# Optimised score plot by principal components of predictions.

Ø. Langsrud[*] and T. Næs[*,+]

* MATFORSK, Osloveien 1, N-1430 Ås , Norway
+ Division of Statistics, Department of Mathematics, University of Oslo, Norway

## Abstract

A common problem in statistics/chemometrics is to relate two data matrices (*X* and *Y*) to each other, with the purpose of either prediction or interpretation. Usually one is interested in understanding which directions in *Y*-space that can be predicted by which directions in *X*-space. Several methods exist for this, for instance PLS regression and canonical correlation. The present paper presents a new plot for visualising the relationship between *X* and *Y*. The plot can be used for any regression method.

## Introduction

A common problem in statistics/chemometrics is to relate two data matrices, usually denoted by *X* and *Y*, to each other. The purpose is usually either prediction or interpretation. Several methods exist for finding such relationships, both linear, non-linear, parametric and non-parametric. In this paper we will concentrate on methods, which relate linear combinations of *X* to *Y* or to linear combinations of *Y*. The most commonly used methods in this group of techniques are canonical correlation analysis (Mardia *et al.*, 1979), multivariate linear regression, principal component regression and partial least squares regression (Martens and Næs, 1989), but a number of alternative methods and variants also exist. Most of these methods also provide plotting tools to facilitate interpretation of the data. An important question is whether these established plots are the most natural to use for interpretation of the relationship between X and Y.

In this paper we propose a new plotting technique for multivariate linear regression models. The plot can be used for any regression method, both for those mentioned above and for all other linear methods, and is based on the final validated model relationship between *X* and *Y*. The main idea behind the new plot, based on what we call principal components of predictions (PCP), is to find the most dominating directions of *Y* that can be predicted from *X* (usually denoted by $\hat{Y}$), and to plot these directions together with information about which subspace of *X* these directions correspond to. The procedure goes as follows:

- First build a predictor of *Y* based on *X* using for instance one of the methods mentioned above. Validate the usual way by either cross-validation (CV) or prediction testing.
- Then run PCA on the predicted *Y*-values. The first two components represent the two main directions of the part of *Y* that can be predicted from *X* (the same for three etc.). PCP scores and *Y*-loadings are found directly from this decomposition.
- Finally calculate the corresponding *X*-loadings and make plots of scores and loadings the regular way.

Note that this plot is not meant to replace regular PLS and PCR plots, it is rather to be considered complimentary. We propose that the regular plots are still used for diagnostic procedures, detection of non-linearities etc. The present plot focuses on the validated relationship and is therefore primarily meant for visualising the final estimated model relationship between X and Y. All good models, whether they are based on PLS, PCR or any other method will therefore have almost identical PCP plots. We hope this plot can be used to

gain extra insight into how X and Y are related and that it can solve some of the controversy regarding what type of plot that should be used for interpretation.

In the next section we will describe the mathematics of the plot and show how it is related to regular tools such as principal components and projections. Then we will present an example based on near infrared (NIR) and sensory analysis of sausages. Finally we present some concluding remarks and give the MATLAB code for the plot.

## The optimised score plot

Assume that we are in a situation where a number of response variables ($Y$) can be adequately related linearly to as set of explanatory variables ($Y = XB + E$). Let us further assume that the regression model is estimated. Any method can be used for the estimation, but for many situations it is natural and advantageous to use a method like PCR or PLS which both handle collinearity problems among a large set of $X$-variables.

The optimised score plot proposed here is based on first decomposing the predicted values of $Y$ by the use of an singular value decomposition (SVD), i.e

$$\hat{Y} = USV^T = \sum_i u_i s_i v_i^T \quad (1)$$

This decomposition has the property that for any dimension $r$, defined by the first $r$ elements in the sum above, account for as much as possible of the variation of $\hat{Y}$ that can be described by an $r$ dimensional decomposition. Thus, the $r$ first components will then describe as much as possible of that part of $Y$ that can be predicted by $X$. The $U$-matrix is the matrix of PCP scores. Accordingly,

