1. PLS*-Regression: Y ~ X

Fit $Y = c_0 + c_1 \cdot T_1 + \ldots + c_m \cdot T_m$

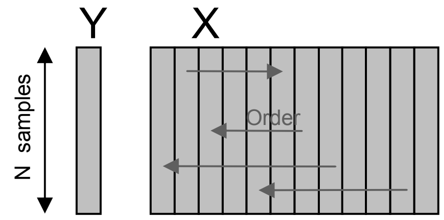
with $T_i = w_{i1} X_1 + \ldots + w_{ip} X_p$ for $i = 1, \ldots, m$.

Here Y is the given response vector (N x 1) and X the given matrix (N x p) of describing variables.

2. Computation of VIP** for each variable $X_1, \ldots, X_p$

$$VIP(X_j) = \sqrt{\frac{\sum_{i=1}^{m} \left(w_{ij} \cdot cor(Y, T_i)\right)^2}{\sum_{i=1}^{m} \left(cor(Y, T_i)\right)^2} \cdot p}$$

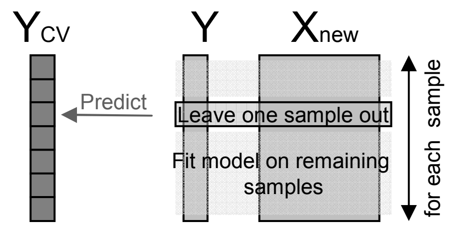
3. Order variables $X_1, \ldots, X_p$ with respect to VIP (descending). → $X_{1:p}, \ldots, X_{p:p}$



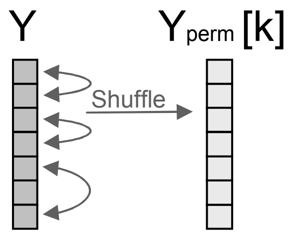
4. For j = 1, ..., p do PLS*-Regression: Y ~ $(X_{1:p}, \ldots, X_{j:p})$

5. Select number j.opt such that the corresponding model Y ~ $(X_{1:p}, \ldots, X_{j.opt:p})$ has highest R^2 in CV. → $X_{new} := (X_{1:p}, \ldots, X_{j.opt:p})$

6. PLS*-Regression: Y ~ X_{new}



7. Generate 5000 different permutations of response Y. → $Y_{perm}[1], \ldots, Y_{perm}[5000]$



8. Test significance of the correlation R = cor(Y , Y_{CV}) by comparing to cor(Y_{perm} [k] , Y_{CV}) for k = 1, ..., 5000:

Compute mean and s of cor(Y_{perm} [k] , Y_{CV}), k = 1, ..., 5000. Compute distance of R and mean as measured by s; estimate P-value, assuming N(mean, s^2).

* Partial Least Squares Regression: Projecting describing variables to latent variables $T_1, \ldots, T_m$, such that R^2 is maximized in CV of the OLS-model Y ~ T with respect to m

** Variable Importance in the Projection of X_j: Contribution to Y of those latent variables with high weights in X_j

Abbreviations:

- CV : Cross Validation (Leave-one-out)
- Y_{CV} : Response predicted in CV
- R^2 : Squared correlation of Y and Y_{CV}
- s : Sample standard deviation
- OLS: Ordinary Least Squares