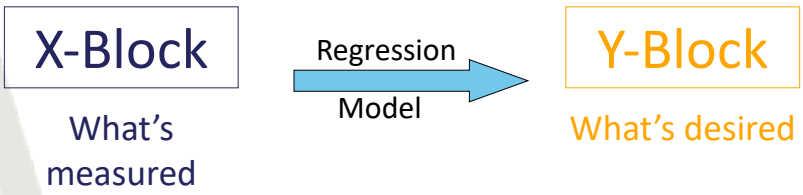
# Building a PLS Regression Model



EIGENVECTOR RESEARCH INCORPORATED

## Regression




- Regression analysis creates a mapping between two blocks of data.
- In contrast, PCA was used to explore the correlation structure within a single data block.
- Regression models are often used to obtain estimates (or predictions) for one block of data from the other.



EIGENVECTOR RESEARCH INCORPORATED

# Outline

- Inverse Least Squares (ILS) Models
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
  - Cross-validation
- Partial Least Squares Regression (PLS)
  - Model Quality Measures
  - Determining of the Number of factors
  - Outlier Detection and Model Diagnostics
- Model Validation
- Conclusions



EIGENVECTOR RESEARCH INCORPORATED

# Inverse Least Squares

- Inverse least squares (ILS) models assume that the model is of the form:

$$\mathbf{Xb} = \mathbf{y} + \mathbf{e}$$

where:

- $\mathbf{y}$ $(M \times 1)$ is a property to be predicted
- $\mathbf{X}$ $(M \times N_x)$ is the measured response
- $\mathbf{e}$ $(M \times 1)$ is an error vector
- $\mathbf{b}$ $(N_x \times 1)$ is a vector of coefficients



EIGENVECTOR RESEARCH INCORPORATED

# Advantage of ILS Methods

- ILS methods (including MLR, PCR, PLS, CR) don't require the concentration of all analytes, including interferents, be known …
- …however, interferents must vary in the calibration data set for the ILS regression model to be robust against them
  - this has important implications for experimental design
  - *Interferent: Any substance whose presence interferes with an analytical procedure and generates incorrect results*
  - interferents that correlate with an analyte during calibration but don't in the future are to be avoided. Don't confound your design!



EIGENVECTOR RESEARCH INCORPORATED

# Estimation of b: MLR

- It is possible to estimate **b** from

$$\mathbf{b} = \mathbf{X}^{+}\mathbf{y}$$

  where $\mathbf{X}^{+}$ is the pseudo-inverse of $\mathbf{X}$
- There are many ways to obtain a pseudo-inverse; the most obvious is multiple linear regression (MLR), a.k.a., ordinary Least Squares (OLS)
- In this case, $\mathbf{X}^{+}$ is estimated from

$$\mathbf{X}^{+} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$$



EIGENVECTOR RESEARCH INCORPORATED

# Problem with MLR

- Inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ only exists if …
  a) $\mathrm{rank}(\mathbf{X}) = N_{\mathrm{x}}$
     - however $\mathrm{rank}(\mathbf{X}) \leq \min(M, N_{\mathrm{x}})$
  b) $\mathbf{X}$ has more samples than variables *i.e.,* if $M > N_{\mathrm{x}}$, and
     - this is a problem with spectra
  c) columns of $\mathbf{X}$ are not co-linear.
- Inverse may exist but be highly unstable if $\mathbf{X}$ is nearly rank deficient (a.k.a., ill-conditioned).
  - In these cases, small perturbations in the data (possibly due to noise) can produce very different results.



EIGENVECTOR RESEARCH INCORPORATED

# Principal Components Regression

- Principal Components Regression (PCR) is one way to deal with
  - Having more variables (wavelengths) than samples
  - Ill-conditioned problems, *i.e.* highly correlated variables
- Property of interest $\mathbf{y}$ is regressed on PCA scores:

$$\mathbf{X}^{+} = \mathbf{P}_K\left(\mathbf{T}_K^{\mathrm{T}}\mathbf{T}_K\right)^{-1}\mathbf{T}_K^{\mathrm{T}}$$

- Problem is to determine *K* the number of factors to retain in the formation of the model



EIGENVECTOR RESEARCH INCORPORATED

# Model Quality Measures

- Root Mean Square Error (RMSE) Metrics
  - RMSE**Calibration**
  - RMSE**Cross-Validation**
  - RMSE**Prediction**
  - In units of the $y$ variable
- Correlation Coefficient ($r$)
  - Unit-less
  - Considers the range of $y$