$$C = VS \quad (2)$$

is the matrix of PCP $Y$-loadings. Note that this matrix can be obtained by regressing the $\hat{Y}$ matrix onto the scores $U$. Note also that since the predicted $\hat{Y}$-values can be written as linear functions of $X$ ($\hat{Y} = X\hat{B}$), the PCP components can also be written as linear functions of $X$:

$$U = X\hat{B}VS^{-1} \quad (3)$$

We may then want to see how the original $X$ projects onto a space of PCP components. In other words, we will calculate the PCP $X$-loadings (as for the $Y$-loadings) by regressing $X$ onto $U$. This can be expressed as

$$P = X^T U \quad (4)$$

So, we now have a set of scores and loadings, which can be plotted and interpreted in a similar way as scores and loadings calculated from e.g. PLS or PCR.

In for instance PLS the explained $Y$-variance associated with each component is usually calculated by cross validation. To calculate comparable measures for the PCP components, we have to use PCP as a new regression method (but this is not the main purpose of PCP). To calculate scores for a new $X$-observation, we use the weights that are presented in equation (3). Thereafter, these scores (for a fixed number of components) are multiplied by the transpose of the $Y$-loadings in (2) to obtain the predicted $Y$-values. Note that this special regression method may be useful for prediction purposes in some cases. In fact, reduced rank regression (Davies and Tso, 1982) is a special case, which is obtained when the basis for PCP is classical multivariate regression.

A special case of PCP occurs when there is only one response variable. Only one PCP component can be extracted. In this case, we suggest computing the second component as the

one that maximises the explained *X*-variance orthogonal to the single PCP component. This procedure results in a nice score plot where the predicted values ($\hat{Y}$) can be read directly from the first axis.

## Example

The example is taken from production of sausages. Fifty-seven different types of sausages were produced (according to an experimental design) and they where measured by both near infrared (NIR) reflectance spectroscopy and by sensory analysis. In total 16 sensory attributes were measured by a trained sensory panel. The main focus in the present paper is to consider the relationship between the NIR data (=*X*) and the sensory data (=*Y*). We will focus on differences and similarities between plots rather than on the interpretation itself. A detailed description of this data set is given in Ellekjær *et al*. (1994).

In Figure 1 is presented the NIR data from the experiment.

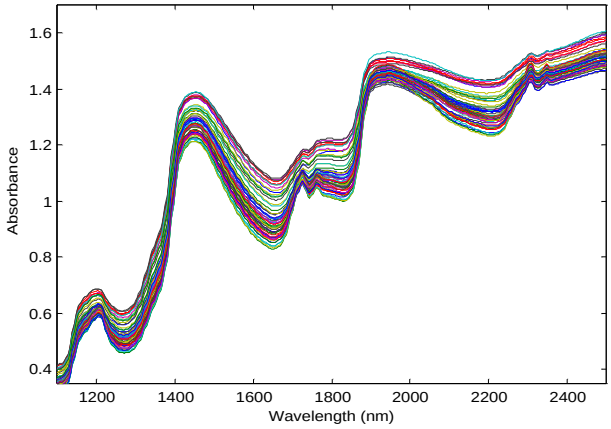
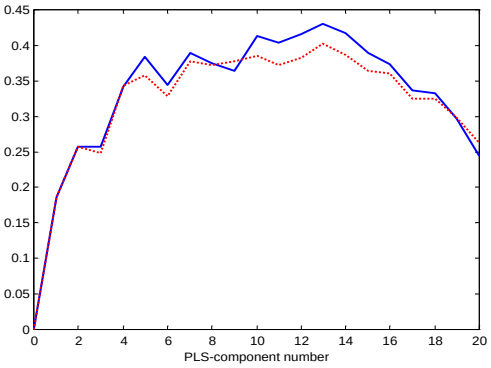
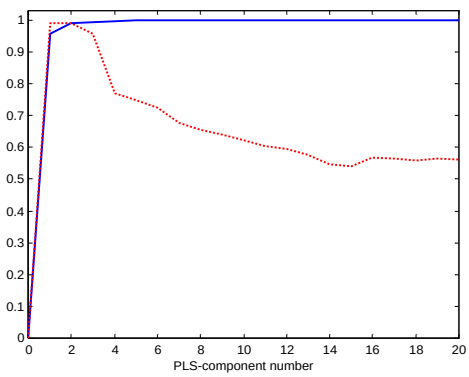



