# CHAPTER II

# LITERATURE REVIEW

## Chemometrics

*Chemometrics* can briefly be described as the interaction of certain mathematical and statistical methods to chemical problems. It has developed as a consequence of the change in the data obtained within chemistry with the emergence of new analytical techniques as well as microprocessors. The applications using chemometric techniques in analytical chemistry are now numerous and applications have been revealed in spectroscopy, chromatography and other disciplines of analytical chemistry. A major strength of chemometric techniques lies in their ability to find and extract information given large amounts of data. As mentioned above, with the development of analytical instruments the type of data has changed from being uni- and low -variate (≤ 2 variables) to truly become multivariate. The field of chemometrics has thus found its natural connection with analytical chemistry.

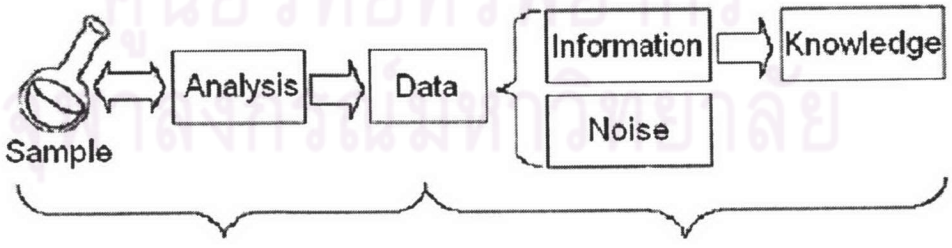
Chemometrics involves the use of multivariate calibration methods applied to spectroscopic data (*6,25,29-30*). These methods are classified according to the type of spectral information used and as to whether the calibration process is direct or inverse (Table 1). The common task of all multivariate methods is to efficiently extract information concerning certain analytes of interest from spectra of multicomponent mixtures. Perhaps the simplest of these methods is classical least squares (CLS). CLS is an extension of the well-known method for resolving mixtures of *I* components by measuring the absorbance at *I* wavelengths. In CLS, measurements are performed at *n* wavelengths (in practice $n >> I$), and, thus, it is considered to be a full-spectrum method. This generally leads to higher precision as compared to using only a small number of wavelengths (*25*); however, it also has some drawbacks. First, it uses a *direct* calibration step, which requires the knowledge of all sample components. The term *direct* refers to the usual definition of Beer's law ($A = kc$) extended to multicomponent

mixtures. Second, it is very sensitive to spectral noise, baseline drift, and spectral overlap of the sample components (*25*). Inverse least squares (ILS) circumvents these problems. Because the inverse Beer's law, $c = k'A$, (extended to multicomponent mixtures) is used, it only requires the knowledge of the concentration of the analyte of interest for calibration (*25*); however, it is restricted to a small number of wavelengths. Better performance is obtained with methods which use spectral factors, such as principal component regression (PCR) and partial least squares (PLS) (*30*). These methods use inverse calibration steps, combined with a prior optimization of the information contained in the calibration spectra. They display the following advantages: (1) they use full spectra, (2) they require knowledge of only the concentrations of the analytes of interest, and (3) the spectra can be decomposed into factors, avoiding the problems associated with overlapping, collinearities, noise, drifts, and other spectral artifacts (*25*). They are ideally suited for the study of complex biological samples, such as drug or metabolite monitoring in blood (*31*) or pharmaceutical analysis of multicomponent preparations where the excipients may not be known (*32*). A common requirement to all these multivariate methods is that the matrix should be modeled during calibration; that is, all compounds in which the chemist is not interested should be present in the calibration samples (although one need not know their concentrations). A summary of the advantages and drawbacks of these methods is shown in Table 1.

Table 1 Summary of Some Popular Multivariate Calibration Methods.

| Method | Advantages | Drawbacks |
|---|---|---|
| CLS<br>Classical<br>Least Squares | Uses full spectra. | All components should be known.<br>Sensitive to collinearities, baseline drifts, noise, etc. |
| ILS<br>Inverse<br>Least Squares | Only the components of interest need to be known. | Uses a small number of Wavelengths.<br>Sensitive to collinearities. |
| PCR<br>Principal<br>Components<br>Regression | Only the components of interest need to be known.<br>Uses spectral factors (less sensitive to collinearities).<br>Uses full spectra. | The matrix should be modeled in the calibration. |
| PLS<br>Partial Least<br>Squares | Uses spectral and concentration factors (less sensitive to collinearities).<br>Uses full spectra. | The matrix should be modeled in the calibration. |

A firm connection has also been established between chemometrics and spectroscopy. Figure 2 shows in a simplified manner why this is a powerful combination.