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(y_m - \hat{y}_m)^2}$$

$$r = \frac{\sum_{m=1}^{M}(\hat{y}_m - \hat{y})(y_m - y)}{\sqrt{\sum_{m=1}^{M}(\hat{y}_m - \hat{y})^2}\sqrt{\sum_{m=1}^{M}(y_m - y)^2}} \equiv \frac{\sigma_{\hat{y}y}}{\sigma_{\hat{y}}\sigma_y}$$

54

EIGENVECTOR RESEARCH INCORPORATED

# Cross-Validation

- Divide data set into $J$ sample subsets
- For ***each subset*** ($j = 1, \ldots, J$):
  - Build regression model using samples in the ***remaining*** subsets
  - Apply the model to subset $j$ samples
  - Calculate PRESS (Predictive Residual Error Sum of Squares) for the subset samples:

$$\mathbf{e}_j^2 = (\mathbf{y} - \mathbf{Xb})_j^2$$

- Look for minimum or "knee" in cumulative PRESS curve

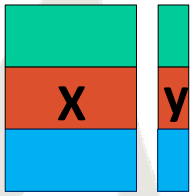
$$\text{RMSECV} = \left(\frac{1}{M}\sum_{j=1}^{J}\mathbf{e}_j^2\right)^{1/2}$$

55

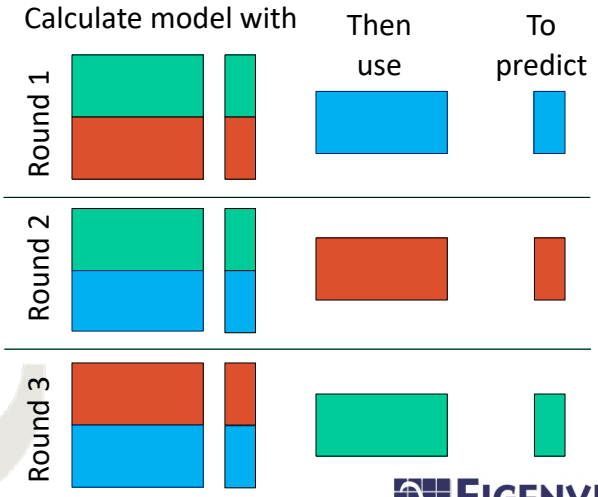
EIGENVECTOR RESEARCH INCORPORATED

# Cross-Validation Graphically

Break data into subsets



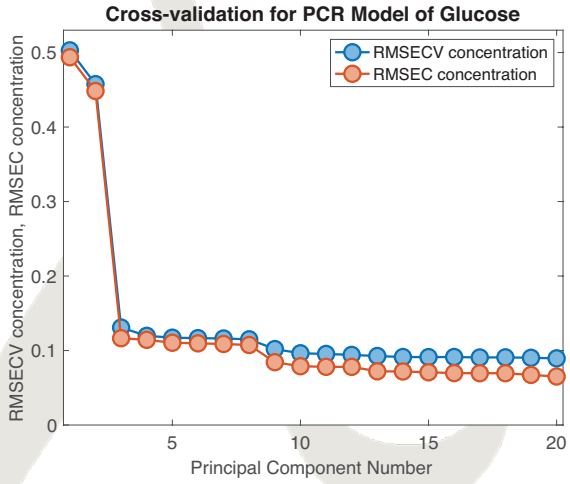

Calculate prediction error for each subset as a function of number of PCs (or LVs for PLS)

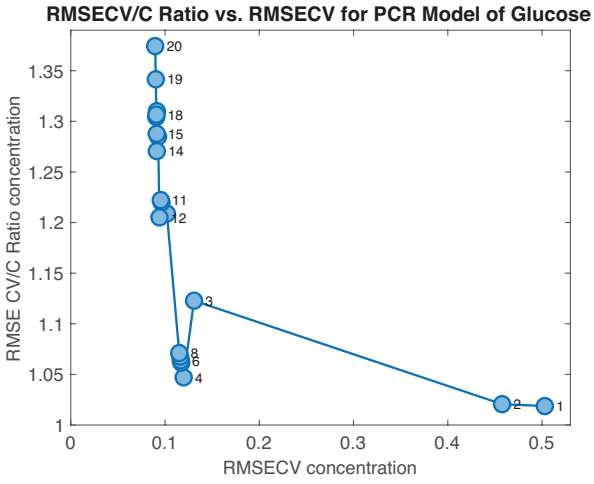



56

EIGENVECTOR RESEARCH INCORPORATED

# Typical CV & CV/C Ratio Plot




EIGENVECTOR RESEARCH INCORPORATED

# Problems with PCR

- Some PCs not relevant for prediction, but are only relevant for describing variance in $\mathbf{X}$
  - leads to local minima and increase in PRESS
- This is a result of PCs determined without using information about property to be predicted $\mathbf{y}$
- A solution is to find factors using information from both $\mathbf{y}$ and $\mathbf{X}$



