Multimatrix Extension of Partial Least Square

Raju Rimal (Supervisor: Solve Sæbo)

Norwegian University of Life Sciences (NMBU), Ås, Norway

Norwegian University of Life Sciences

Partial Least Square Regression

Lets consider a model with multiple response $\mathbf{Y}$, predictor matrix $\mathbf{X}$ and error matrix $\mathbf{E}$ as,

$$\mathbf{Y} = \mathbf{B}_\circ + \mathbf{XB} + \mathbf{E}$$

Usual Problems

► Multicolinear Predictors
► High Dimensional data
► Large Number of predictors

Partial Least Square Regression

Partial Least Square (PLS) regression is a,

► Method that can explore the underlying relation of predictor with response through latent structure
► Method that can gather maximum covariance structure between $\mathbf{X}$ and $\mathbf{Y}$ in first few components
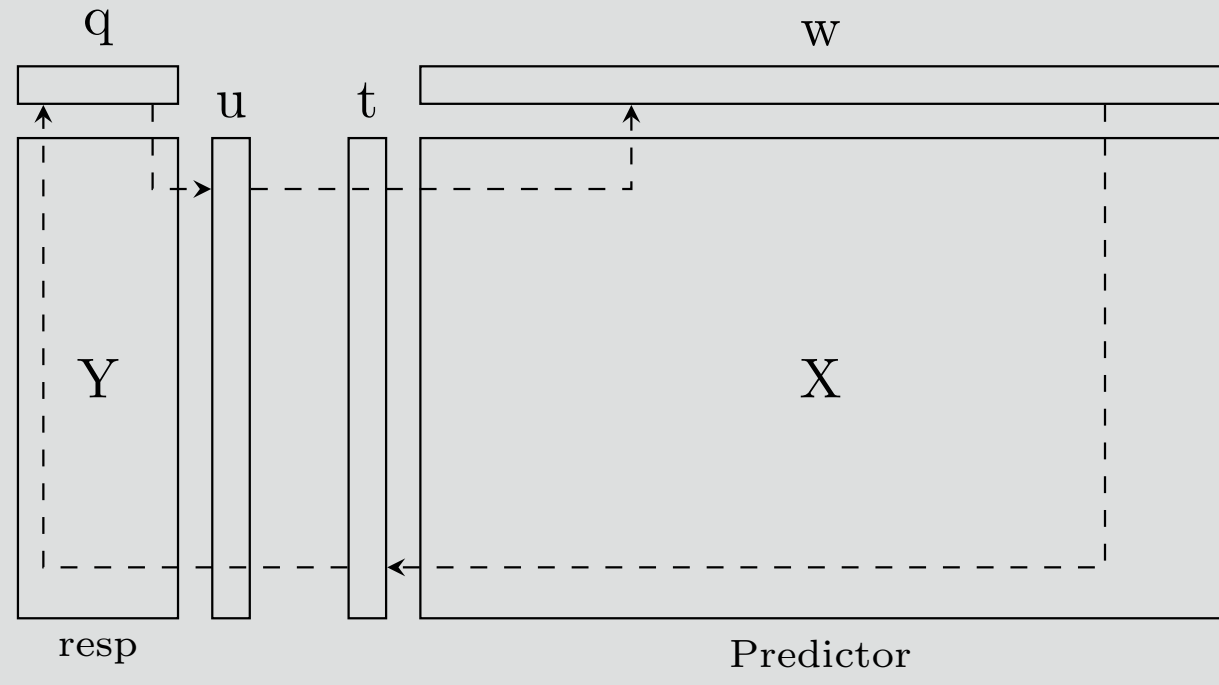

Figure: NIPALS algorithm for performing PLS regression. Starting from any arbitrary column $\mathbf{u}$, the algorithm converges after some loops. For each component required, the algorithm creats loadings, scores for $\mathbf{X}$ and $\mathbf{Y}$ with coefficients estimates. After extraction of each component, the $\mathbf{X}$ and $\mathbf{Y}$ matrices are deflated ensuring that the components obtained are independent to each other.

NIPALS Algorithm

$$[\text{X0},\text{Y0}] \leftarrow \texttt{scale}(\text{X},\text{Y})$$
$$w \leftarrow \text{X}_a^t$$
$$w \leftarrow w/|w|$$
$$t \leftarrow \text{X}_0 w$$
$$p \leftarrow \text{X}_a^t t(t't)^{-1}$$
$$q \leftarrow \text{Y}_a^t t(t't)^{-1}$$
--- Deflation ---
$$\text{X}_{a+1} \leftarrow \text{X}_a - tp^t$$
$$\text{Y}_{a+1} \leftarrow \text{Y}_a - tq^t$$
--- Coefficients Estimate ---
$$\hat{\mathbf{B}} \leftarrow \mathbf{W}(\mathbf{P^tW})^{-1}\mathbf{Q}$$

References

[1] Solve Sæbø, Magni Martens, and Harald Martens. "Three-block data modeling by endo- and exo-LPLS regression". In: *Handbook of Partial Least Squares*. Springer, 2010, pp. 359–379.

