Module 3: Partial Least Squares

3.1 The mathematics behind Partial least squares regression with one response-variable (PLS1)

The method of partial least squares (PLS) is an algorithm, or rather several related algorithms, originally invented by Herman Wold and developed in the field of chemometrics by several others (S. Wold, H. Martens and T. Næs, among others). We deal here only with the PLS1 algorithm which is the (main) one used for prediction of one response variable, \( Y \), on the basis of a set of explanatory variables, \( x_1, \ldots, x_p \). A PLS2 algorithm exists for prediction of several response variables on the basis of the same set of explanatory variables, but it plays a less important role, since a reasonable alternative is simply to predict the response variables one at a time. A precise mathematical description and discussion of the PLS-algorithms and references to the literature on the subject may be found in Helland (1990).

As a typical example of application of the PLS1 algorithm (in the sequel just denoted PLS) one may think of the prediction of some property of a food sample, like protein content, on the basis of a NIR-spectrum (Near Infrared Reflectance). Each \( x \)-variable then corresponds to the NIR measured at a particular wavelength and there may well be in the order of \( p = 100 \) recorded wavelength in the entire NIR-spectrum.

The PLS-algorithm is designed to deal with situations with many variables compared to the number of observations, as is typically the case with an example like the NIR-measurements above, or other cases where the \( x \)-variables represent readings from some continuous spectrum. It has not in any strict way been proven that the PLS algorithm does well in such (or other) situations, the documentation lies in the examples where it has worked. Experience seems to suggest that the PCR (principal components regression) and PLS methods are quite similar in performance although the underlying algorithms are quite different.

Let it be emphasized once and for all that the PLS algorithm (like, for example, PCR and MLR) is a linear method, that is, it results in a linear predictor,

\[
\hat{y} = a_0 + a_1 x_1 + \cdots + a_p x_p.
\]

The problem is to estimate a good set of coefficients \( (a_1, \ldots, a_p) \).

The estimation of the constant \( a_0 \) presents no problem. Usually one gets rid of this by centering all \( x \)-variables as well as the \( y \)-variable.
3.1 The mathematics behind Partial least squares regression with one response-variable (PLS1)

Actually we also need the constant term \((a_0)\), but that is the easiest part. If, as the very first step, the \(x\)-variables are centered, that is, their means (over all the observations) are subtracted, we should simply choose the constant as the mean of the response variable, \(a_0 = \bar{y}\), or equivalently center \(y\) also and omit \(a_0\). In the sequel we assume that all variables have been centered so that \(a_0\) is not needed.

If there are \(n\) observations and \(p\) \(x\)-variables, and if \(p > n\), ordinary multiple linear regression (MLR) breaks down because there are infinitely many sets, \((a_1, \ldots, a_p)\), of exact solutions to the \(n\) available equations. In fact, even if \(p\) is not substantially smaller than \(n\) MLR may lead to extremely uncertain predictors because of the uncertainty of the estimates. PLS has been claimed to work for such situations and there is no doubt that it does in several cases, although there is no guarantee and it is easy to construct cases where it performs poorly. The situation is somewhat paradoxical because the existence of the infinitely many exact solutions cannot be denied, but the algorithm has to find some (non-exact) solution which should be better for prediction purposes.

Let us turn now to the specification of the algorithm. PLS constructs a number of components each of which is just a new variable which is a linear function of the \(x\)-variables, that is, a function of the form

\[
w = c_1 x_1 + \cdots + c_p x_p.
\]

Once a (usually low) number of such components have been constructed, say \(w_1, w_2\) and \(w_3\), an ordinary multiple linear regression of \(y\) on the components is performed, leading to a predictor of the form

\[
\hat{y} = b_1 w_1 + b_2 w_2 + b_3 w_3,
\]

which may then be rewritten in terms of the \(x\)-variables. Two problems remain: the (successive) construction of the components, and a criterion for the number of components to use.

Consider now the successive construction of the components \(w_1, w_2, \ldots\), and to be even more specific, consider the construction of the first PLS-component \(w_1\). To obtain that we have to obtain the \(c\)-coefficients. If we perform a simple linear regression analysis of the \(j\)th variable \(x_j\) on \(y\) we get a relation of the form

\[
\hat{x}_j = \hat{\alpha} + \hat{\beta} y.
\]

The slope, \(\hat{\beta}\), from this equation is denoted \(\hat{\beta}_j\). Now choose \(c_j = \hat{\beta}_j\) and you have the first PLS-component. (Disregard, for the moment, whether this is reasonable or not — this is the way it works.)

The next step is to use the component we have just constructed to predict \(y\), as in the regression \(\hat{x}_j\). If we obtain a good fit (typically judged by cross validation) we stop, otherwise we continue to construct the second component. To do this take the residuals from the regression \(\hat{x}_j\) on the first component, and let us name this new variable \(y_1\). This may be thought of as the part of \(y\) which is not explained by the first PLS-component. We now construct \(w_2\) in the same way as \(w_1\), except that we use \(y_1\) instead of \(y\). Now we have two components and can perform the regression \(\hat{x}_j\) with two predictors, judge the quality of the fit, stop here or calculate the residuals, use these to construct the third component, and so on. That’s it!
3.1 The mathematics behind Partial least squares regression with one response-variable (PLS1)

The PLS-calculations can be performed ‘semi-manually’ by use of any multiple linear regression program, but it is, of course, a lot easier if you have a PLS-routine available.

In most places the choice of coefficients (the $c_j$’s) are written in another way, namely as the estimate of $\text{Cov}(X_j, Y)$. Then the entire set of $c$-coefficients will be multiplied by the factor $\text{Var}(Y)$ compared to the definition above, but since this is the same constant for all the coefficients it simply means that the component is multiplied by a constant. When it is used as a predictor in the regression (3.1) this does not matter because the regression coefficient will automatically be scaled to compensate for this.

A modification which is frequently used is that the $x$-variables are scaled to have unit variance before the algorithm is used. Especially when the variables are measured in non-comparable units this makes sense. If, however, the $x$-variables are originally on the same scale, like intensities in a NIR-spectrum, the choice is more difficult and both approaches can be reasonable.