Spectroscopic analysis
*short time of analysis
*large amount of data

Chemometrics
*extract the information in large amounts of data

Figure 2 Illustration of why spectroscopy and chemometrics work well in conjunction.

Spectroscopic techniques are generally fast, with the analysis taking from a few seconds to a few minutes. As previously mentioned, spectroscopy also produces large amounts of data for each sample analyzed. Roughly speaking, this data can be said to consist of two parts: information and noise. The information part of the data is what eventually leads to knowledge about the sample, while the noise is a non-information part. A matter of concern is always to minimize and, if possible, to get rid of disturbing noise in the data since it impairs the information gained. This is where chemometrics comes in, since multivariate methods are constructed to extract the information from large sets of data. Using multivariate data with many variables instead of univariate data offers many advantages in qualitative and quantitative spectroscopic analysis. The methods generally become more robust, precise and less sensitive to spectral artifacts. One could therefore say that multivariate methods are the optimal choice for the evaluation of spectroscopic data and that the conjunction of spectroscopic analysis techniques with multivariate data analysis offers further possibilities in analytical chemistry.

## Multivariate data analysis

The term multivariate can be explained as "*multi*" meaning numerous or many and "*variate*" meaning variation or change. Multivariate data analysis is thus the analysis of data consisting of multiple variables measured from many samples. The aim of chemometric methods for data decomposition and reduction is to find a small number of *latent variables* that can explain all the variation in the data matrix studied. Multivariate data analysis thus tries to find the relationships between the variables and samples in the data set and capture them in new latent variables. Some important definitions in multivariate data analysis are the terms *variables* and *observations*, as shown in Figure 3 below. The rows of the matrix are generally called *objects* or observations and thus comprise the samples. The columns are in turn called variables and consist of the entities measured for each object. The variables are generally divided

into X and Y. As previously mentioned, in experimental design the X variables are called factors, the y-variables responses and the objects experiments.

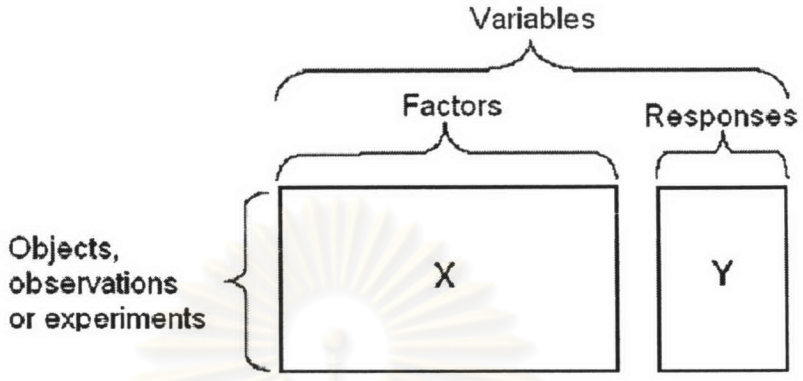



Figure 3 Illustration of the terms objects, observations and experiments, variables, factors and responses.

Multivariate methods that find the relationship between x and y-variables are generally called regression methods. Examples of these methods are PCR (*6*), PLS (*29,33*) and MLR (*6*).

Often differing pre-treatment of the data, as in Figure 4, is carried out before the multivariate data analysis takes place. The reasons for performing data pre-treatment are generally to reduce the effect of noise, improve the predictive ability of the model and simplify the model (lower model dimensionality) by making the data more normally distributed. Transformation is pre-treatment by means of a mathematical formula to change the distribution of the data, examples being logarithmic and exponential transformation. If the average of each variable is calculated and then subtracted from the variables, the data is said to be mean-centred. Unit variance scaling means that the standard deviation (*s*) for each variable is calculated and by multiplying each value of that variable by 1/s, all variables are then given an equal weight. This is also sometimes called variance scaling or auto-scaling.

