Mathematical Appendix

A statistical framework to model the meeting-in-the-middle principle using metabolomic data: application to hepatocellular carcinoma in the EPIC study.

1 PLS regression

1.1 Introduction

PLS (partial least squares) regression is a widely used method in multivariate statistics to relate two sets of variables while reducing their dimensionality. It was first developed as a method to predict a set of variables Y from another set X; and also to depict their common structure. The main aim of PLS is to regress a set Y of $\mathbf{q}$ variables $(y_1, y_2, \ldots, y_q)$ of interest, which are called responses, on a set X of $\mathbf{p}$ predictor variables $(x_1, x_2, \ldots, x_p)$ that may display high levels of correlation. PLS combines and generalizes features of principal component analysis (PCA) and multiple linear regression (MLR); and results in a set of PLS latent factors as linear combinations of variables, in turn, in the X- and Y-sets. By simultaneously decomposing X and Y, PLS finds components that explain as much as possible of the inter-relations of X and Y. The latent factors obtained from the decomposition can be used to predict Y. The following details of the algorithm are adapted from Michel Tenenhaus' book *La régression PLS, Théorie et Pratique* [1].

1.2 The PLS algorithm

Two different, but closely related, techniques exist under the name of PLS regression. The canonical or symmetric PLS regression assumes that the X- and Y- sets play a symmetrical role. The version presented here is the regression mode where latent variables are computed from a succession of singular value decompositions (SVD) followed by deflation of both the X- and Y- matrices. These sets are assumed to play the asymmetric roles of predictors and responses, respectively. Next, we briefly describe the landmark algorithm NIPALS Nonlinear estimation by Iterative Partial Least Squares. As a first step, two substitute matrices X_0 and Y_0 are initialized with $X_0 = X_{(n\times p)}$ and $Y_0 = Y_{(n\times q)}$, where variables were standardized to have means and standard deviations equal to zero and one, respectively. For $h = 1, \ldots, H$, where $H = min(p, q)$, the PLS factors are obtained iteratively. PLS regression focuses on finding two sets of weights, $w_{h(p\times 1)}$ and $c_{h(q\times 1)}$, in order to create respectively a linear combination of the columns of X and Y, known as the PLS factors, such that these two linear combinations have maximum covariance and are unique. These weights define a first pair of vectors, called the X- and Y-scores, $t_h = Xw_h$ and $u_h = Yc_h$ where we have $t_h^\intercal u_h$ maximal. PLS can be written as the following optimisation problem where maximum covariance is sought between $t_{h(1\times n)}$ and $u_{h(1\times n)}$ for each $h = 1 \cdots H$:

$$\text{Max } cov(Xw_h, Yc_h) \tag{1}$$

under the following normality constraints

$$\|w_h\| = 1 \tag{2}$$

$$\|c_h\| = 1 \tag{3}$$

and the following orthogonality constraint

$$t_h^\intercal(t_1, \ldots, t_{h-1}) = 0 \tag{4}$$

By construction we also have the following property:

$$u_h^\intercal(t_1, \ldots, t_{h-1}) = 0 \tag{5}$$

The first pair of X- and Y- scores can equivalently be obtained via a singular value decomposition. Indeed, the SVD of the cross-product matrix $X_{h-1}^\intercal Y_{h-1}$ leads to the identification of the first left and right singular vectors and of the weights w_h and c_h. The scores t_h and u_h are obtained as follows:

$$t_h = X_{h-1}w_h \tag{6}$$

$$u_h = Y_{h-1}c_h \tag{7}$$

The vector t_h is then normalized (a scaling of u_h is optional). Regressing the predictor and response matrices on the t_h vector yields the corresponding loadings.

$$p_h = X_{h-1}^\intercal t_h \tag{8}$$

$$c_h = Y_{h-1}^\intercal t_h \tag{9}$$

Next is the deflation step, where information based on the extracted latent factor h is subtracted from the current data matrices.

$$X_h = X_{h-1} - t_h p_h^\intercal \tag{10}$$

$$Y_h = Y_{h-1} - t_h c_h^\intercal \tag{11}$$

The described steps of the algorithm are iterated until one of the following criteria is met:

- If H is specified, and the algorithm stops when the H-th PLS factor is extracted and its associated statistics computed.
- If H is not specified, the algorithm stops when X_H becomes a null matrix. In this case however, H cannot exceed $min(p, q)$.

