Partial Least Squares: When Ordinary Least Squares Regression Just Won't Work

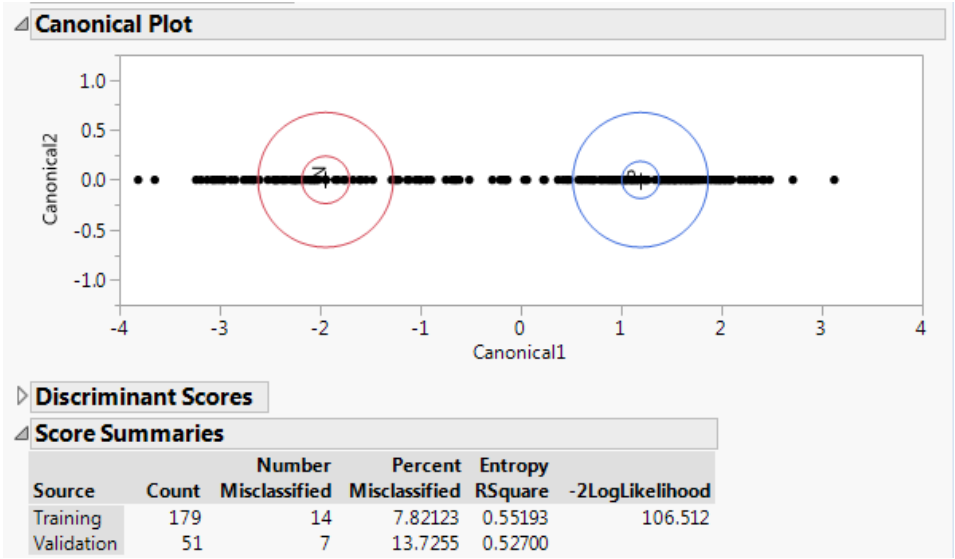
Peter Bartell

JMP Senior Systems Engineer

Peter.bartell@jmp.com

When OLS Just Won't Work

- OLS (Ordinary Least Squares) in JMP/JMP Pro = Fit Model -> Standard Least Squares.
- 230 rows…almost 11,000 columns…can we create a model that can be used to classify a person as 'positive' or 'negative' for a potentially life altering condition?



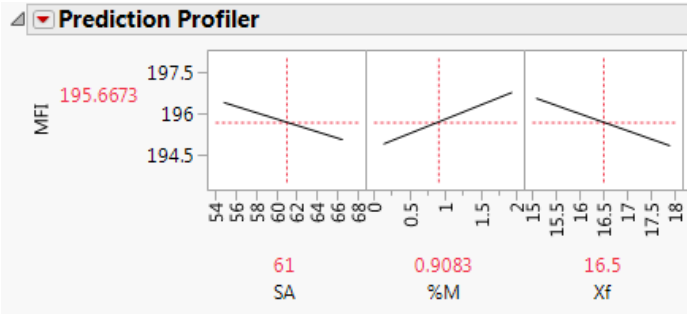
Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	179	14	7.82123	0.55193	106.512
Validation	51	7	13.7255	0.52700	

Objectives

- At the end of this presentation you will be able to
  - List the practical situations when partial least squares regression (PLS) is a viable option for empirical modeling.
  - List the key steps in using JMP and JMP Pro to construct, evaluate and use PLS models.
  - List resources to learn more about partial least squares in JMP and JMP Pro.

What is Partial Least Squares?

- An empirical linear modeling technique, y = f(x), that ***leverages correlation/covariance in x and y variables.***
  - Introduced and formalized by Herman and Svante Wold and others beginning in the 1960's.
- Scenarios for use.
  - Most commonly used with historical/happenstance or pre-existing data.
    - Process optimization, control strategies, variable identification.
  - Precursor to Design of Experiments.
    - Variable identification/elimination, factor ranges, functional relationship to responses.



Practical Situation for PLS

- High degree of correlation (multicollinearity) between and among the x and or y variables.
  - Historical process data.

