World Academy of Science, Engineering and Technology
International Journal of Chemical and Molecular Engineering
Vol:13, No:04, 2019

Rd-PLS Regression: From the Analysis of Two Blocks of Variables to Path Modeling

Authors : E. Tchandao Mangamana, V. Cariou, E. Vigneau, R. Glele Kakai, E. M. Qannari

Abstract : A new definition of a latent variable associated with a dataset makes it possible to propose variants of the PLS2 regression and the multi-block PLS (MB-PLS). We shall refer to these variants as Rd-PLS regression and Rd-MB-PLS respectively because they are inspired by both Redundancy analysis and PLS regression. Usually, a latent variable t associated with a dataset Z is defined as a linear combination of the variables of Z with the constraint that the length of the loading weights vector equals 1. Formally, $t=Zw$ with $\|w\|=1$. Denoting by Z' the transpose of Z, we define herein, a latent variable by $t=ZZ'q$ with the constraint that the auxiliary variable q has a norm equal to 1. This new definition of a latent variable entails that, as previously, t is a linear combination of the variables in Z and, in addition, the loading vector $w=Z'q$ is constrained to be a linear combination of the rows of Z. More importantly, t could be interpreted as a kind of projection of the auxiliary variable q onto the space generated by the variables in Z, since it is collinear to the first PLS1 component of q onto Z. Consider the situation in which we aim to predict a dataset Y from another dataset X. These two datasets relate to the same individuals and are assumed to be centered. Let us consider a latent variable $u=YY'q$ to which we associate the variable $t= XX'YY'q$. Rd-PLS consists in seeking q (and therefore u and t) so that the covariance between t and u is maximum. The solution to this problem is straightforward and consists in setting q to the eigenvector of $YY'XX'YY'$ associated with the largest eigenvalue. For the determination of higher order components, we deflate X and Y with respect to the latent variable t. Extending Rd-PLS to the context of multi-block data is relatively easy. Starting from a latent variable $u=YY'q$, we consider its ‘projection’ on the space generated by the variables of each block X_k (k=1, ..., K) namely, $t_k= X_kX_k'YY'q$. Thereafter, Rd-MB-PLS seeks q in order to maximize the average of the covariances of u with t_k (k=1, ..., K). The solution to this problem is given by q, eigenvector of $YY'XX'YY'$, where X is the dataset obtained by horizontally merging datasets X_k (k=1, ..., K). For the determination of latent variables of order higher than 1, we use a deflation of Y and X_k with respect to the variable $t= XX'YY'q$. In the same vein, extending Rd-MB-PLS to the path modeling setting is straightforward. Methods are illustrated on the basis of case studies and performance of Rd-PLS and Rd-MB-PLS in terms of prediction is compared to that of PLS2 and MB-PLS.

Keywords : multiblock data analysis, partial least squares regression, path modeling, redundancy analysis
Conference Title : ICCAC 2019 : International Conference on Chemometrics in Analytical Chemistry
Conference Location : Rome, Italy
Conference Dates : April 09-10, 2019