistics that are only marginally larger than the absolute minimum. To address this, van der Voet (1994) has proposed a statistical test for comparing the predicted residuals from different models; when you apply van der Voet's test `CVTEST`, the number of factors chosen is the fewest with residuals that are insignificantly larger than the residuals of the model with minimum PRESS.

van der Voet's Test

To see how van der Voet's test works, let $R_{t,k}$ be the predicted residual for response k for the model with t extracted factors; the PRESS statistic is $\sum_k R_{t,k}^2$. Also, let t_{min} be the number of factors for which PRESS is minimized. The critical value for van der Voet's test is based on the differences between squared predicted residuals

One alternative for the critical value is $C_t = \sum_k R_{t,k}^2 - R_{t_{min},k}^2$, which is just the difference between the squared PRESS residuals for t and t_{min} factors. The significance level is obtained by comparing C_t with the distribution of values that result from randomly exchanging $R_{t,k}^2$ and $R_{t_{min},k}^2$. In practice, a Monte Carlo sample of such values is simulated and the significance level is approximated as the proportion of simulated critical values that are greater than C_i .

Another alternative uses Hotellings T^2 as an approximation.

If you apply van der Voet's test by specifying the CVTEST option, then, by default, the number of extracted factors chosen is the least number with an approximate significance level that is greater than 0.10.

You choose which critical value you'd like to use with the `cvtest(stat=option)` command.

SAS Code and Output for PCR

```
proc pls data=UScrime method=pcr cv=one cvtest(stat=press);
model r = Age So Ed Ex0 Ex1 LF M N NW U1 U2 W X;
run;
```

The PLS Procedure

Cross Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS	Prob > PRESS
0	1.021739	0.0100
1	0.960904	0.0030
2	0.943789	0.0030
3	0.961149	0.0030
4	0.709936	1.0000
5	0.713109	0.3970
6	0.728181	0.1050
7	0.729494	0.1460
8	0.766771	0.0240
9	0.83655	0.0030
10	0.722135	0.4190
11	0.732126	0.3610
12	0.711253	0.4840
13	0.717465	0.4320

Minimum root mean PRESS	0.7099
Minimizing number of factors	4
Smallest number of factors with p > 0.1	4

Percent Variation Accounted for by Principal Components

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	43.4751	43.4751	15.4707	15.4707
2	19.3863	62.8614	7.7896	23.2603
3	14.6609	77.5223	2.9128	26.1731
4	7.5052	85.0276	35.9542	62.1274

We are lucky here and the minimizing number of factors is also significantly better according to van der Voet's test. In the case where we obtained a contradiction, we would have to decide whether to use van der Voet's test to choose the number of factors, or whether we preferred to choose the number of factors that minimized the PRESS, as we did in ordinary linear regression.

The default value for significance in van der Voet's test is 0.10. We can specify a different value using `pval=` in the `Proc PLS` line.

2.4.2 Fitting the Model using a Selected Number of Factors

If we wish to fit a PCR using our own criterion for the number of factors we can specify the number in the `Proc PLS` line. First we would perform the PCA and look at the percent variance explained. For example, say we wished to explain at least 90% of the variance in the model effects. Then we would fit a PCR using 6 factors. Below is output for a model choosing 4 factors.

SAS Code and Output for Specifying the Number of Factors

```
proc pls data=UScrime method=pcr nfact=4;
model r = Age So Ed Ex0 Ex1 LF M N NW U1 U2 W X;
run;
```

The PLS Procedure

Percent Variation Accounted for by Principal Components

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	43.4751	43.4751	15.4707	15.4707
2	19.3863	62.8614	7.7896	23.2603
3	14.6609	77.5223	2.9128	26.1731
4	7.5052	85.0276	35.9542	62.1274
5	4.8794	89.9069	0.0422	62.1696
6	2.7103	92.6173	0.0290	62.1986

2.4.3 Interpretation in PCR

Interpretation in PCR is similar to that in PCA. We would often be interested in which variables loaded heavily in which factors. Usually, in PCR analyses interest is primarily in prediction. That

is, we would have some “unknown” for which we would like to use our model to predict the response variable.

To ascertain which variables are loading heavily in the different components, we simply look at the eigenvectors as we did in factor analysis. If it makes sense, we can also plot the loadings versus variable number.