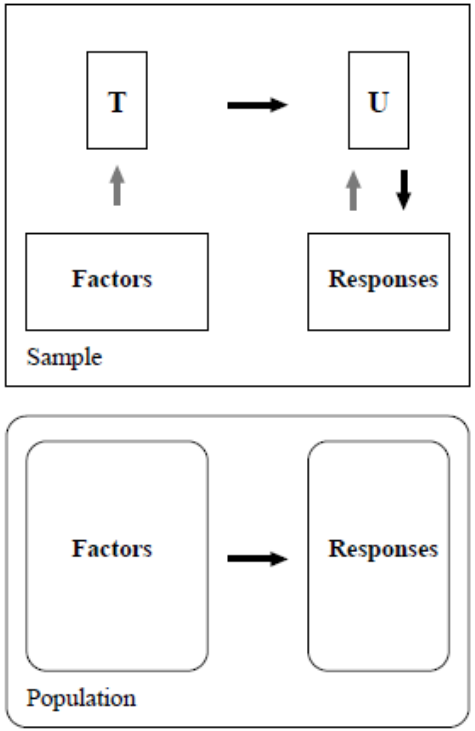
- The wide (x) and shallow (y) situation.
  - Many more x variables than there are observations of y (the response).
  - Common in batch processes.

Practical Situations for PLS

- An ENORMOUS number of x variables…hundreds to thousands! Dimensionality/variable reduction is key focus.

How Does PLS Work?

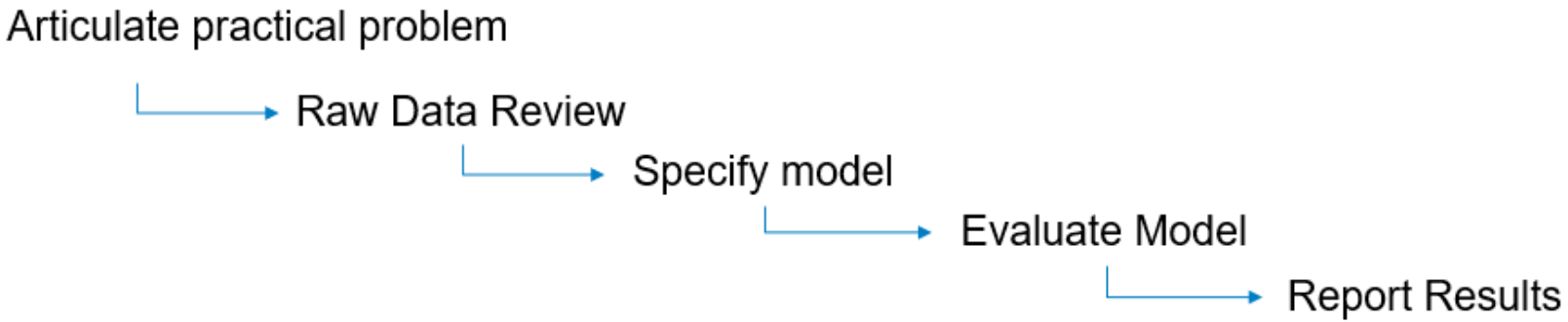
- A nonmathematical view…
  - Latent structures (variables) are at the heart.
    - Latent variables are created from the original variables.
      - A projection of the original variables…PLS (projection to latent structures).
    - Linear models are fit using these latent variables…then used for the intended purpose.
      Variable identification, optimization, control, etc.



"An Introduction to Partial Least Squares", Tobias, SAS Institute

PLS Implementation in JMP/JMP Pro

- Three PLS methods in JMP and JMP Pro.
  - PLS – Discriminant Analysis (categorical responses).
  - NIPALS (nonlinear iterative partial least squares).
  - SIMPLS (statistically inspired modification of PLS).
  - For a single response NIPALS and SIMPLS methods yield identical results.
- Similar workflow to other JMP modeling platforms.

