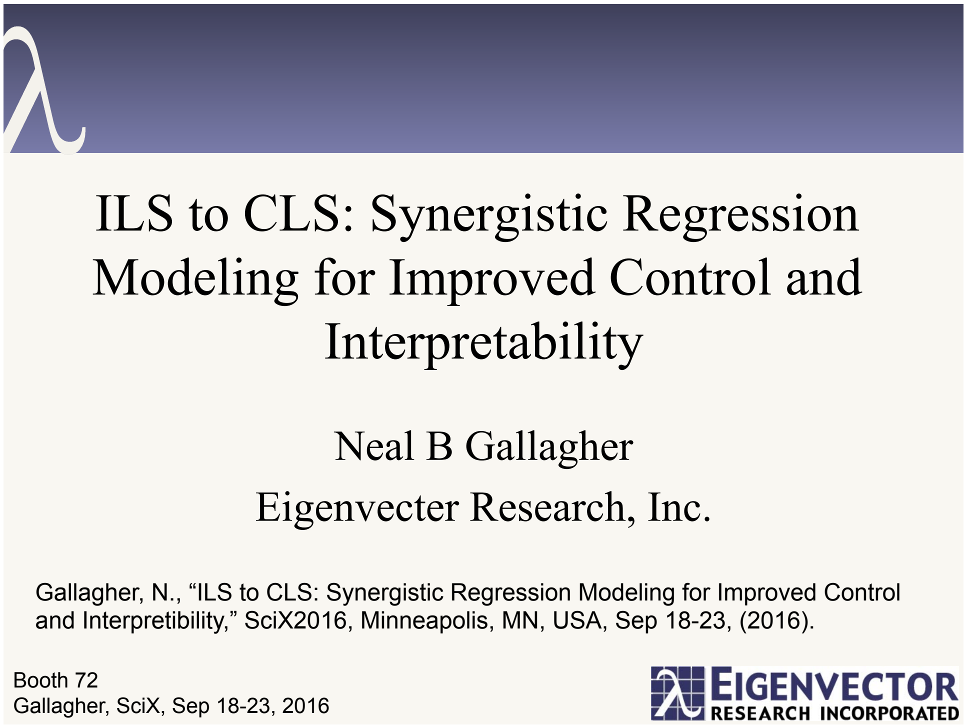


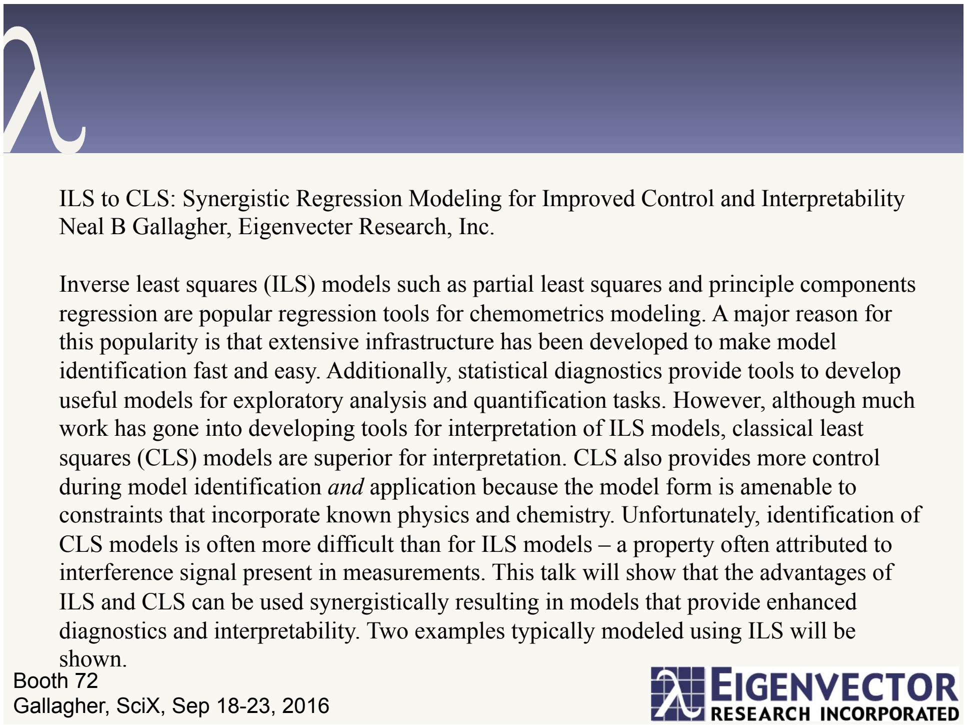
ILS to CLS: Synergistic Regression Modeling for Improved Control and Interpretability

Neal B Gallagher

Eigenvecter Research, Inc.