SAS Code and Output

```

                                The PLS Procedure

                                Model Effect Loadings

Number of
Extracted
Factors      Age      So      Ed      Ex0      Ex1      LF
-----
1      -0.317175    -0.340006    0.356104    0.316981    0.319519    0.183097
2      -0.124021     0.179417   -0.214396    0.299853    0.297130   -0.400916
3      -0.126503   -0.098462   -0.029560   -0.176204   -0.182127   -0.151188
4       0.332991     0.218088    0.059579    0.323438    0.302156    0.307190

```

```

                                Model Effect Loadings

Number of
Extracted
Factors      M      N      NW      U1      U2      W
-----
1      0.126491     0.107309   -0.303570    0.044308    0.016648    0.392071
2     -0.358458     0.453749    0.222782    0.118836    0.400497    0.094476
3      0.313943   -0.196424   -0.160714    0.679475    0.508462   -0.053782
4      0.575033     0.017623    0.411523    0.082724    0.109227    0.029474

```

```

                                Model Effect Loadings

                                Number of
                                Extracted
                                Factors      X
-----
1      -0.381132
2       0.008553
3       0.016404
4       0.166619

```

We can also plot the loadings versus principal component number, though for the US crime data, this is not particularly more helpful than just looking at the loadings, but in some data sets where there is a natural ordering or grouping of variables, and a very large number, such as in hyperspectral reflectance, data these plots can be revealing.

SAS Code for Plotting Loadings versus PC number

```

ods listing close;
ods output XLoadings=xloadings;

proc pls data=UScrime nfac=4 details method=pcr;

```

```

model r = Age So Ed Ex0 Ex1 LF M N NW U1 U2 W X;
      output out=xscores XSCORE=T;
run;

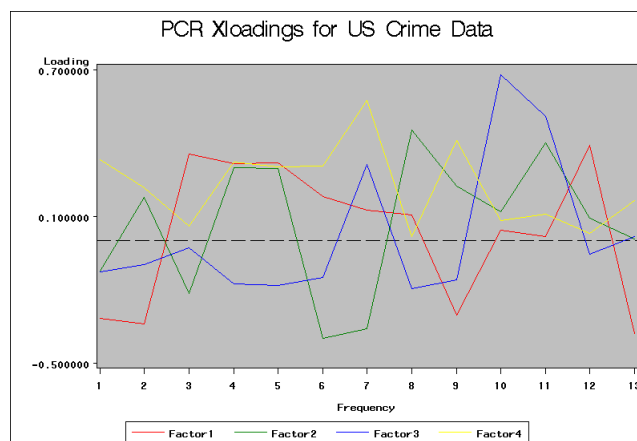
ods listing;
proc transpose data=xloadings(drop=NumberOfFactors)
      out=xloadings;
data xloadings; set xloadings;
obsno = _n_;
      rename col1=Factor1 col2=Factor2 col3=Factor3 col4=Factor4;
run;

goptions border;
axis1 label=("Loading" ) major=(number=3) minor=none;
axis2 label=("Frequency")          minor=none;
symbol1 v=none i=join c=red      l=1;
symbol2 v=none i=join c=green    l=1;
symbol3 v=none i=join c=blue     l=1;
symbol4 v=none i=join c=yellow   l=1;

legend1 label=none cborder=black;
proc gplot data=xloadings;
      plot (Factor1 Factor2 Factor3 Factor4)*obsno
      / overlay legend=legend1 vaxis=axis1
      haxis=axis2 vref=0 lvref=2 frame cframe=ligr;
      title 'Xloadings for US Crime Data';
run; quit;

```

The graph produced by the above code is shown below.



3 Partial Least Squares

Partial least squares is sometimes referred to a “soft modeling”. This is in contrast with ordinary least squares regression that makes “hard” assumptions including lack of multicollinearity among the predictor variables, with well-understood relationships to the response variable. “Soft modeling” refers to modeling when these assumptions cannot be made. PLS can handle multicollinearity, many predictor variables, and because its focus is prediction, not explanation, lack of well-understood relationships of the response to the predictor variables is not a problem.

3.1 Examples

1. Partial least squares was used by Davies in the multivariate calibration of octane using 226 near infra-red wavelengths.
2. Nguyen and Rocke used partial least squares to classify tumors using micro-array gene expression data.
3. Karp et al used partial least squares discriminant analysis to classify two-dimensional difference gel studies in expression proteomics.
4. Wilson et al used partial least squares to reduce the dimension of hyperspectral reflectance data to derive components to use in a logistic discrimination of salt marsh plants exposed to petroleum and heavy metal contamination.