EIGENVECTOR RESEARCH INCORPORATED

# Partial Least Squares

- PLS is related to PCR and MLR and can be used to finds a basis for $\mathbf{X}$
  - PCR captures maximum variance in $\mathbf{X}$
  - MLR achieves maximum correlation between $\mathbf{X}$ and $\mathbf{y}$
  - PLS tries to do both by maximizing covariance between $\mathbf{X}$ and $\mathbf{y}$
- Requires addition of weights $\mathbf{W}$ to maintain orthogonal scores
- Factors calculated sequentially by projecting $\mathbf{Y}$ through $\mathbf{X}$

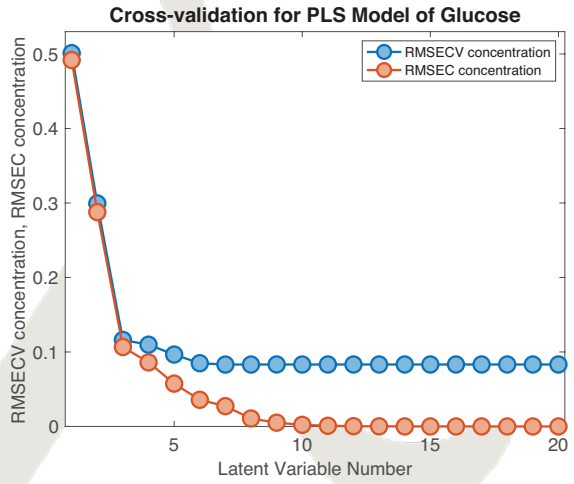
$$\mathbf{X}^{+} = \mathbf{R}_K\left(\mathbf{T}_K^{\mathrm{T}}\mathbf{T}_K\right)^{-1}\mathbf{T}_K^{\mathrm{T}} = \mathbf{W}_K\left(\mathbf{P}_K^{\mathrm{T}}\mathbf{W}_K\right)^{-1}\left(\mathbf{T}_K^{\mathrm{T}}\mathbf{T}_K\right)^{-1}\mathbf{T}_K^{\mathrm{T}}$$

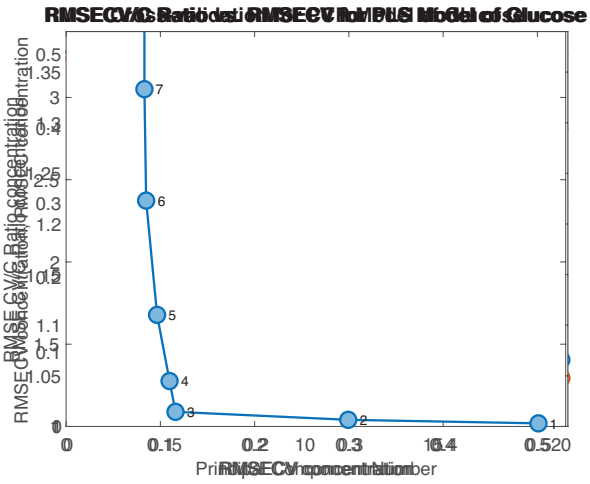
- Note that $\mathbf{P}_K$ in PLS is different than $\mathbf{P}_K$ in PCA. Here $\mathbf{T}_K = \mathbf{X}\mathbf{R}_K$.



EIGENVECTOR RESEARCH INCORPORATED

# Typical CV & CV/C Ratio Plot




60

EIGENVECTOR RESEARCH INCORPORATED

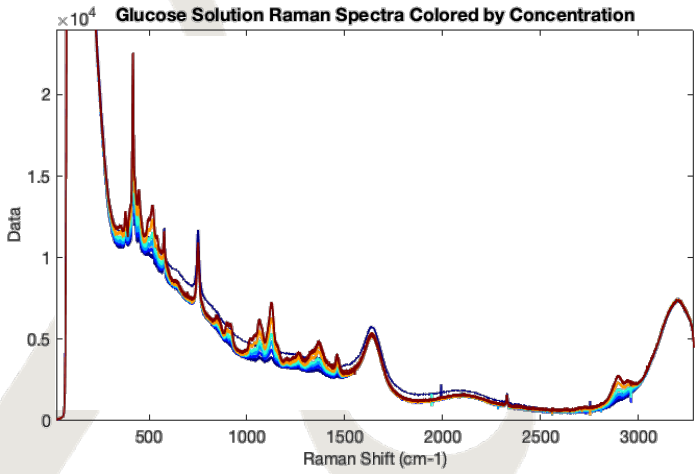
# Number of PCs or LVs

- Choice is not always simple (Future prediction is the goal, not fit.)
- A few rules of thumb
  - Be conservative, models are more often over-fit than under-fit.
  - The best choice is often not the global minimum RMSECV.
  - Look for minimum of RMSECV and work backwards if improvement is not at least 2%.
  - If RMSEC<RMSECV by more than ~20% suggests overfit.
  - Use RMSECV/RMSEC ratio plot
  - Look at the variance captured in **X** and **Y** and ask if it is significant with respect to what is known about errors in the data?