Gallagher, N., "ILS to CLS: Synergistic Regression Modeling for Improved Control and Interpretibility," SciX2016, Minneapolis, MN, USA, Sep 18-23, (2016).

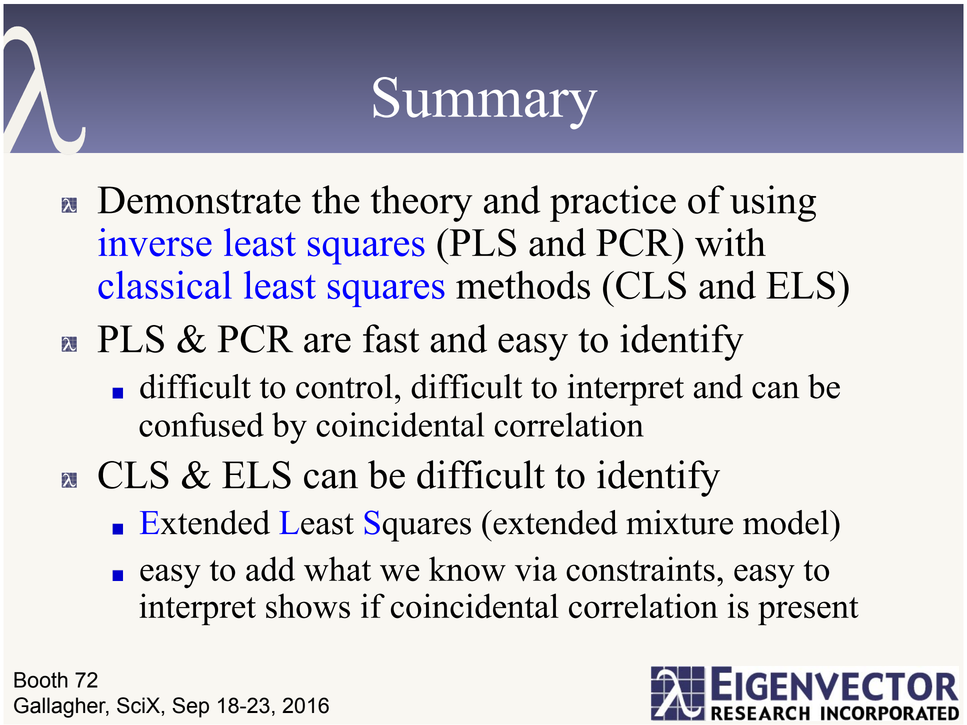
Booth 72
Gallagher, SciX, Sep 18-23, 2016



ILS to CLS: Synergistic Regression Modeling for Improved Control and Interpretability
Neal B Gallagher, Eigenvecter Research, Inc.

Inverse least squares (ILS) models such as partial least squares and principle components regression are popular regression tools for chemometrics modeling. A major reason for this popularity is that extensive infrastructure has been developed to make model identification fast and easy. Additionally, statistical diagnostics provide tools to develop useful models for exploratory analysis and quantification tasks. However, although much work has gone into developing tools for interpretation of ILS models, classical least squares (CLS) models are superior for interpretation. CLS also provides more control during model identification *and* application because the model form is amenable to constraints that incorporate known physics and chemistry. Unfortunately, identification of CLS models is often more difficult than for ILS models – a property often attributed to interference signal present in measurements. This talk will show that the advantages of ILS and CLS can be used synergistically resulting in models that provide enhanced diagnostics and interpretability. Two examples typically modeled using ILS will be shown.