3.2 Basic Concepts

Partial least squares (PLS) was first developed by Herman Wold, with an interest in applications for the social sciences, especially economics. However, its first popularity was among chemists and used for high dimensional calibration problems, for example using a large number of reflectance measurements to estimate a concentration.

The object of PLS regression is to predict the response vector, \mathbf{Y} , from the matrix of explanatory variables, \mathbf{X} , and to ascertain their common structure. This is often referred to as a *latent variable* approach to modeling the covariance structure between a response vector and the design matrix. As mentioned earlier, when \mathbf{X} is full rank (i.e., fewer predictors than observations), this goal could be accomplished using ordinary multiple regression. If the number of predictors is large compared to the number of observations, \mathbf{X} is likely to be singular (impossible to estimate the variances) and the regression approach is no longer feasible (i.e., because of multicollinearity).

Several approaches have been developed to cope with this problem. We have seen the use of principal component regression on the \mathbf{X} matrix in order to use principal components of \mathbf{X} as regressors on \mathbf{Y} . The mutual independence of the principal components eliminates the multicollinearity problem. But the principal components are chosen to capture the variability in \mathbf{X} rather than \mathbf{Y} , so a principal components regression may include components that are in fact irrelevant to the prediction of \mathbf{Y} or may load sub-optimally for predicting \mathbf{Y} .

By contrast, PLS regression finds components from \mathbf{X} that are the most relevant for predicting \mathbf{Y} . Specifically, PLS regression searches for a set of components (sometimes called latent vectors)

that performs a simultaneous decomposition of \mathbf{X} and \mathbf{Y} with the constraint that these components explain as much as possible of the *covariance* between \mathbf{X} and \mathbf{Y} . As in principal components, the derived components are used in a regression to predict \mathbf{Y} .

It is difficult to show the math behind PLS without using matrix algebra. Articles describing the underlying math are posted on the website. We will outline the basic procedure below.

As with factor analysis and principal components analysis, we'll work with the standardized variables, i.e.: $x_1 = \frac{(X_1 - \bar{X}_1)}{S_1}$ and $x_2 = \frac{(X_2 - \bar{X}_2)}{S_2}$. Standardization helps equalize the variances and so stabilizes the analysis.

The components in a partial least squares decomposition are similar to those in PCA, linear functions of x_1, x_2, \dots, x_p and we have a principal component for every observation and for every predictor. They can be written as

$$\begin{aligned} C_{1i} &= v_{11}x_{1i} + v_{12}x_{2i} + \dots + v_{1p}x_{pi} \\ C_{2i} &= v_{21}x_{1i} + v_{22}x_{2i} + \dots + v_{2p}x_{pi} \\ &\vdots \\ C_{pi} &= v_{p1}x_{1i} + v_{p2}x_{2i} + \dots + v_{pp}x_{pi} \end{aligned}$$

where i is the subscript indicating the observation number. Recall also that this system of equations can be inverted thusly:

$$\begin{aligned} x_1 &= l_{11}C_1 + l_{21}C_2 + \dots + l_{p1}C_p \\ &\vdots \\ x_p &= l_{1p}C_1 + l_{2p}C_2 + \dots + l_{pp}C_p \end{aligned}$$

In principal components, the coefficients are the same in both systems of equations. That is $a_{ij} = l_{ij}$ for all i, j . In partial least squares this is not necessarily the case, hence the change in notation. In matrix representation we have

$$\begin{aligned} \mathbf{C} &= \mathbf{XV} \\ \mathbf{X} &= \mathbf{CL}' + \mathbf{E} \end{aligned}$$

The matrix \mathbf{E} contains the residuals from the decomposition of \mathbf{X} , i.e., the variance in the predictors left unexplained by the principal components procedure. The matrix \mathbf{V} contains the “weights” and the matrix \mathbf{L} contains the “loadings”. Finally, in the bilinear regression we consider

