MedeA QT: An Interactive QSPR Toolbox

Contents

1 Introduction

MedeA QT, the *MedeA* QSPR Toolbox, employs an interactive graphical user interface to allow you to explore and analyze the relationships between descriptors and system properties. *MedeA QT* employs statistical methods, such as partial least squares (PLS), to compute correlations.

2 The Partial Least Squares (PLS) and Multiple Linear Regression (MLR) Methods

Partial Least Squares [1], [2], [3], [4], [5] was developed by Herman and Svante Wold and is now widely used in chemometrics and related areas. A PLS model is based on the determination of multidimensional directions in descriptor space that provide the maximum variance in both descriptor and activity space. Additional information on PLS modeling is provided below. PLS regression is particularly suited to situations that are challenging for multiple linear regression (MLR) methods. For example, where there is collinearity in supplied descriptors, and where there are relatively few activities relative to descriptors, ordinary least squares will likely require the handling of ill-conditioned matrices, and the MLR algorithm may not yield stable models. The PLS method is generally numerically stable and can handle duplicated descriptor columns, for example, that would lead to difficulties in matrix operations for MLR least-squares model-building algorithms.

The difference between MLR and PLS can be summarized as follows. MLR adjusts a set of descriptor coefficients, given a fixed descriptor space, to minimize the squared deviations between estimated and observed activities. In contrast, PLS seeks to maximally span descriptor and activity space, and generate correlations between descriptor space and activity space. PLS tackles these three objectives in the inherent optimization problem that the algorithm solves to create a predictive model.

The numerical basis of the PLS method can be summarized as follows. In the MLR case, the essential equation to be solved is as follows:

$$y = Xb + e \tag{1}$$

[1] A. Boulesteix, K. Strimmer, “Partial least squares: a versatile tool for the analysis of high-dimensional genomic data”, *Briefings in bioinformatics*, **8** (2006)

[2] R. Rosipal, N. Krämer, “Overview and recent advances in partial least squares”, In *Subspace, latent structure and feature selection*, (pp. 34-51), Springer, Berlin, Heidelberg (2006)

[3] R. Tobias, “An introduction to partial least squares regression”, In *Proceedings of the twentieth annual SAS users group international conference* (pp. 1250-1257), SAS Institute Inc. Cary, NC (1995)

[4] S. Wold, A. Ruhe, H. Wold, W.J. Dunn, “The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses”, *SIAM Journal on Scientific and Statistical Computing*, **5**, 735 (1984)

[5] S. Wold, P. Geladi, K. Esbensen, J. Öhman, “Multiway principal components and PLS analysis”, *Journal of chemometrics*, **1**, 41 (1987)

where *X* is a matrix of descriptors for each measured activity, *b* is a vector of regression coefficients, *y* is the vector of activities, and *e* is a vector of errors. This equation is typically and straightforwardly solved forming normal equations and using matrix inversion, yielding the vector of regression coefficients, *b*.

PLS uses a decomposition of the *X* matrix and the *y* vector in the following manner:

$$X = TP^T + E \tag{2}$$

$$y = Tq + f \tag{3}$$

Here *T* is known as a score matrix and *P* as a loading matrix, terminology which originates in principal component regression (PCR) which is related to PLS. *E* and *f* are matrix and vector of residuals, respectively. The PLS algorithm determines the vectors that comprise *T* and *U* based on the maximization of their variance, and this maximization provides the simultaneous multi-objective optimization that provides PLS with its properties.

Given (2), then

$$XW = TP^TW \tag{4}$$

where X is a weight matrix. Hence *T* is

$$T = XW(P^TW)^{-1} \tag{5}$$

and from (3),

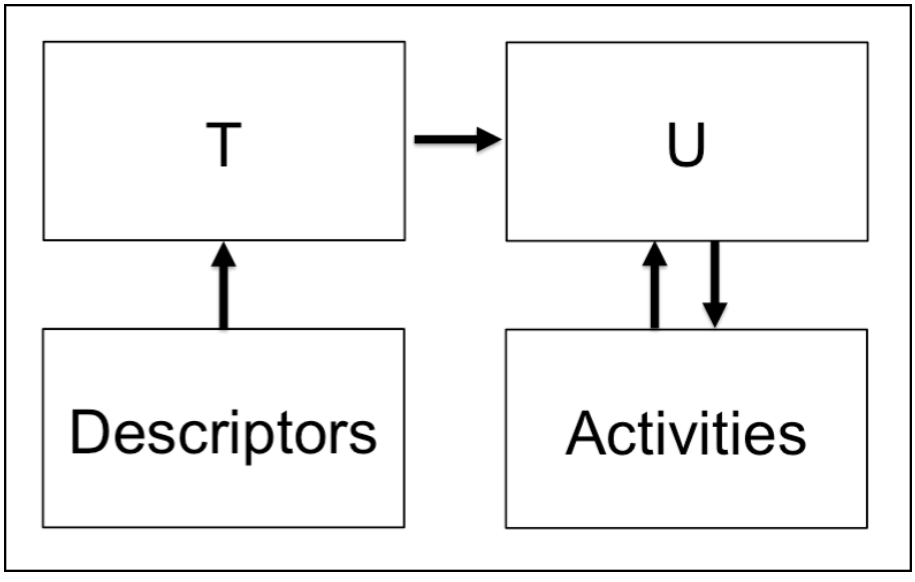
$$y = Tq = XW(P^TW)^{-1}q \tag{6}$$

as *y* = *Xb*, the PLS correlation coefficients, *b*, are:

$$b = W(P^TW)^{-1}q \tag{7}$$

Residual terms, which do not affect the transformations, are omitted from the equations above.

Herman and Svante Wold and other researchers have shown that focusing on the three objectives in the PLS algorithm (spanning descriptor and activity spaces as well as maximization of correlation) results in a modeling approach with desirable characteristics, including numerical stability and robustness with respect to outlying descriptors. Recognizing that effectively PLS transforms the supplied descriptor space has resulted in an alternative interpretation of the PLS acronym: 'Projection to Latent Structures', which is less widely used than 'Partial Least Squares' but perhaps more accurately describes the method. 'Partial Least Squares' is employed here as this term is more widely employed in the literature.



The diagram above illustrates the essential operation of the PLS method. The 'descriptor' space supplied to describe the systems of interest are converted into a set of latent variables, labeled **T** in the diagram.

Activities are also transformed into a latent or derived form, labeled **U**. The space of the transformations may be reduced relative to the dimensionality of the original descriptor space dimensions and the effect of the transformation is to reduce collinearity problems, as discussed above.

Several algorithms have been described to generate the score and loading matrices. These are typically either iterative or recursive and from an initial starting build the matrices step by step.

The algorithm employed by *MedeA QT* is known as Nonlinear Iterative Partial Least Squares and proceeds through the simple regression of descriptor vectors from the *X* matrix to create the vectors of the *P* and *Q* matrices. The algorithm terminates when a defined number of latent variables have been employed or no further correlation between *y* and *t* vectors is obtained. Extensive additional information on the PLS method is available in the literature.

The technical literature related to PLS is diverse for several reasons. The method has been developed comparatively recently, it is applied in many fields each with differing terminologies, and a wide range of algorithmic variants of PLS have been, and continue to be, developed. However, several selected publications are listed below that provide an overview of the method.