Algorithm 1 PLS1 classic algorithm steps - When Y is univariate.

```
$X_0 \leftarrow X$ ; $y_0 \leftarrow y$
for $(h = 1; h \leq H; h++)$ do
    $w_h = {}^{X_{h-1}^\intercal y_{h-1}}\big/_{y_{h-1}^\intercal y_{h-1}}$
    $w_h = {}^{w_h}\big/_{\sqrt{w_h^\intercal w_h}}$
    $t_h = {}^{X_{h-1} w_h}\big/_{w_h^\intercal w_h}$
    $p_h = {}^{X_{h-1}^\intercal t_h}\big/_{t_h^\intercal t_h}$
    $X_h = X_{h-1} - t_h p_h^\intercal$
    $c_h = {}^{y_{h-1}^\intercal t_h}\big/_{t_h^\intercal t_h}$
    $u_h = {}^{y_{h-1}}\big/_{c_h}$
    $y_h = y_{h-1} - c_h t_h$
```

When Y is univariate, the PLS algorithm carried out is PLS1 (See Algorithm 1, following the notation of M. Tenenhaus [1]). PLS2 (Algorithm 2) is used when Y is multivariate. When there are missing data in either the X- or Y- sets, the coordinates of the vectors w_h, t_h, c_h, u_h, and p_h are computed as slopes of the least squares straight line that passes through the origin, using the available data as follows:

Algorithm 2 PLS2 classic algorithm steps - When Y is multivariate.

```
$X_0 \leftarrow X$ ; $Y_0 \leftarrow Y$
for $(h = 1; h \leq H; h{+}{+})$ do
    $u_h = Y_{h-1}[,1]$ i.e. the first column of the matrix
    while $w_h$ has not converged do
        $w_h = X_{h-1}^\intercal u_h / u_h^\intercal u_h$
        $w_h = w_h / \sqrt{w_h^\intercal w_h}$
        $t_h = X_{h-1} w_h / w_h^\intercal w_h$
        $c_h = Y_{h-1}^\intercal t_h / t_h^\intercal t_h$
        $u_h = Y_{h-1} c_h / c_h^\intercal c_h$
    $p_h = X_{h-1}^\intercal t_h / t_h^\intercal t_h$
    $X_h = X_{h-1} - t_h p_h^\intercal$
    $Y_h = Y_{h-1} - t_h c_h^\intercal$
```

- $w_h = (w_{h1}, \ldots, w_{hp})^\intercal$, is a normalized vector, where w_{hj} is the slope of the least squares line passing through the origin of the plane defined by $(u_h,\ X_{h-1,j})$. $X_{h-1,j}$ is the j-th X variable of the $h-1$ PLS factor.

- $t_h = (t_{h1}, \ldots, t_{hn})^\intercal$, where t_{hi} is the slope of the least squares line passing through the origin of the plane defined by $(w_h,\ x_{h-1,i})$. $x_{h-1,i}$ is the i-th x observation of the $h-1$ PLS factor.

- $c_h = (c_{h1}, \ldots, c_{hq})^\intercal$, where c_{hk} is the slope of the least squares line passing through the origin of the plane defined by $(t_h,\ Y_{h-1,k})$. $Y_{h-1,k}$ is the k-th Y variable of the $h-1$ PLS factor.

- $u_h = (u_{h1}, \ldots, u_{hn})^\intercal$, where u_{hi} is the slope of the least squares line passing through the origin of the plane defined by $(c_h,\ y_{h-1,i})$. $y_{h-1,i}$ is the i-th y observation of the $h-1$ PLS factor.

- $p_h = (p_{h1}, \ldots, p_{hp})^\intercal$, where p_{hj} is the slope of the least squares line passing through the origin of the plane defined by $(t_h,\ X_{h-1,j})$. $X_{h-1,j}$ is the j-th X variable of the $h-1$ PLS factor.

1.3 Tools for interpretation

1.3.1 Choice of number of components

The number of PLS latent factors or components to be retained can be decided based on a cross-validation.