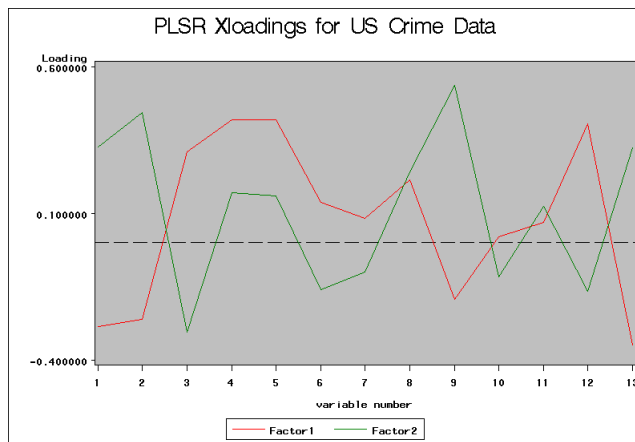
$$\mathbf{Y} = \mathbf{CQ}' + \mathbf{F}$$

where \mathbf{F} contains the residuals. This represents the regression of the response variable(s) by the derived components. The matrix \mathbf{Q} represents the regression coefficients. In principal components, the loadings and the weights will be the same. We solve this system by deriving the components to maximize the explained variance in the \mathbf{X} . In partial least squares, we consider also the variance in \mathbf{Y} , or the covariance between \mathbf{X} and \mathbf{Y} . Hence $\mathbf{V}, \mathbf{L}, \mathbf{Q}$ change as we iteratively solve each of these equations.

Let's see this in matrix notation for an example with 3 observations and 2 predictor variables.

3.3 Fitting the PLS Model

The code for fitting a PLS model is very similar to that for PCR: we simply substitute `method=pls` where we had `method=pcr`. Output for the US crime data for a PLSR predicting crime rate using the other variables is shown below.



SAS Output

The PLS Procedure

Cross Validation for the Number of Extracted Factors

Number of Extracted Factors	Root Mean PRESS	Prob > PRESS
0	1.021739	0.0210
1	0.865846	0.0280
2	0.761182	0.2020
3	0.70568	0.4800
4	0.725099	0.3340
5	0.745355	0.0390
6	0.706153	0.4280
7	0.704784	1.0000
8	0.70883	0.3810
9	0.709483	0.3880
10	0.733169	0.0820
11	0.723116	0.2210
12	0.720464	0.2570
13	0.717465	0.2890

Minimum root mean PRESS 0.7048
 Minimizing number of factors 7
 Smallest number of factors with p > 0.1 2

Percent Variation Accounted for
by Partial Least Squares Factors

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	38.8223	38.8223	37.9603	37.9603
2	16.8383	55.6606	23.3863	61.3466

Notice that we achieve a similar proportion of variation explained for the dependent variable, but not as much as for the `model effects`, which are simply the variation in the predictors. In the PLS model we achieve almost the same amount of predictive ability for \mathbf{Y} with only two factors, rather than with 4, as were required in the PCR.

4 Predicting the Unknown using PCR or PLSR

For both procedures we merely add the observations with unknown response variables to our training data and follow through the procedures as outlined. We can then print out the output data set we created for the PLS procedure to see the predicted response and the `xscores` upon which it is based.

SAS Code and Output

```
data UScrimePred;
input R Age So Ed Ex0 Ex1 LF M N NW U1 U2 W X;
datalines;
. 120 0 110 115 116 500 966 101 106 77 35 657 170
;
run;
data UScrime;
set UScrime UScrimePred;
run;
proc pls data=UScrime nfac=4 details method=pcr;
  model r = Age So Ed Ex0 Ex1 LF M N NW U1 U2 W X/solution;
  output out=outpls
          predicted = yhat
          yresidual = yres
          xscore    = xscr;
run;
proc print data=outpls;
var yhat yres xscr1-xsc4; run;
```

Obs	yhat	yres	xscr1	xscr2	xscr3	xscr4
1	75.924	3.1756	-0.24286	0.13707	0.06811	0.02217

2	117.243	46.2569	0.06282	-0.06426	0.02935	0.16703
3	54.949	2.8510	-0.24166	-0.00185	0.02556	-0.05942
4	169.873	27.0272	0.22183	0.22415	-0.08492	0.23056
5	103.193	20.2066	0.11350	-0.12296	-0.12852	0.02685
6	96.025	-27.8249	0.17612	0.05652	-0.11088	-0.13832
...						
45	39.819	5.6808	-0.17995	0.14333	0.36370	-0.16742
46	97.485	-46.6846	0.11104	-0.05909	-0.13388	-0.03909
47	117.491	-32.5908	0.15255	-0.15837	0.18787	0.19867
48	106.727	.	0.13120	0.21337	-0.12553	-0.11800

If observations with an unknown response are taken at a future date, they can simply be merged with the earlier data and the procedure run again to obtain the predictions.