Figure 1. NIR data from the experiment

In Figure 2 is shown the cross-validated explained variance for PLS and PCP based on two components. The PCP results show how much of the Y-variance that is explained by basing the PCP on the PLS solution indicated by the number of components given. As can be seen, as many as 13 PLS components are needed in order to obtain the best possible results. A reasonable result is however obtained already after 5 PLS components. We can see that the two lines follow each other closely. This indicates that the part of the *Y*-data that can be predicted by *X* is essentially two-dimensional. The large number of components needed to explain *Y* must come from the complex relationship between the NIR and the sensory data. Figure 3 presents the corresponding cross-validated error measures for the *X*-data.




Figures 2 and 3: *Y* and *X*-variance respectively (CV); explained variance for PLS (———) and for PLS followed by two-component PCP (--------).

The PCP scores, for the optimal PLS model, are compared with the PLS scores in Figure 4. In order to facilitate comparison, the scores are standardised to equal variance (note that the description of PCP above already uses orthonormal scores) and they are matched by using so-called procrustes rotation (Gower, 1975). This is done in order to focus on the important differences and to avoid interpreting differences due to differences in rotation only. Straight lines connect corresponding scores. The numbered end corresponds to PCP. As can be seen, there are substantial differences between the scores. The change from one plot to the second seems to be rather chaotic, meaning that the two plots will give a quite different first impression.

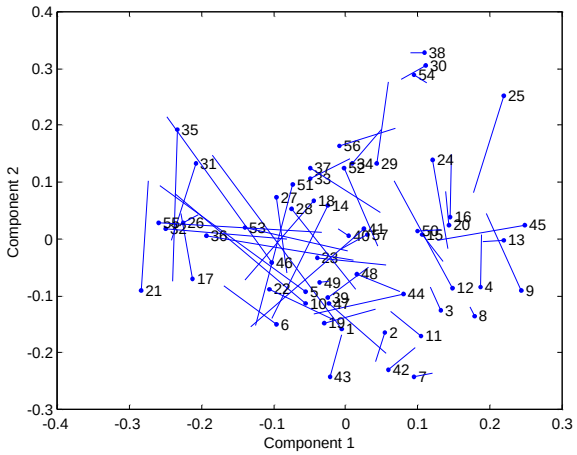



Figure 4: Scores for PLS and rotated PCP (The numbered end of line is rotated PCP)

The *Y*-loadings are presented in Figure 5. Here, the differences are smaller than they are for the scores. The general trend is that the PCP loadings are further away from the centre. This corresponds to the fact that the amount of explained variance with two components is larger for PCP than it is for PLS.

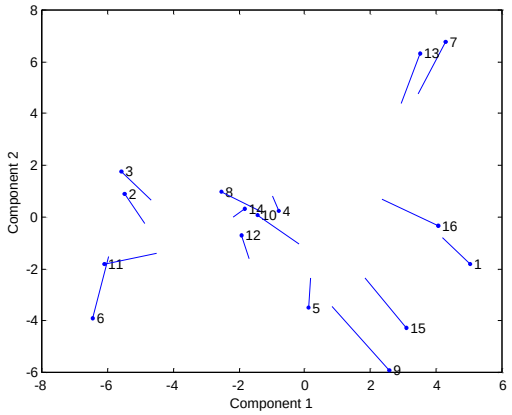
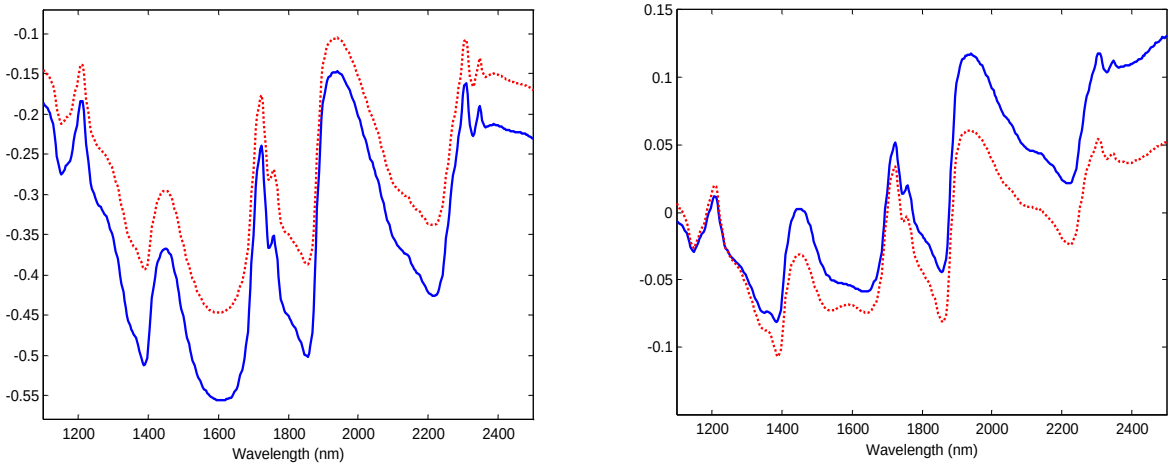



