Multimatrix Extension of Partial Least Square

Raju Rimal (Supervisor: Solve Sæbo, Trygve Almøy)

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,

$$\boldsymbol{Y} = \boldsymbol{B}_\circ + \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$

Usual Problems

- Multicollinear Predictors
- High Dimensional dataset
- Large Number of predictors

Partial Least Square Regression

Partial Least Square (PLS) regression is a method that can,

- explore the underlying predictor-response relationship through latent structures
- capture maximum covariance structure between $\mathbf{X}$ and $\mathbf{Y}$ in first few principle components

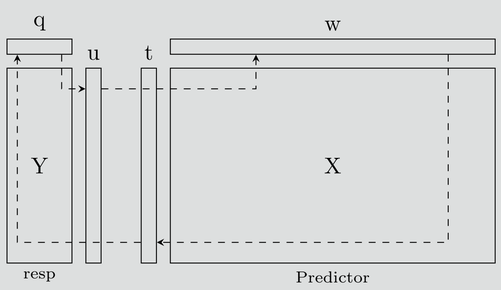

Figure: NIPALS algorithm for performing PLS regression. Starting from any arbitrary column $\mathbf{u}$, the algorithm converges after few loops. For each required components, the algorithm creates loadings and scores for $\mathbf{X}$ and $\mathbf{Y}$ with coefficients estimates. After extraction of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

PLS Algorithm

$$
\begin{aligned}
[\mathtt{X0}, \mathtt{Y0}] &\leftarrow \mathtt{scale}(\mathtt{X}, \mathtt{Y}) \\
\mathtt{w} &\leftarrow \mathtt{Obtained\ from} \\
&\quad\ \mathtt{NIPALS\ algorithm} \\
\mathtt{w} &\leftarrow \mathtt{w}/|\mathtt{w}| \\
\mathtt{t} &\leftarrow \mathtt{X}_0\mathtt{w} \\
\mathtt{p} &\leftarrow \mathtt{X}_\mathtt{a}^\mathtt{t}\mathtt{t}(\mathtt{t}'\mathtt{t})^{-1} \\
\mathtt{q} &\leftarrow \mathtt{Y}_\mathtt{a}^\mathtt{t}\mathtt{t}(\mathtt{t}'\mathtt{t})^{-1} \\
\text{--- Deflation ---} & \\
\mathtt{X}_{\mathtt{a}+1} &\leftarrow \mathtt{X}_\mathtt{a} - \mathtt{t}\mathtt{p}^\mathtt{t} \\
\mathtt{Y}_{\mathtt{a}+1} &\leftarrow \mathtt{Y}_\mathtt{a} - \mathtt{t}\mathtt{q}^\mathtt{t} \\
\text{--- Coefficients Estimate ---} & \\
\hat{\mathbf{B}} &\leftarrow \mathbf{W}(\mathbf{P}^\mathbf{t}\mathbf{W})^{-1}\mathbf{Q}
\end{aligned}
$$

References

[1] H Martens. *Domino PLS: a framework for multi-directional Path Modelling*. 2005, pp. 125–132.

[2] Solve Sæbø, Trygve Almøy, and Helge Brovold. "Does academia disfavor contextual and extraverted students?" In: *Uniped* 38.4-2015 (2015), pp. 274–283.

[3] 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.

[4] 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 variants (Exo-LPLS and Endo-LPLS)

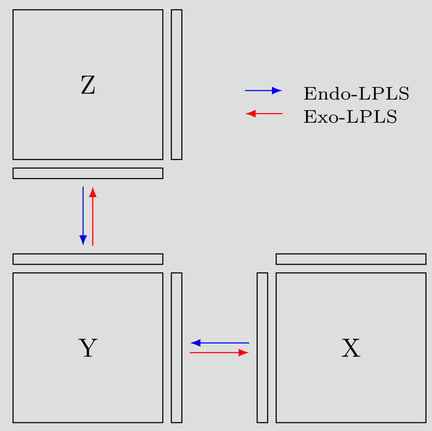

Figure: Extension of PLS including some background information

- Partial Least Square can also be used in understanding the covariance structure and the relation between different matrices.
- More information can give better insight in solving any problem. So, 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 three 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)

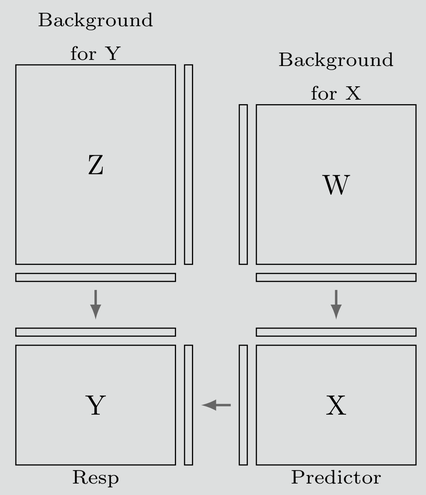

Figure: UPLS structure

Scenario

- Large amount of chaotic yet informative data streams, if organized, can give light on any research process
- Integration of such easily available data source can supply background information to the variables under study and helps, not only to understand the causal relationship and covariance structure, but also to visualize them 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 that are from entirely different situations, i.e. the relation between $\mathbf{W}$ and $\mathbf{Z}$ through the eye of $\mathbf{X}$ and $\mathbf{Y}$ (Figure above)

Application Example

Hypothesis

- The personality-type of students have influence on their performance.
- The better/poor score in a course of a student may be due to the inappropriate choice of teaching methods which does not correspond to their personality-types.

Materials

- Personality test data for 50 variables from 288 students through questionnaire ($\mathbf{X}$)
- Scores of 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 have not taken all the 9 courses 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 neighbor