The `solution` and `detail` options in the above code will output the weights, the loadings, and the regression coefficients. The regression coefficients are given as VQ rather than as Q and so there is a regression coefficient listed for each variable. This can aid in interpretation. Examination of the coefficients can suggest the relative importance of a variable. Low absolute values for a given variable suggest consideration for deletion.

The `inner regression coefficients` shown with the weights are Q . Hence, if you have the `xscores` you can calculate the final regression by hand, though there is usually no reason to ever do this. It's better to let SAS do this for you.

SAS Ouput

```

                                The PLS Procedure

                                Model Effect Loadings

Number of
Extracted
Factors      Age          So          Ed          Ex0          Ex1          LF
-----
1      -0.285757    -0.261075    0.310817    0.418482    0.418563    0.138876
2       0.327895     0.445523   -0.304032    0.170797    0.159147   -0.161348

                                Model Effect Loadings

Number of
Extracted
Factors      M          N          NW          U1          U2          W
-----
1       0.084454     0.213983   -0.194323    0.019083    0.068576    0.407020
2      -0.099071     0.242554    0.537925   -0.117661    0.125406   -0.166181

                                Model Effect Loadings

                                Number of
                                Extracted
                                Factors          X
-----
1          -0.350075
2           0.325314

```

Model Effect Weights

Number of Extracted Factors	Age	So	Ed	Ex0	Ex1	LF
1	-0.082841	-0.083919	0.298907	0.636640	0.617298	0.174868
2	0.520279	0.460223	-0.182536	0.184227	0.148522	-0.005707

Model Effect Weights

Number of Extracted Factors	M	N	NW	U1	U2	W
1	0.198059	0.312461	0.030183	-0.046737	0.164178	0.408610
2	0.164929	0.070197	0.512653	-0.130713	0.140082	-0.207479

Model Effect Weights

Number of Extracted Factors	X	Inner Regression Coefficients
1	-0.165755	0.274254
2	0.519382	0.326858

Dependent Variable Weights

Number of Extracted Factors	R
1	1.000000
2	1.000000

Parameter Estimates for Centered and Scaled Data

	R
Intercept	0.000000000
Age	0.1333412073
So	0.1132334078
Ed	0.0728162855
Ex0	0.3423835367
Ex1	0.3221406868
LF	0.0756385298
M	0.1416909793
N	0.1614313412
NW	0.1809422604
U1	-.0634389897
U2	0.1185528198
W	0.1132851961
X	0.0962994772

Parameter Estimates

R

Intercept	-353.4358161
Age	0.4103510
So	9.1433673
Ed	0.2517442
Ex0	0.4455779
Ex1	0.4455870
LF	0.0723901
M	0.1859711
N	0.1639970
NW	0.0680565
U1	-0.1360926
U2	0.5429173
W	0.0454079
X	0.0933552

5 Appendix

5.1 Introduction to Bilinear Modeling using Matrix Algebra

Given a sample of size l , let \mathbf{X} and \mathbf{Y} represent the centered and scaled input data where \mathbf{X} is a high-dimensional (l by p , with $p \geq l$) predictor matrix and the response vector (l by 1) indicating the level or classification of the observation, respectively. Then, the data-compressed model becomes the linear model

$$\mathbf{T} = \mathbf{X}\mathbf{V} \quad (1)$$

and

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}' + \mathbf{F} \quad (2)$$

where \mathbf{F} represents the variation in \mathbf{Y} that cannot be explained by the model, the matrix \mathbf{Q} represents the regression coefficients of \mathbf{Y} on \mathbf{T} , and $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_m\}$ for $m \leq p$. The problem for multivariate calibration is to find the best estimates $\hat{\mathbf{V}}$ and $\hat{\mathbf{Q}}$, i.e., those that minimize the residual sums of squares of the elements in \mathbf{F} , which is assumed to have mean zero, and where with high probability, $\hat{\mathbf{Y}}$ is close to any future, unknown \mathbf{Y} given any future measured \mathbf{X} . For a more complete discussion of the above issues and those to follow in this section, see Martens and Naes, 1989.