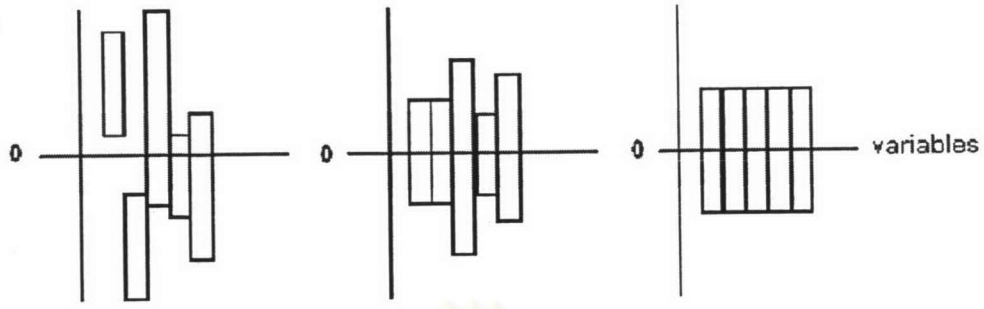



Figure 4 Data pretreatment. First column mean-centering (a) followed by variance scaling (b)

## Multivariate calibration

The term calibration can be defined as the use of empirical data and prior knowledge for determining how to predict unknown quantitative information Y from available measurements X, via a mathematical transfer function of some kind (*6*). Calibration can hence be described as the process of establishing this mathematical function (*f*) between the measured variable *x* and a dependent variable *y*:

$$f(x) = y \qquad (1)$$

One of the simplest forms of calibration is *linear regression*.

$$y = kx + l \qquad (2)$$

where *k* is the regression coefficient and *l* is the intercept of the linear approximation. In linear regression one x-variable and one y-variable are used. In multivariate calibration (*6,34,35*), however, numerous variables are used and the term multivariate calibration refers to the process of constructing a mathematical model that relates a property such as content or identity to the absorbances of a set of known reference samples at more than one wavelength. Multivariate calibration thus means using many variables simultaneously to quantify one or many target variables Y. A calibration model is determined from a set of samples of known content and identity, the calibration set. This can be done by means of PCR or PLSR and the resulting model is used to predict the content of new unknown samples from their digitized spectra. The calibration set could

consist, for instance, of *m* samples of known content (*y*). From these samples the *n* spectral variables are measured. If the resulting first-order data arrays of each sample are put together in a table, the result is as illustrated in Figure 5. From the X and Y matrices of the calibration set, the calibration model is then constructed and subsequently validated. The best way to perform this validation is by using new samples not previously used, a validation set consisting of new samples (*p*) from which the same variables have been measured. By predicting the Y values of the samples in the validation set and then comparing the results with the true values, an estimate of the predictive ability of the model is obtained.

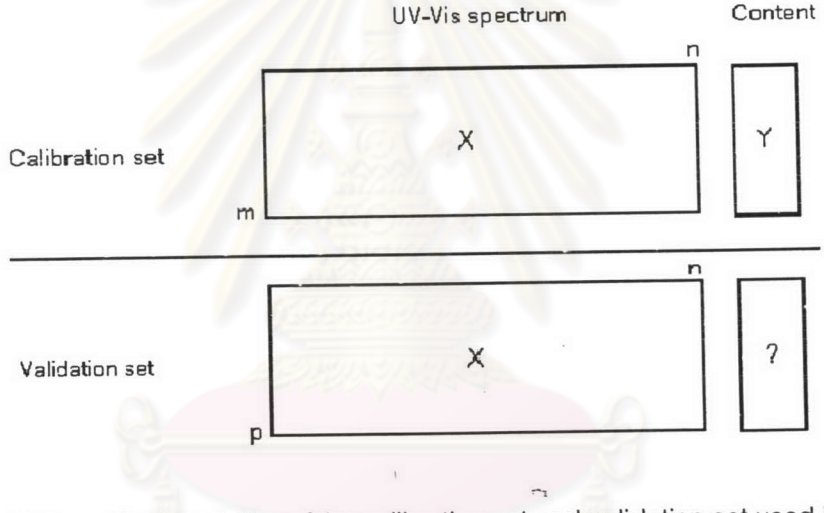



Figure 5 Schematic description of the calibration set and validation set used in multivariate calibration.

