Madeleine Oudin Lab at Tufts University

Biomedical Engineering Department

**PLSR in Matlab: “PLSRCode.m”**

* Outputs: variance explained in X and Y per PC; VIP scores; PLS loading and scores plots; prediction of new data condition
* Section 1 – Loading the data
  + X0i – unscaled cell shape values (n-by-m matrix)
  + Y0i – unscaled cell response values (n-by-p matrix)
* Section 2 – Run PLS
  + Set number of principal components [default is the min(no. ECM conditions -1 or no. X vars -1)

[Xloadings,Yloadings,Xscores,Yscores,beta,PctVar,mse,stats] = plsregress (X0,Y0,Pcomps);

* + X loading: p-by-pcomp matrix; coefficients define linear combination of the PCs that approximate the original predictor variables (cell shape parameters)
  + Y loading: m-by-pcomp matrix; coefficients define linear combination of the PCs that approximate the original response variables (speed, persistence, invasion)
    - We can plot the X and Y loadings to see how the different cell shape parameters and cell responses map out together
  + X scores: n-by-pcomp matrix; map out each ECM factor on PCs based on cell shape parameters
  + Y scores: n-by-pcomp matrix; map out each ECM factor on PCs based on both cell shape and cell response parameters
    - We can plot Y scores to see how the model relates the different ECMs
  + Beta: regression coefficients, which are used to make predictions
  + PctVar: percent variation explained from each PC (row 1 for X and row 2 for Y)
  + Mse: estimated mean square error of the model (row 1 for X and row 2 for Y)
  + Stats: structure with different statistics (we use this to calculate the VIP score)
* Section 3 – Plot Variance explained by each PC
  + R^2 is the cumulative variance explained in Y
* Section 4 – Plot Loadings and Scores
* Section 5 – Calculate VIP Scores
  + Need to input the number of PCs to use in calculation (ideal model)
  + Calculate VIP scores for predicting each cell response separately
  + Formula for VIP Score: A close up of a logo

    Description automatically generated

Janes KA, Reinhardt HC, Yaffe MB. Cytokine-induced signaling networks prioritize dynamic range over signal strength. Cell. 2008;135(2):343-54. Epub 2008/10/30. doi: 10.1016/j.cell.2008.08.034. PubMed PMID: 18957207; PMCID: PMC2635014.

* Section 6 – Predict new ECM
  + Define new data
  + Ypred variable outputs the model’s prediction
* More resources for using plsregress on Matlab:
  + <https://faculty.virginia.edu/lazzara/protocols/PLSR%20using%20MATLAB.pdf>
  + <https://www.mathworks.com/help/stats/plsregress.html>

**Model Evaluation: cross validation and Q2 metric**

* “PLS\_CV.m”, which calls functions “CrossValFunc.m” and “q2calc.m”
* Output: Predictions from cross validation; Q2 values
* Section 1 – Load Data
* Section 2 – Leave One Out Cross Validation
  + Leaves out one of the rows, run PLS model and predict the left out row, repeats this until each row has been left out and predicted
  + Ypred variable: if predicting all 3 responses, column 1-3 are with PC1, 4-6 with PC2, 7-9 with PC3, and 10-12 with PC4
* Section 3 - Calculate Q2
  + - Gaudet S, Janes KA, Albeck JG, Pace EA, Lauffenburger DA, Sorger PK. A compendium of signals and responses triggered by prodeath and prosurvival cytokines. Mol Cell Proteomics. 2005;4(10):1569-90. Epub 2005/07/21. doi: 10.1074/mcp.M500158-MCP200. PubMed PMID: 16030008.
* More information: <https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Leave-one-out_cross-validation>
* <https://stats.stackexchange.com/questions/292673/validation-metrics-r2-and-q2-for-partial-least-squares-pls-regression>