In bilinear modelling we have a model structure $(\mathbf{X}, \mathbf{Y}) = h(\mathbf{T}) + (\mathbf{E}, \mathbf{F})$. For centered and scaled data, as above, we have the model:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \quad (3)$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}' + \mathbf{F} \quad (4)$$

where $\mathbf{T} = \mathbf{X}\mathbf{V}$ as above. This equation looks odd because \mathbf{X} occurs on both sides of the equation. However, at this point we should think of the model as if we have \mathbf{T} , but do not have \mathbf{V} separately. We call the matrix \mathbf{P} the loading matrix and it represents the regression coefficients of \mathbf{X} on \mathbf{T} . Hence, during model fitting $\hat{\mathbf{T}}$ (or more precisely, $\hat{\mathbf{V}}$) is chosen according to its relevance to both \mathbf{X} and \mathbf{Y} . The residuals for these regressions, \mathbf{E} and \mathbf{F} , represent the variation not explained by the bilinear model. The underlying assumption of bilinear modelling is that the calibration data are generated by a combination of physical or mechanistic relationships between the variables and the response and a probability distribution which includes the existence of a degree of collinearity in the variables. The collinearity can be estimated and accounted for via data compression and the mechanistic relationship, while possibly not well characterized, can be elucidated and estimated from the data themselves.

The ‘best’ $\hat{\mathbf{V}}$, $\hat{\mathbf{P}}$ and $\hat{\mathbf{Q}}$ can be found using several methods. In general, the estimated \mathbf{V} matrix, $\hat{\mathbf{V}}$, is found by optimizing some set criterion that defines the method. Then the scores $\hat{\mathbf{T}} = \mathbf{X}\hat{\mathbf{V}}$ are computed. Once the scores have been found, the loadings \mathbf{P} and \mathbf{Q} are estimated by the multiple least squares regression of each individual variable, \mathbf{x}_k on the scores. That is

$$\hat{\mathbf{P}}' = (\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}'\mathbf{X} \quad (5)$$

$$\hat{\mathbf{Q}}' = (\hat{\mathbf{T}}'\hat{\mathbf{T}})^{-1}\hat{\mathbf{T}}'\mathbf{Y}. \quad (6)$$

Here we discuss two methods of bilinear modelling: principal components regression (PCR), and its extension (and improvement), partial least squares regression (PLSR). For this chapter, we are concerned with classification rather than regression, but the former is a generalization of the latter. In chapter 2, we will consider the regression case. For now, we first review PCR, by way of introduction, since most readers will be familiar with this analysis technique, and PLSR is then an easy extension of the same principles.

In PCA, orthogonal linear combinations of the predictor variables (components) with maximal variance are constructed sequentially, i.e.,

$$\mathbf{v}_k = \mathbf{argmax}\{var^2(Xv)\} \forall \mathbf{v} \in \mathbf{V} \quad (7)$$

subject to the constraint

$$\mathbf{V} = \{\mathbf{v} : \mathbf{v}'\mathbf{S}\mathbf{v}_j = 0, 1 \leq j \leq k\} \quad (8)$$

i.e., the k th component, $\mathbf{X}\mathbf{v}_k$, is orthogonal to all previously obtained components, where $\|\mathbf{v}\| = 1$ and where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The number of components can be no more than the rank of \mathbf{X} . Here $\mathbf{argmax}\{f(\mathbf{x})\}$ is the value x that maximizes $f(x)$. Usually, the analysis is performed on the PCA correlation matrix $\mathbf{R}_{p \times p} = (1/(l-1))(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}})'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$ to eliminate affects due solely to the scale of the predictors. The components with maximal variance that satisfy the constraint are obtained in the spectral decomposition of \mathbf{R} . That is, we find $\mathbf{\Delta}$ and \mathbf{V} such that

$$\mathbf{R} = \mathbf{V}\mathbf{\Delta}\mathbf{V}' \quad (9)$$

where $\mathbf{\Delta}$ is a diagonal matrix holding the eigenvalues of R , i.e., $\mathbf{\Delta} = \mathit{diag}\{\gamma_1 \geq \dots \geq \gamma_p\}$, and where $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_p)$ are the corresponding eigenvectors. The i^{th} PC is the linear combination $\mathbf{X}\mathbf{v}_i$. Recall the definition of an eigenvalue: given a matrix \mathbf{A} if there exist scalars λ and vectors $\mathbf{x} \neq 0$ such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \quad (10)$$