Figure 5: The *Y*-loadings for PLS and PCP

The *X*-loadings are given in Figure 6 and 7. The loadings are different, but still they represent almost the same spectral shapes.




Figures 6 and 7: *X*-loadings for components 1 and 2; PLS (———) and PCP (--------).

The same plots were made using PCR instead of PLS. The best PCR model gave almost as good results as the optimal PLS model. The PCP results based on PCR were very similar to those obtained by PLS. Figure 8 compares the PCP scores based on PCR to those from PLS (Procrustes rotation is used as above). The two sets of scores are not identical, but apart from a few objects, the differences are negligible compared to PCP vs PLS (Figure 4). There are also small differences between the *X* and *Y*-loadings from the two PCP decompositions (not shown) and together these results illustrate the unifying potential of the present methodology.

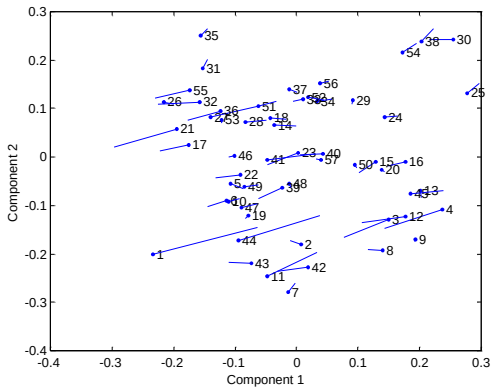
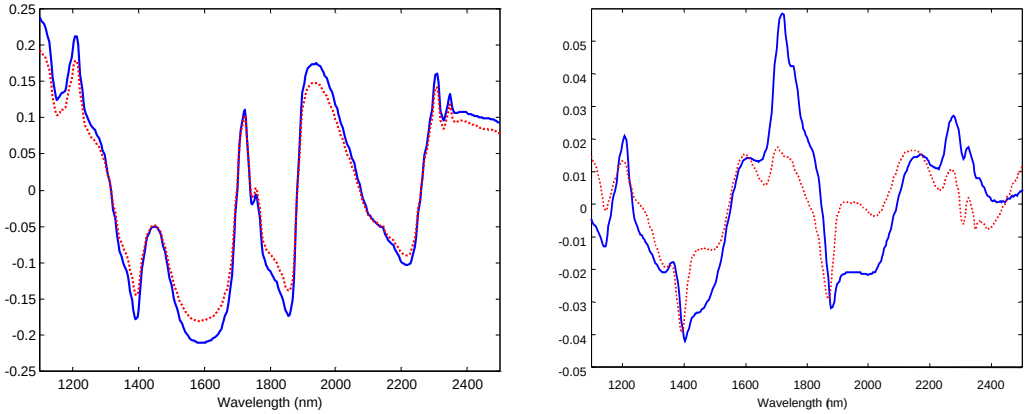



Figure 8. Comparison of PCP components based on PLS and PCR.