There are a number of reasons for using multivariate calibration. Univariate calibration works well as long as no other compounds in the solution analysed absorb light at the wavelength used, i.e. the wavelength is selective for the compound under study. If this is not the case, all the other absorbing compounds, i.e. the interferences in the solution must be known. In multivariate calibration, however, this is not the case since using many x-variables automatically corrects for each other's selectivity problem, and the x-variables used thus do not need to be totally selective. The precision of multivariate calibration is generally high as long as there is a linear relationship between

the x and y-variables. Multivariate calibration is also generally more robust to small changes in the experimental or instrumental parameters such as small changes in pH, temperature or lamp intensity. Samples that differ in some way, outliers, are also more easily spotted and evaluated with multivariate data than with univariate data since multivariate residuals will show how an unknown sample fits into the calibration model. In addition, in multivariate calibration numerous graphical diagnostic tools are generally available.

The calibration samples should in the calibration set span the calibration experimental domain as far as possible and experimental design can be used to find suitable samples to include in the calibration set. Also the best calibration is generally obtained when the calibration model contains all the sources of variation that can occur in the actual measurement. The order in which the samples are analyzed is one source of variation and randomization of the order of analysis minimizes the error caused by day-to-day variations. The evaluation of the predictive ability of a quantitative multivariate calibration model can be made, by means of the root mean square error of prediction (RMSEP) (*6*) and relative standard error of prediction (RSEP) (*36,37*).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{p} (y_{pred} - y_{obs})^2}{(p)}} \qquad (3)$$

$$RSEP(\%) = 100 * \sqrt{\frac{\sum_{i=1}^{p} (y_{pred} - y_{obs})^2}{\sum (y_{obs})^2}} \qquad (4)$$

where $y_{pred}$ is the predicted concentration in the sample, $y_{obs}$ is the observed or reference value of the concentration in the sample and $p$ is the number of samples in the test set. RMSEP gives an estimate of the prediction error in the same unit as the initial data, while RSEP gives a relative measure of the prediction error in terms of

percentage. RSEP has also been defined as the relative standard deviation of residuals of the concentrations (*37*). In recent years the use of multivariate calibration has become well established and standardised guidelines have been published (*38*).

## Principal component analysis

In the following text capital boldface letters (X) represent two-way matrices, underlined capital bold face letters (X) three way arrays and lower-case boldface letters (x) vectors (one-way arrays).

PCA is a well-known chemometric method for the decomposition of two-way matrices (*29,39*). The variance in the data matrix X, with *m* observations and *n* variables, is decomposed by successively estimating principal components (PCs) that capture the variance in the data in scores and loadings. In Figure 6 the geometrical interpretation of PCA on a small data set consisting of twelve observations and three variables is shown. The steps in the geometrical interpretation of PCA are as follows. Firstly the X-space is given a coordinate system where each variable gets an axis whose length corresponds to its scaling. In this example three variables mean three coordinate axes. Each observation in this space is represented by a point. The average of each variable is then calculated and subtracted (mean centring). This is equivalent to moving the swarm of points to the centre of the coordinate system. Thereafter a function is fitted to the data that describes as closely as possible the variance of the observations in the X-space. This is the first PC and is represented by the line in Figure 6b.

By projecting each point down to the line (Euclidian distance) and measuring the distance between the centre point and the projection point, the score value (t) of each observation is obtained. Since the data set consists of twelve objects, the same number of score values (t1) are obtained for the first PC. The angle between the line and each

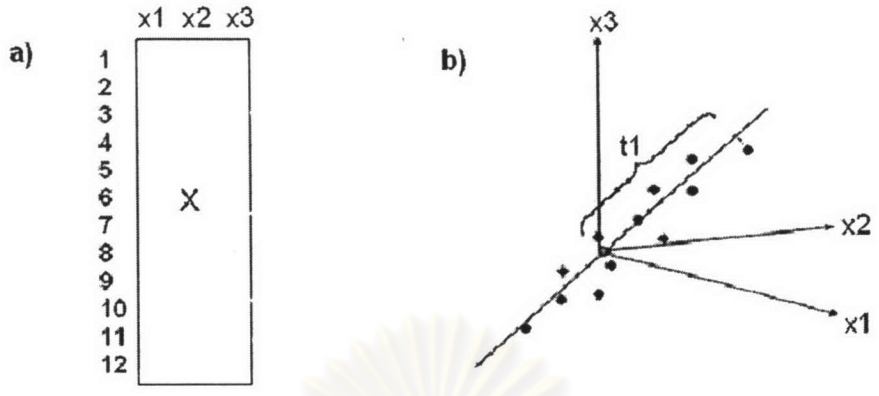



Figure 6 Geometrical interpretation of a PCA model consisting of one PC: a) data set consisting of twelve objects and three variables, b) a geometric interpretation of the first PC calculated from the data set.