For each model with a number h of extracted factors, this is done by running the PLS analysis on only a part of the data called the training set, and then evaluating how well the model fits observations in the test set. This includes the part of the data not involved in the PLS modelling of the training set.

The dataset comprised of n observations is split into z approximately equal sets of observations. The training set consists of the data in the first $z-1$ folds and the remaining fold is used as test set. Predicted values for the Y-set are computed on this test set along with the sum of the squared error of prediction. This process is repeated z times so that each fold can in turn serve as a test set. In practice, for each number of possible latent factors $h = 1, \ldots, H$, we compute the prediction of y_i by the PLS model with results obtained on the training set with a number h of components applied to observations in the test set in order to yield $\hat{y}_{h(-i)}$. The Prediction Error Sum of Squares (PRESS) is the resulting sum of all squared errors of prediction statistic computed across all test sets as defined in the following equation:

$$PRESS_h = \sum (y_i - \hat{y}_{h(-i)})^2 \tag{12}$$

The Residual Sum of Squares (RSS) is computed in a standard way:

$$RSS_h = \sum (y_i - \hat{y}_{hi})^2 \tag{13}$$

Different criteria can be used to determine the number of components h to retain. One such criterion, Q^2_h was first introduced by H. Wold [2] and is mainly used in the software SIMCA-P. It is based on the following statistic:

$$Q^2_h = 1 - \frac{PRESS_h}{RSS_{h-1}} \tag{14}$$

As pointed out by M. Tenenhaus, the initial value for RSS when y is univariate centred-scaled and $h = 0$ is:

$$RSS_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n - 1 \tag{15}$$

In the software SIMCA-P the PLS component is kept when the following condition is met:

$$\sqrt{PRESS_h} \leq 0.95\sqrt{RSS_{h-1}} \tag{16}$$

$$\Longleftrightarrow Q^2_h \geqslant 0.0975 \tag{17}$$

The default threshold 0.0975 is equal to $1 - 0.95^2$. In SAS, the criteria to select the number h of components to be retained is by minimizing the $PRESS_h$ statistic.

The above described formulae can be generalized for multivariate Y, thus we have for any given variable y_k, $k = 1, \dots, q$:

$$Q^2_{kh} = 1 - \frac{PRESS_{kh}}{RSS_{k(h-1)}} \tag{18}$$

$$Q^2_h = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} RSS_{k(h-1)}} \tag{19}$$

The criteria for keeping a PLS factor are identical to what was established for the univariate case. One can alternately use one of the following rules, where the equivalence defined in formula (17) still holds true:

- $Q^2_h \geqslant 0.0975$
- At least one value of $Q^2_{hk} \geqslant 0.0975$

If the criteria are met by several values of h, the one retained is the smallest h, to achieve a better dimensionality reduction.

The Q^2 and $PRESS$ criteria are relatively robust to the choice of number of folds (blocks) used for cross-validation. A number of folds between 5 and 10 is recommended (Tenenhaus 1998, p.238) [1]. The default choice in the SIMCA-P and SAS softwares is 7, and is the parameter used in this study.

1.3.2 Variable Importance in the Projection (VIP)

The Variable Importance in the Projection (VIP) is a measure of the explanatory power of a given variable x_j over Y. The VIP_{hj} of a given component h of the j-th variable x_j is defined as:

$$VIP_{hj} = \sqrt{\frac{p}{Rd(Y; t_1, \dots, t_h)} \sum_{l=1}^{h} Rd(Y, t_l) w^2_{lj}} \tag{20}$$

and one has:

$$\sum_{j=1}^{p} VIP^2_{hj} = p \tag{21}$$

where $Rd(Y; t_1, \dots, t_h)$ is the redundancy of Y with respect to the t scores $(t_1, \dots, t_h)$. It describes the amount of variance of Y explained by the component t_h of the X-set. It is defined as follows:

$$Rd(Y, t_h) = \frac{1}{q} \sum_{k=1}^{q} cor^2(y_k, t_h) \tag{22}$$

It can be equivalently computed as:

