SI Text

Relation between canonical correlation analysis (CCA), ordinary least squares (OLS) regression, partial least squares (PLS) and principal component regression (PCR). There are several multiple regression methods, which are applied in different scientific domains like chemometrics and statistical bioinformatics. Here, we explain the differences between and the relative merits of three of the most important methods: CCA/OLS, PLS, and PCR.

The input data for each of the regression methods consist of a predictor *X* and a response matrix *Y*. While their number of rows, which is equal to the number of samples, must be the same for both, the number of columns is generally different. These regression methods can be described by the criteria they maximize (or correspondingly minimize). CCA finds the linear combinations of columns (*Xw, Yv*) (called canonical variates) of both matrices, which have the maximal correlation (1):

$$(\mathrm{w}, \mathrm{v}) = \arg\max_{w,v} corr(\mathrm{Xw}, \mathrm{Yv})$$

The vectors ***w*** and ***v*** are the vectors of regression coefficients. Often the response has only one column i.e., is represented by a vector. In this case OLS regression detects the linear combination of the predictor variables, which has the least squares difference with the response.

$$\mathrm{w} = \arg\min_{w}(\mathrm{Xw} - \mathrm{Y})^2$$

The canonical variate has the same direction as this estimation. In other words the minimum squared distances yield the maximum correlation. This method has the advantage of yielding an unbiased estimate of the regression coefficients. However, there is a tradeoff: The mean square error (MSE), i.e., the difference between true and

estimated regression coefficients is often high especially as the ratio of the number of predictor variables and sample size increases. This failure is apparent in a cross-validation procedure. The coefficients that maximize the correlation in the training set, often give poor results in the test set.

Therefore alternative regression methods were developed, which accept some bias in the estimation for the sake of a lower MSE. These methods effectively reduce the number of dimensions, i.e., the number of predictor variables. The coefficient vector ***w*** is chosen by these methods in such a way that directions, for which the spread of predictor variables is small, can be omitted.

The most extreme method is the PCR. The first step is to find a vector ***w,*** for which the variance of *Xw*, is maximal:

$$\mathrm{w} = \arg\max_{w} \mathrm{var}(\mathrm{Xw})$$

The OLS regression is then performed on those of the new variables that have the highest variance. The disadvantage of the procedure is that the new variables are determined without considering the response. This is especially the case, if the predictor variables are largely uncorrelated.

Therefore PLS is often proposed as an alternative (2). The underlying maximization principle is that of the covariance between prediction and the response: Find ***w*** such that the covariance of *Xw* and *Y* is maximal:

$$\mathrm{w} = \arg\max_{w} \mathrm{cov}(\mathrm{Xw}, \mathrm{Y}) = \arg\max_{w} \mathrm{var}(\mathrm{Xw}) corr(\mathrm{Xw}, \mathrm{Y})$$

Since the covariance of two variables is equal to its correlation multiplied with the variance of both variables, PLS occupies an intermediate position between CCA/OLS and

PCR, and it has been shown to be more appropriate for cross-validation in many cases (3).

1. Hotelling H (1935) *J Educ Psychol* 26:139-143.

2. Wold H (1975) *Soft Modelling by Latent Variables* (Academic Press, London).

3. Frank IE, Friedman JH (1993) *Technometrics* 35:109-135.