61

EIGENVECTOR RESEARCH INCORPORATED

# PLS Regression Example

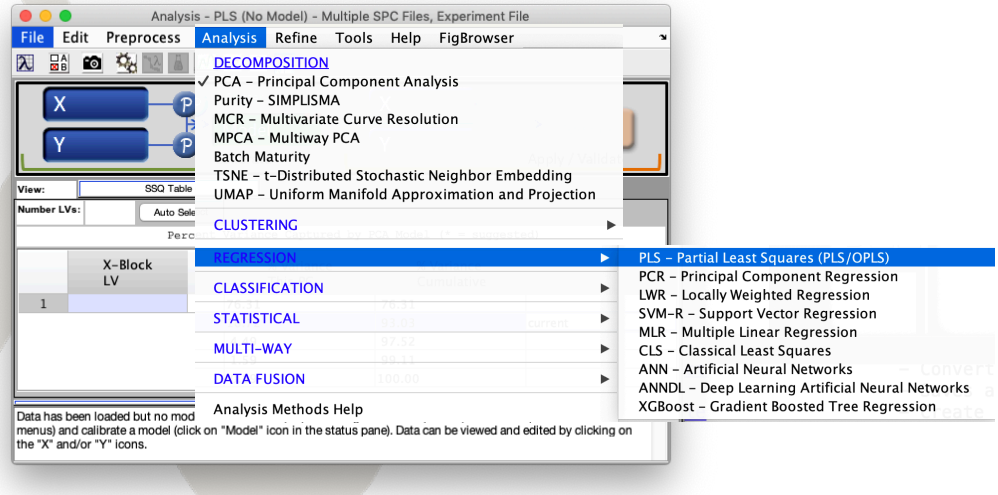



- Develop PLS model for glucose concentration from Raman spectra
  - 120 samples
  - 2048 wavelengths
- Start by selecting wavelength ranges of interest

62

EIGENVECTOR RESEARCH INCORPORATED

# Switch to PLS




EIGENVECTOR RESEARCH INCORPORATED

# Model Validation

- All predictive models, (quantitative and classification), should be validated with an *independent* test set
  - Not used in the development of the model
  - Best to collect this data on different days, different people
- Validation data should span the range of conditions expected during model application
  - This may not be the same range as the calibration set!

EIGENVECTOR RESEARCH INCORPORATED

# Regression Summary

- Regression models can be divided into CLS (used when pure analyte spectra are available) and ILS models (MLR, PCR, PLS, RR, CR, ...)
- PCR and PLS work with ill-conditioned data by reducing to a smaller number of factors
  - has advantage of signal averaging
- Cross-validation is used to determine number of factors
- Fit and Prediction are two different things
- Models should be validated!



EIGENVECTOR RESEARCH INCORPORATED

# Model Development

- Developing PCR or PLS models
  - center and scale the data (as appropriate)
  - cross-validate to determine number of factors
  - check **X**-block Q, $T^2$, leverage, and **Y**-block residuals for outliers
    - remove / explain outliers
  - check RMSEC and RMSECV values for overfit
  - Examine models scores and loadings for interesting trends
  - repeat as necessary



EIGENVECTOR RESEARCH INCORPORATED

# Model Application

- A PCR or PLS model is applied by
  - centering and scaling to the model mean and variance
  - multiply measurements by regression vector to get scaled predictions
  - rescale the predictions back to original units using model mean and variance
- Prediction outliers can be found by
  - calculating Q and $T^2$ values for new samples
- All the modeling and application is packaged:
  - the model is an object that contains all the parameters



EIGENVECTOR RESEARCH INCORPORATED

# Regression Resources

- Webinars
  - https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-about-how-to-do-partial-least-squares-regression/
  - https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-about-cross-validation/
- Eigenvector University courses live and recorded
  - https://eigenvector.com/training/short-course-topics/chemometrics-ii-regression-and-pls/

EIGENVECTOR RESEARCH INCORPORATED