variable axis determines the influence of each variable, the loading value (p). One loading value is given for each variable in the data set (p1). In this example twelve score values and three loading values are thus given for the first PC. When the first PC has been calculated, the remaining unexplained variance is left in the residual matrix, E:

$$X = TP' + E \qquad (5)$$

The decomposition of a matrix with PCA is schematically described in Figure 7, where the initial data matrix X is decomposed with PCA using two PCs.

After the first PC has been calculated, the next is calculated on the residual matrix E1, which contains the variance not explained by the first PC.

$$X = t1p1' + t2p2' + \ldots + tapa' + E \qquad (6)$$

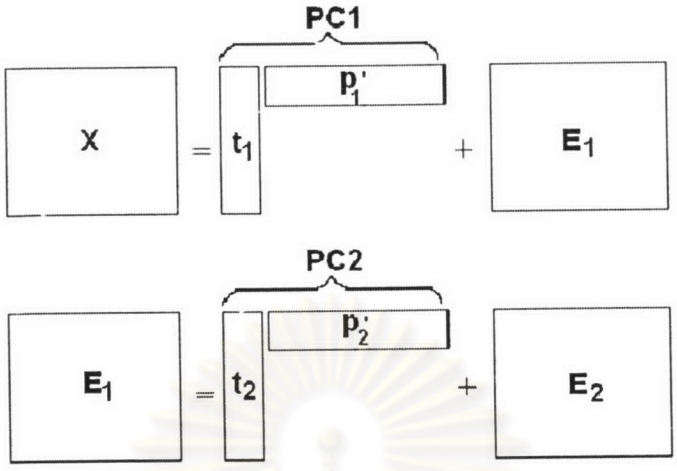



Figure 7 Schematic description of a decomposition of a matrix X with PCA using two PCs.

The second PC is orthogonal to the first. More PCs (*a*) can be calculated as long as unexplained information is left. The significant number of principal components can be estimated by different methods, of which cross validation is an often-used method. The variance of a principal component is described by the eigenvalue, which is proportional to the variance explained by a PC. The eigenvalue ($\lambda$) can be described as the length of the PC and estimated as the sum of squares of the scores:

$$\lambda = \sum_{m=1}^{M} t_{ma}^2 \qquad (7)$$

where $t_{ma}$ is the score of object *m* for component *a*. Hence the length of the score vector is proportional to the importance of the particular PC describing how much of the variance in X that is explained.

Although PCA can be calculated using different algorithms, the two methods most common are non-linear iterative partial least squares (NIPALS) (*29*) and singular value decomposition (SVD) (*39*).

One problem with applying PCA to matrices with a chemical rank larger than one is that is does not give directly interpretable profiles like chromatograms and spectra since scores and loadings become linear combinations of the true analytical profiles. This issue is also called the problem of rotational freedom since the scores and loadings can be rotated without changing the fit of the model.

## Partial least squares regression

PLSR (*28,32*) has been established as a standard data analysis tool for multivariate data in the last ten to fifteen years and there are numerous applications in different fields of analytical chemistry. In PLSR the variance in a data matrix **X** and a dependent matrix **Y** is decomposed by successively estimating PLS components that capture the variance and correlation between **X** and **Y**. In Figure 8 the geometrical interpretation of PLS of two small data sets consisting of twelve objects, three x-variables (x1–x3) and three y-variables (y1–y3) are shown. The steps in the geometrical interpretation of PLS are as follows. Firstly, as in PCA, the X space gets a coordinate system where each variable gets an axis with the length corresponding to them scaling and in PLS this is also given for the Y space. Since the data sets in this example consisted of three x-variables and three y-variables, the X and Y spaces both get three axes.