$$Rd(Y, t_h) = r_h^2 \frac{1}{q} \sum_{k=1}^{q} cor^2(y_k, u_h) \tag{23}$$

where $r_h = cor(Xw_h, Yc_h)$ is called a canonical correlation and r_h^2 is the h^{th} largest eigenvalue of the crossproduct matrix decomposition.
The contribution of a variable x_j to the construction of a component t_l is measured by the weight w_{lj}^2. For each l, with $l = 1, \ldots, h$, the sum of these weights across the p variables x_j equals 1. To measure the contribution of the variable x_j to the construction of Y through the components t_l, one should consider the explanatory power of the component t_l, measured by the redundancy $Rd(Y; t_l)$. An equal weight w_{lj}^2 indicates an explanatory power of the x_j variable over the Y-set whose importance increases with the level of redundancy $Rd(Y; t_l)$.
The VIP enables the ranking of the predictors x_j according to their explanatory power on Y, and summarizes their contribution to the model. A VIP is considered small if its value is less than 0.8 and high when its value is greater than 1. Variables with a high VIP ($VIP > 1$) are the most important for the reconstruction and prediction of Y.

2 Statistical Recoupling of Variables (SRV)

The SRV procedure was introduced by *Blaise et al.(2009)* [3] and for which a matlab toolbox was later implemented [4]. The SRV is an "intelligent bucketing" algorithm that aims at regrouping variables (typically the smallest unit of the NMR spectrum) in clusters corresponding to a wider biological and chemical entity.
SRV exploits the spectral structure of data, without forming any metabolic hypothesis to reduce the dimensionality of spectra. A typical NMR ^1H 9 ppm spectrum is often partitioned into $9,000$ buckets of 0.001 ppm width. The main idea of the algorithm is to exploit the spectral dependency landscape L which is the covariance to correlation ratio between two neighbouring variables along the chemical shift axis to assemble them within a cluster. If one considers a matrix Z of serum spectra acquired by NMR with n observations and r columns $(z_1, \ldots, z_r)$ corresponding to neighbouring bins of NMR signals. The first bin-variable starts the first

cluster, then L is computed for each z_i as follows with $i = 1, \ldots, r$:

$$\begin{aligned} L(z_i) &= \frac{cov(z_i, z_{i+1})}{cor(z_i, z_{i+1})} \\ &= sd(z_i) * sd(z_{i+1}) \end{aligned} \tag{24}$$

where sd is the standard deviation.

The variable then joins a cluster according to the following rules:

- $L(z_i)$ values are used to locate local minima i.e. borders between clusters.
- If $L(z_{i-1}) > L(z_i)$ then z_{i-1} and z_i are associated in the same cluster, otherwise z_i and z_{i+1} start a new cluster.
- The minimum number of variables belonging to a cluster is set a priori as it is based on the resolution of the NMR spectra. When acquired at 700 MHz, the typical peak base width of a well-resolved singlet is equal to 7 Hz. Therefore, the threshold was set to 10 in our analysis, meaning that if a cluster has less than 10 variables, it is discarded.
- The super-cluster intensity is computed as the mean of the intensities of the signal in the bins assigned to the super-cluster.
- If two neighbouring clusters have a correlation > 0.9, they are aggregated to form a super-cluster. In these analyses, the association is limited to 3 clusters per super-cluster (this value is empirical and was discussed in the original paper [3]).

References

[1] Michel Tenenhaus. *La régression PLS: Théorie et Pratique.* Paris, 1998.

[2] Herman Wold. *Multivariate Analysis*, chapter Estimation of principal components and related models by iterative least squares, pages 391–420. Academic Press, New York, 1966.

[3] Benjamin J Blaise, Laetitia Shintu, Bénédicte Elena, Lyndon Emsley, Marc-Emmanuel Dumas, and Pierre Toulhoat. Statistical recoupling prior to significance testing in nuclear magnetic resonance based metabonomics. *Analytical Chemistry*, 81(15):6242–6251, August 2009.

[4] Vincent Navratil, Clément Pontoizeau, Elise Billoir, and Benjamin J Blaise. Srv: an open-source toolbox to accelerate the recovery of metabolic biomarkers and correlations from metabolic phenotyping data sets. *Bioinformatics*, 29(10):1348–1349, May 2013.