[2] Solve Sæbø et al. "LPLS-regression: a method for prediction and classification under the influence of background information on predictor variables". In: *Chemometrics and Intelligent Laboratory Systems* 91.2 (2008), pp. 121–132.

[3] Svante Wold, Michael Sjöström, and Lennart Eriksson. "PLS-regression: a basic tool of chemometrics". In: *Chemometrics and intelligent laboratory systems* 58.2 (2001), pp. 109–130.

L-PLS and its varients (Exo-LPLS and Endo-LPLS)
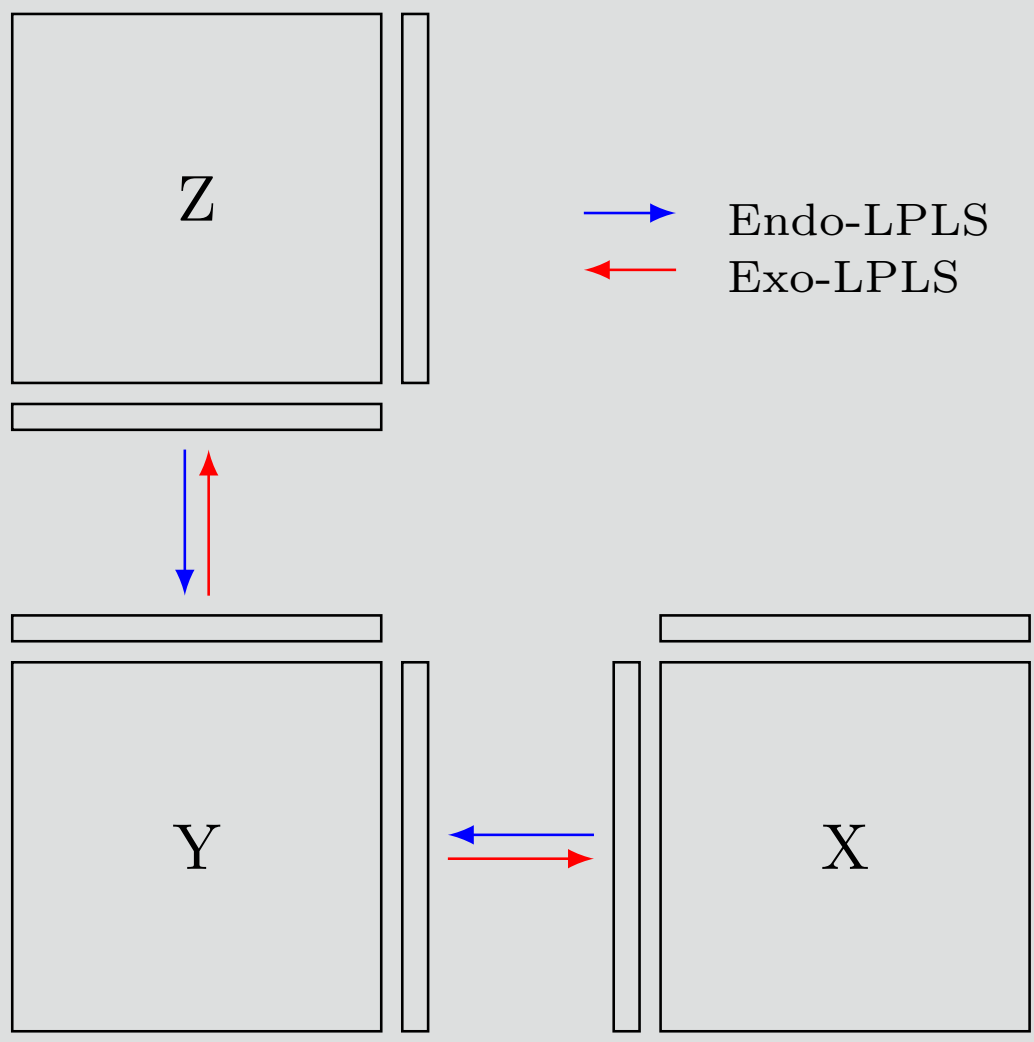

Figure: Extension of PLS including some background information

► Partial Least Square can also be used in understanding the covariance structure and the relation between different matrices.
► Information can give better insight in solving any problem. A background information on variation under study helps to model the complex dynamics of the real world phenomena.
► L-shaped PLS (LPLS) helps in exploring covariance structure of matrices
► LPLS with its variant Endo and Exo LPLS, enables to see the relationship between two matrices with no direct connection.