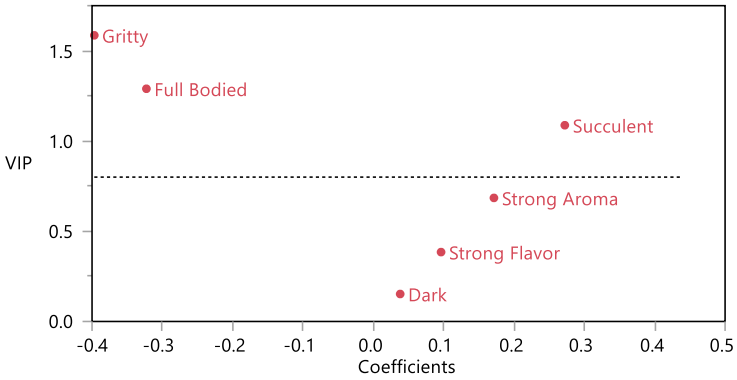


PLS in JMP Pro

- It’s own Analyze -> Fit Model personality.
  - In JMP, PLS is accessible through Analyze -> Multivariate Methods -> Partial Least Squares.
  - Will NOT show the JMP workflow…ONLY the JMP Pro workflow today.
- Fit responses with nominal or continuous JMP data type.
- Fit polynomial, interaction, and categorical effects.
- Larger set of validation and cross validation methods.
  - Train/validate/test construct, Kfold, Leave One Out, Holdback %.
- Optional missing data imputation.
- Bootstrap estimates of distributions of select statistics.

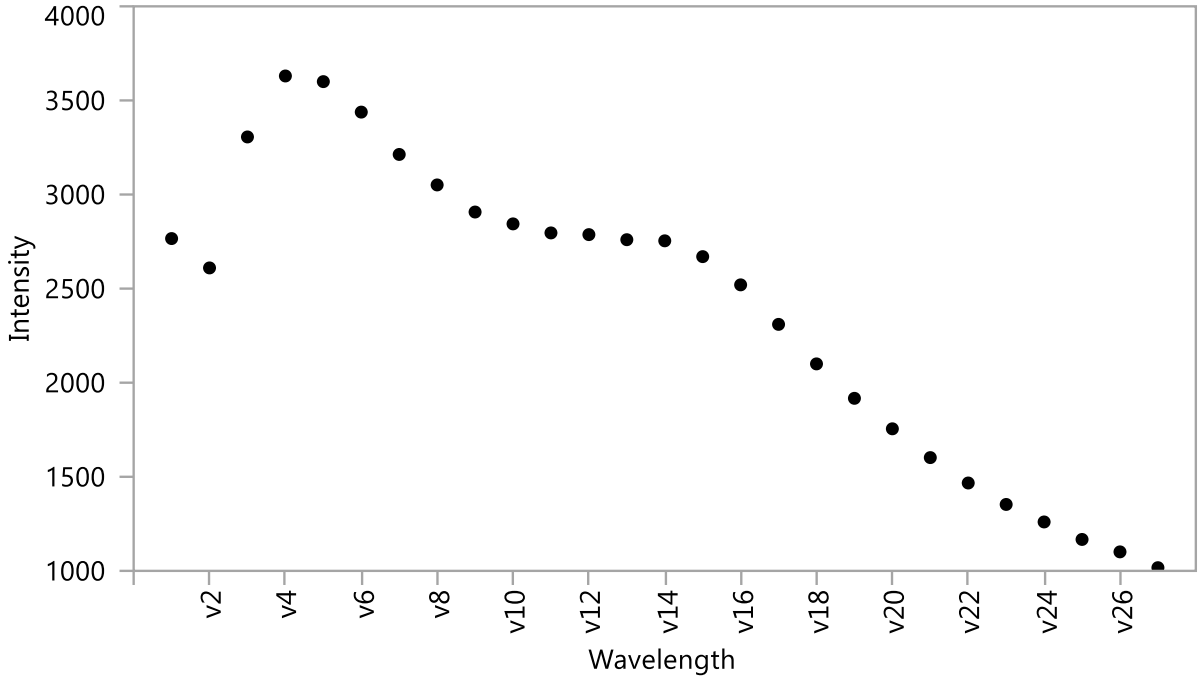
Case #1 – Making Great Bread

- Inspired by chapter 8 of “Discovering Partial Least Squares using JMP®” by Cox and Gaudard.
- The problem at hand…
  - Can we identify product attributes that help guide product formulation and design processes?
- 50 participants on a consumer panel rate 24 types of bread on a ‘likability’ scale (y) using ratings of 6 (x’s) attributes.