In this data set the predictions can be improved using multiplicative scatter correction (Geladi *et al.*, 1985). In that case, comparison between PLS and PCP shows much of the same tendencies as above, but a change is that differences along the first PLS dimension are less important. The two first *X*-loadings from PLS and rotated PCP are shown in Figures 9 and 10. The loadings for component 1 are very similar (Figure 9). Figure 10, however, shows that there are some loading differences that may be of some importance for component 2.




Figures 9 and 10: *X*-loadings for components 1 and 2 when the spectra are scatter corrected; PLS (———) and PCP (--------).

## Concluding remarks

If the optimal predictor in a multiresponse situation is found by PLS or PCR with two components, the new PCP score plot will obviously be essentially the same as the traditional plots. The new plot will simply be a rotated version. As was observed above, in a situation where several PLS components are needed, the scores plots may be very different from each other. In other examples considered (not reported here) we have also found other differences between PCP and PLS plots. If the optimal PLS predictor is found by one component, then the first PCP component will equal the first PLS component.

The new decomposition can be based on any regression method. The above procedure assumes that there are at least 2 dimensions in the $\hat{Y}$-data. If this is not the case, i.e. if there is only on *Y*-variable or the practical rank of $\hat{Y}$ is 1, additional components can be extracted by optimising *X*-variance instead. This methodology is of special importance when there is one response variable only. The fitted values are simply plotted along the first axis and the second axis gives additional information about the *X*-data.

Note that the new plot can be of special importance for process monitoring. Interesting quality parameters (*Y*) may for instance be measured indirectly by some online instrument (*X*). If a regression model is estimated from historical data, one can extract PCP scores for new observations as indicated above. The process can be monitored as a PCP score plot and the two dimensions are optimal with respect to quality information.

## Appendix: MATLAB code for the PCP decomposition

The decompositions of X and Y is written as

$$X = (XR)P^T + F = TP^T + F \quad (5)$$

$$Y = (XR)C^T + E = TC^T + E \quad (6)$$

where *R* is a matrix of weights made so that $T^TT = I$, and where $C = Y^TT$ is the Y-loadings and $P = X^TT$ is the X-loadings. The following code starts from the matrix of estimated regression parameters (B) together with the *X*-matrix (X).

```
[U,S,V]=svd(X*B,0)               % SVD of Yhat
r1 = rank(S)                     % rank of Yhat
T=U(:,1:r1)                      % scores
R=B*V(:,1:r1)*inv(S(1:r1,1:r1))  % weights
P=X'*T                           % X-loadings
C=V(:,1:r1)*S(1:r1,1:r1)         % Y-loadings
r2 =rank(X) - r1                 % Complete the
if r2 > 0                        % decomposition
  [U_,S_,V_] = svd(X-T*P',0)
  T=[T,U_(:,1:r2)]
  R=[R,(V_(:,1:r2)-R*(P'*V_(:,1:r2)))*inv(S_(1:r2,1:r2))]
  P=[P,V_(:,1:r2)*S_(1:r2,1:r2)]
  C=[C,zeros(size(C,1),r2)]
end
```

The first six lines make the *ordinary* PCP decomposition based on PCA (or SVD) of $\hat{Y}$. The other lines complete the decomposition of $X$ ($F$=0 in (5)) by running PCA on the $X$-residuals ($X - TP^T$). Note that the second line may be changed so that a *practical rank* is used instead of the *numerical rank.*

## References


Ellekjær, M.R., Isaksson, T., Solheim, R. (1994), Assessment of Sensory Quality of Meat Sausages Using Near Infrared Spectroscopy, *Journal of Food Science*, **59**, 456-464.

Davies, P. T. and Tso, M. K-S. (1982), Procedures for Reduced-rank Regression, *Applied Statistics*, **31**, 244-255

Geladi, P., MacDougall, D. and Martens, H. (1985), Linearization and scatter-correction for NIR reflectance spectra of meat, *Applied Spectroscopy*, **39**, 491-500.

Gower, J. C. (1975), Generalized Procrustes Analysis, *Psychometrika*, **40**, 33-51.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

Martens, H. and Næs, T. (1989), *Multivariate Calibration*, New York: John Wiley and Sons.