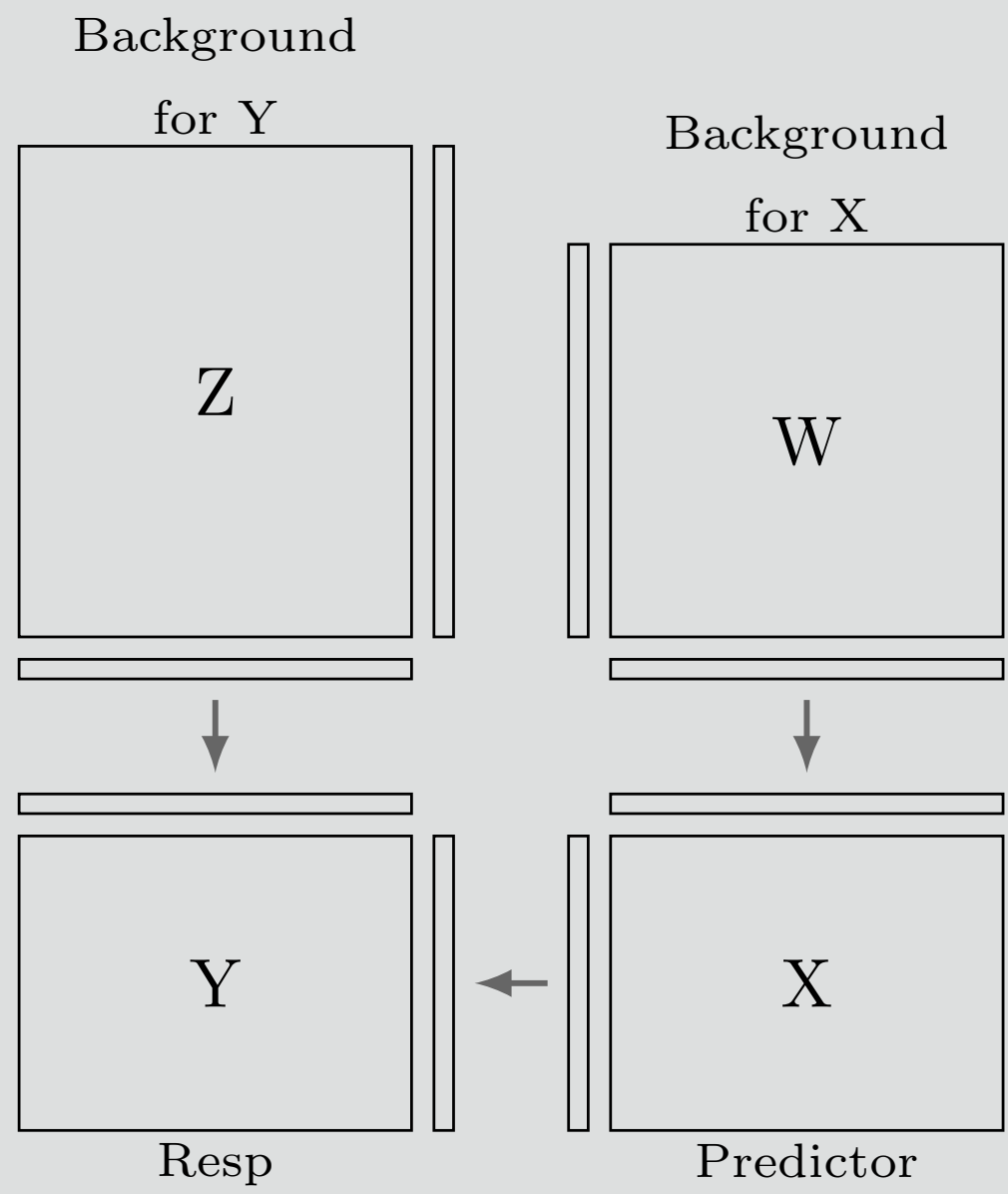## U-shaped Partial Least Square Regression (UPLS)

Figure: UPLS structure

Scenerio

► Large amount of caotic yet informative data streams, on organized, can give light on any research process
► Integraion of such easily availiable datasource can supply background information to the variables under study and helps not only to understand the causal relation and covariance structure but also to visualize in understandable form and generate new hypothesis

What is UPLS

UPLS is an extension of LPLS that enable us to,

► include background information for predictor matrix $\mathbf{X}$ and response matrix $\mathbf{Y}$
► Foresee the relationship between two matrices form entirely different situation, i.e. the relation between $\mathbf{W}$ and $\mathbf{Z}$ through the eye of $\mathbf{X}$ and $\mathbf{Y}$ (Figure alongside)

Application Example

Hypothesis

► The personality type of a student have influence on there performance on the courses.
► The better/poor score in a course of a student may be due to the inappropriate choice of course or the teaching methods which does not corresponds to the personality types of that student.

Materials

► Personality test data fro 50 variables from 288 students through questionnaire ($\mathbf{X}$)
► Scores 50 students in 9 courses constitute matrix $\mathbf{Y}$
► Personality test data ($\mathbf{W}$) for 2200 persons from various fields are used as background information for $\mathbf{X}$
► Scores of 5000 students in the same 9 courses are also included as background information for $\mathbf{Y}$

Missing Value Problem

All the student considered may have taken all the 9 courses considered which creates lots of missing values. Following approaches are intended to apply,

► Ignore the missing observations within the NIPALS loop
► Replacing the missing value through interpolation or with nearest neighbour

Scores in nine courses

56 personality type