then any scalars and vectors satisfying this equation are called the eigenvalues and eigenvectors of \mathbf{A} , respectively. Conceptually, since the components are these unique linear combinations of \mathbf{X} , they contain as much information as the original p predictors. Hence, the compression (or weight) matrix \mathbf{V} is estimated by optimizing the variance of the linear combinations of \mathbf{X} as described above to yield the compressed data $\hat{\mathbf{T}} = \mathbf{X}\hat{\mathbf{V}}$. Once $\hat{\mathbf{V}}$ has been found, we can perform a PC regression by using the principal components in the regression of \mathbf{Y} , that is $\mathbf{Y} = \hat{\mathbf{T}}\hat{\mathbf{Q}} + \mathbf{F}$, where $\hat{\mathbf{Q}}$ is the estimated regression coefficient matrix.

Note that PCA chooses the components without respect to the information contained in the response. Hence, we can perhaps improve upon PCA by seeking components that not only explain as much of the information in the predictors as possible, but also take into account how well the components predict Y . PLS sequentially maximizes the covariance between the response variable and a linear combination of the predictors. In PLS we seek a weight vector \mathbf{w} satisfying

$$\mathbf{w}_k = \mathbf{argmax}\{cov^2(\mathbf{X}\mathbf{w}, \mathbf{y})\} \forall \mathbf{w} \in \mathbf{W} \quad (11)$$

subject to the constraints

$$\mathbf{W} = \{w : \mathbf{w}'\mathbf{S}\mathbf{w}_j = 0 \forall j \text{ where } 1 \leq j \leq k\} \quad (12)$$

and $\|\mathbf{w}\| = 1$. As before, the maximum number of components is the rank of \mathbf{X} . The i th PLS component is also a linear combination of the original predictors, $\mathbf{X}\mathbf{w}_i$, but was obtained by taking into account information in the predictors and its correlation with the response. Once the \mathbf{w}_i have been found, we compute $\hat{\mathbf{T}} = \mathbf{X}\hat{\mathbf{W}}$, then estimate the loadings by regressing \mathbf{X} on $\hat{\mathbf{T}}$, i.e., $\mathbf{X} = \hat{\mathbf{T}}\hat{\mathbf{P}}' + \mathbf{E}$, then estimate the loadings for \mathbf{Y} using the regression $\mathbf{Y} = \hat{\mathbf{T}}\hat{\mathbf{Q}} + \mathbf{F}$. This process is in practice an iterative one, as is PCR. The above discussion does not address certain implementation and algorithmic issues. For a more complete discussion, see Martens and Naes, 1987. Once $\hat{\mathbf{T}}$ has been identified, the investigator can use some criteria for choosing the number of components necessary for the best model. For example, if a large percentage of the variation is explained using only three components, with little improvement using four, then the investigator might choose $\hat{\mathbf{T}} = \{\mathbf{X}\mathbf{w}_1, \mathbf{X}\mathbf{w}_2, \mathbf{X}\mathbf{w}_3\}$, and proceed with the bilinear modelling as discussed above. In practice there are several techniques for choosing the number of components. Several methods can be performed automatically in SAS.

5.2 Bibliography

- Afifi, Abdelmonem, Virginia Clark, Susanne May (2004) *COMPUTER-AIDED MULTIVARIATE ANALYSIS*, 4th ed. Chapman & Hall/CRC.
- Lovett, G.M., Weathers, K.C. & Sobczak, W.V. (2000) *Nitrogen saturation and retention in forest watersheds of the Catskill Mountains, New York*. *ECOLOGICAL APPLICATIONS* **10**:73-84.
- Naiman, R.J., Pinay, G., Johnston, C.A. & Pastor, J. (1994) *Beaver influences on the long-term biogeochemical characteristics of boreal forest drainage networks*, *ECOLOGY* **75**:905-921.
- Quinn, Gerry P. and Michael J. Keough 2002. *EXPERIMENTAL DESIGN AND DATA ANALYSIS FOR BIOLOGISTS*. Cambridge University Press.
- McCullagh, P. and J.A. Nelder 1998. *GENERALIZED LINEAR MODELS*. Chapman & Hall/CRC.
- Thomson, A. & Randall-Maciver, R. (1905) *ANCIENT RACES OF THE THEBAID*, Oxford: Oxford University Press.
- Weber, A. (1973) *AGRARPOLITIK IM SPANNUNGSFELD DER INTERNATIONALEN ERNAEHRUNGSPOLITIK*, Institut fuer Agrarpolitik und marktlehre, Kiel.