Case #2 – Lots of Multicollinearity

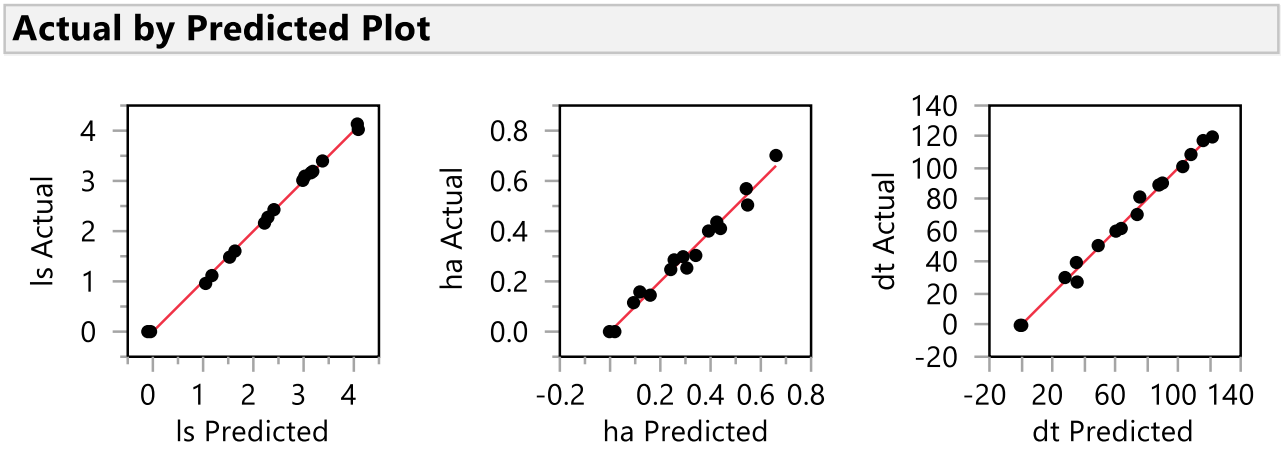
- PLS is especially valuable in the spectral absorption data scenario.



Case #2: Lots of Multicollinearity

- Problem: Can we create a useful model for evaluating the levels of 3 different compounds (ls, ha, dt) based on spectral emissions of samples drawn from a known population?

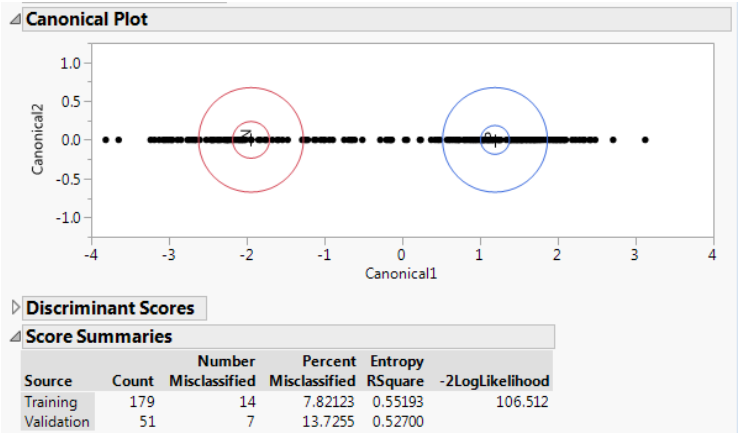
Actual vs. Predicted Plots for ls, ha, dt



Uses Baltic.jmp from the JMP Sample Data Directory

Case #3: PLS – Discriminant Analysis

- PLS – Discriminant Analysis is found at Analyze -> Multivariate Methods -> Partial Least Squares.
  - Micro Array Quality Control Study for classification of individuals based on gene expression characterization.
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3315840/
    - Problem: Can genetic expression data be used to accurately classify estrogen receptor status?
- 230 individuals in the study…over 10,000 co-variates (gene expression characteristics).



Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	179	14	7.82123	0.55193	106.512
Validation	51	7	13.7255	0.52700	

To Learn More

- SAS Education courses:
  - "JMP® Software: Finding Important Predictors"
    - https://support.sas.com/edu/schedules.html?ctry=us&crs=JFIP#s1=3
  - JMP® Software: Analyzing and Modeling Multidimensional Data
    - https://support.sas.com/edu/schedules.html?ctry=us&crs=JAMMD#s1=3
- "Discovering Partial Least Squares with JMP®" – Cox and Gaudard