Mean centering of both the x and y-variables then moves the swarm of points to the middle of the coordinate system. A function is then fitted to the data in a way that best describes the variance in the X and Y spaces as well as maximizing the correlation between **X** and **Y**. This function is a line in both the X and the Y space and it is the first PLS component. PLS can be seen as the regression extension of PCA since it simultaneously fit two ´PCA-like´ models, one for the X space and one for the Y space, in such a way that the correlations between X and Y are maximized. In contrast to PCA, PLS is a maximum covariance method since the main aim of PLS is to predict the y-variables from the x-variables. As in PCA, the PLS decomposition summarizes the

variance in the data sets in new latent variables, scores and loadings. By projecting each point down to the line and measuring the distance between the center point and the projection point, the scores for the X space (t1) and Y space (u1) are

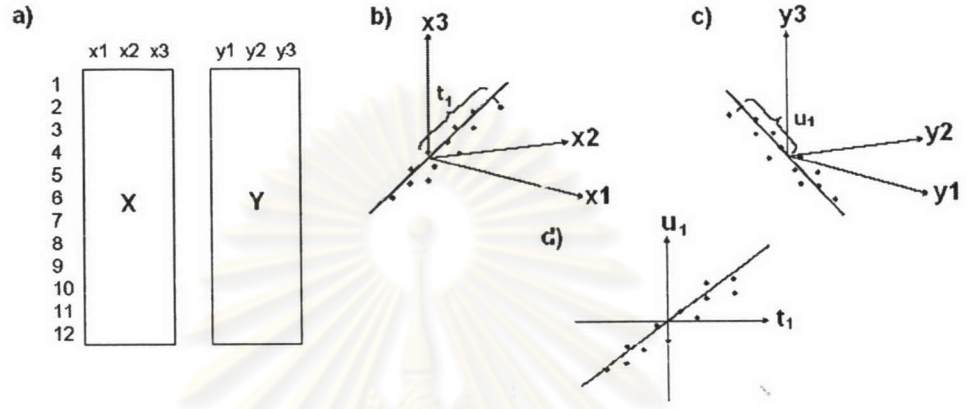



Figure 8 Geometrical interpretation of a PLS model consisting of one PLS component. a) a data set consisting of twelve objects and three x and y-variables, b) geometric interpretation of the first PLS component in the X space, c) geometric interpretation of the first PLS component in the Y space, d) PLS inner relation.

given. The PLSR model can be regarded as consisting of an outer relation and an inner relation, where the outer relation describes the X and Y block individually, while the inner relation links the two blocks together (*29*). The outer relations is given by

$$X = TP' + E \quad (8)$$

$$Y = UC' + F \quad (9)$$

where T is the score matrix and P′ the loading matrix of the X space, U the score matrix and C′ the loading matrix of the Y space. E and F are the residual matrices of the X and Y spaces respectively. By plotting the **t1** values of the twelve objects in the example in Figure 8 against the corresponding **u1** values, the PLS inner relation showing the correlation structure between X and Y is obtained (Figure 8d).

$$U = BT + H \quad (10)$$

where **B** is an identity matrix and **H** is a residual matrix. In PLS a type of additional loadings is calculated that expresses the correlation between **X** and **Y**, the weights, **W**. The weights are related to the PLS regression coefficients, $B_{PLS}$:

$$B_{PLS} = W(P'W)^{-1} C' \qquad (11)$$

The regression coefficients show the direction and magnitude of the influence of an x-variable on a specific y-variable. The prediction of y-variables of new samples is given by

$$Y = XB_{PLS} + F \qquad (12)$$

After the first PLS component has been calculated the next one can be calculated on residual matrices **E** and **F**. The number of significant PLS components (the model dimensionality) in a calibration model can be decided by means of cross validation. The matrices in PLS regression are shown schematically in Figure 9 and, for each new PLS component, **t**, **u**, **w**, **p** and **c** are calculated. The details of PLS regression have been thoroughly described elsewhere *(29,33)*.

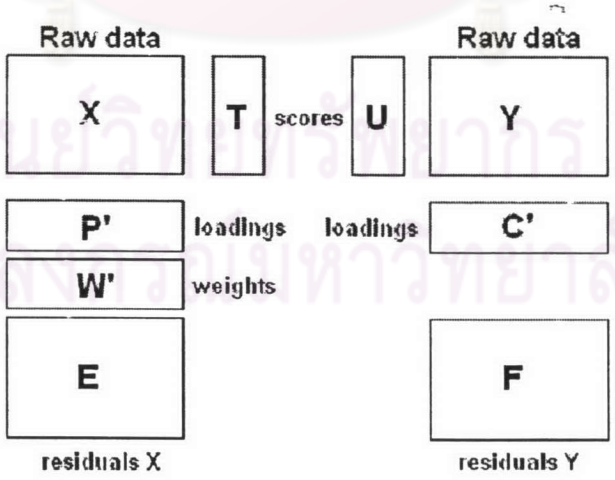



Figure 9 Schematic description of the matrices in PLS regression.

## Validation

Validation is an extremely important part of model building. There are two main purposes of validation.

The first is to make sure that the model will work in the future when the model is applied to new data of future measurements. This is done by estimating the prediction error of the model. It is of course impossible to get a correct value of the prediction value for future measurements but it is important to get an estimate that is as good as possible. This estimate tells us if the model is useful for future measurements and it gives a value of good the model performance is.

The second purpose is to find the optimal (correct) complexity of the model. In the case of PCR or PLSR, the complexity is the number of PCs or latent variables included in the model.

1. Calibration set and validation set

In order to validate a model we have to have both calibration set and validation set. Calibration data is used to build the model and validation set is used to test the prediction ability of the model. It is important that both calibration and validation set covers all possible aspects of future data, i.e. the calibration and validation set should be representative. For example if the variable y is a concentration of some sort, it is important that calibration set and validation set covers the range of the concentrations that we want to predict for future measurements. In order words, the calibration set and validation set must span the data space of X and y. If the results form the prediction testing is satisfactory, the validation model may be used for the prediction of the results of "unknown samples".

2. Number of principal components

If the underlying model for the relation between X and y is a linear model, the number of components needed to describe this is equal to the model dimensionality. Although it is possible to calculate as many principal components as the rank of the X

block matrix, not all of them are usually used. The main reasons for this are the measured data are never noise-free and some of smaller components will only describe noise. As mentioned in earlier paragraphs, it is common to leave out small components because they carry the problems of collinearity.

In any empirical modeling, it is essential to determine the correct complexity of the model. With numerous and correlated X-variables there is a substantial risk for "over-fitting", i.e., getting a well fitting model with little or no predictive power. Hence, a strict test of the predictive significance of each principal component is necessary, and then stopping when components start to be non-significant.

Cross-validation (CV) or leave-one-out (LOO) procedure must be used to establish the number of components, is a practical and reliable way to test this predictive significance. In CV is the part of data used for testing is defined by leaving out one measurement for every sub-model and building the sub-model on the rest of the measurements. Each measurement of the total amount of measurements is left out once and thus we get as many sub-models as we have measurements.

After developing a model, differences between actual and predicted Y-values are calculated for the deleted data. The sum of squares of these differences is computed and collected from all the models to form the predictive residual sum of squares (PRESS), which estimates the predictive ability of the model.

$$\text{PRESS}(h) = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{i,j} - \hat{y}_{i,j})^2$$

where $n$ = total number of calibration samples

$\hat{y}_{i,j}$ = predicted concentration of analyte j in i calibration samples left out during calibration

$y_{i,j}$ = actual concentration analyte j in i calibration sample left out during calibration

$m$ = analyte of mixture sample

$h$ = the number successive principal component

This parameter is a measure of the efficiency for a calibration fit model. One reasonable choice for the optimum number of factors would be the number that yielded the minimum PRESS. However, using the number of factors (h*) that yield a minimum PRESS usually leads to some over-fitting. A better criterion for calculating the optimum number of factors involves the comparison of PRESS from the models with fewer than h* factors. The model selected is that model is nor significantly greater than PRESS from the model with h* factors. We selected an optimum the number of which its F ratio probability drops below 0.75 (*44,45*).

Another often-used parameter is eigenvalue EV (a) or $g_a$. Eigenvalues are calculated as the sum of squares of the score vectors, conventionally used as a measure of the size of a principal component.

$$EV(a) = g_a = \sum_{i=a}^{n} t^2_{ia},\ a = 1, 2, \ldots, n$$

where n = total number of principal component

a = the number successive principal component

This is explained by the fact that successive eigenvalues (a principal component) explain less variance in the data and hence explains the continual drop in the residual percent variance.