Booth 72
Gallagher, SciX, Sep 18-23, 2016

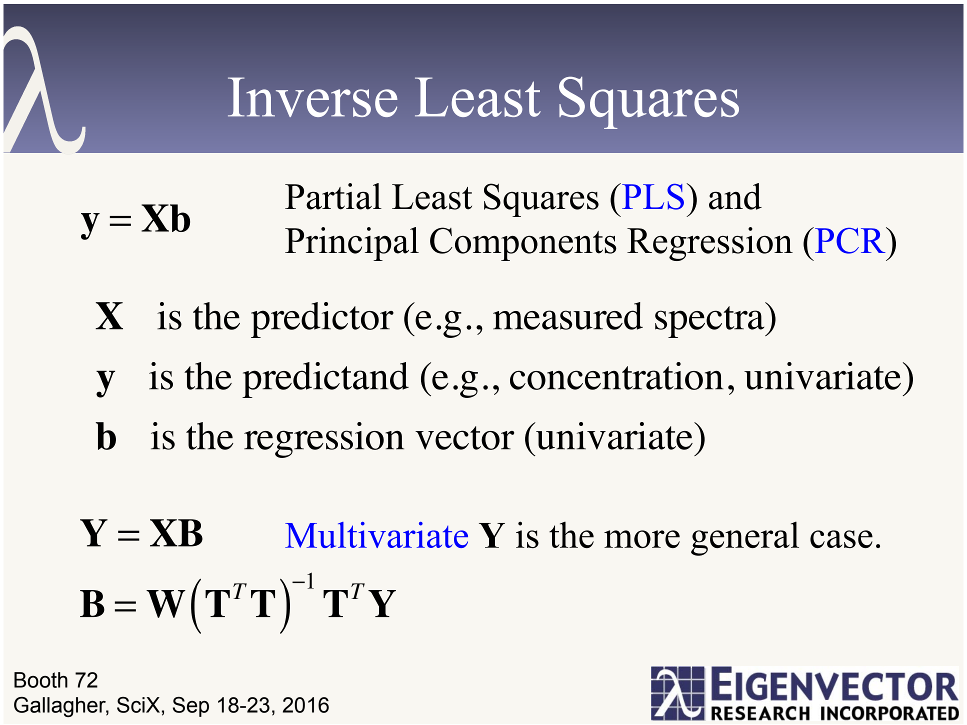


Summary

- Demonstrate the theory and practice of using inverse least squares (PLS and PCR) with classical least squares methods (CLS and ELS)
- PLS & PCR are fast and easy to identify
  - difficult to control, difficult to interpret and can be confused by coincidental correlation
- CLS & ELS can be difficult to identify
  - Extended Least Squares (extended mixture model)
  - easy to add what we know via constraints, easy to interpret shows if coincidental correlation is present

Booth 72
Gallagher, SciX, Sep 18-23, 2016

EIGENVECTOR RESEARCH INCORPORATED



Inverse Least Squares

$\mathbf{y} = \mathbf{Xb}$ Partial Least Squares (PLS) and Principal Components Regression (PCR)

X is the predictor (e.g., measured spectra)

y is the predictand (e.g., concentration, univariate)

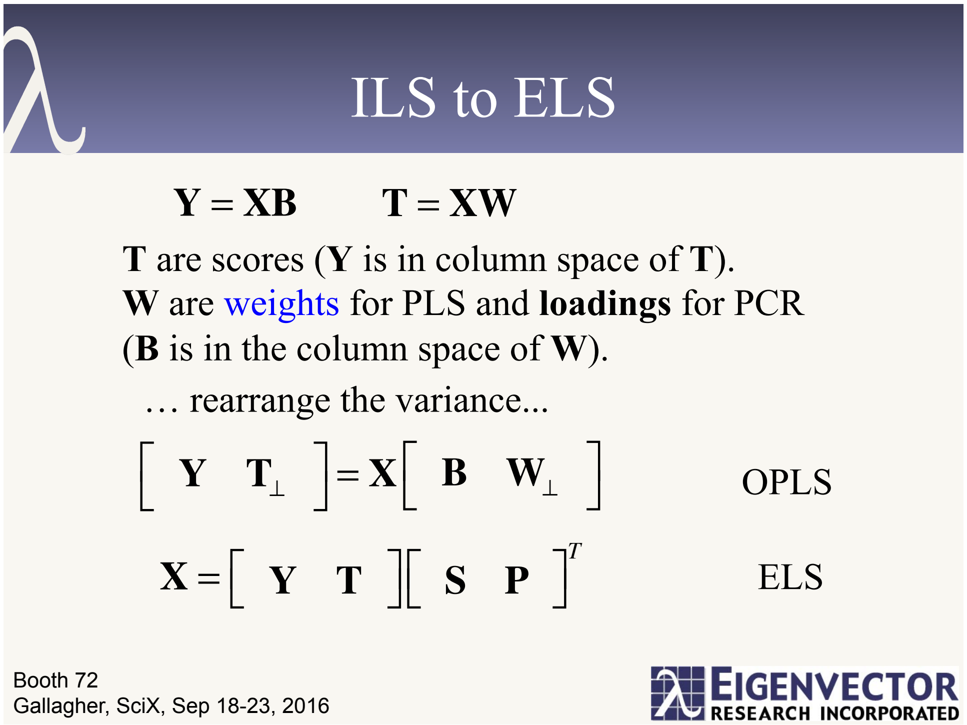
b is the regression vector (univariate)

$\mathbf{Y} = \mathbf{XB}$ Multivariate **Y** is the more general case.

$$\mathbf{B} = \mathbf{W}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y}$$

Booth 72
Gallagher, SciX, Sep 18-23, 2016

EIGENVECTOR RESEARCH INCORPORATED



ILS to ELS

$$\mathbf{Y} = \mathbf{X}\mathbf{B} \qquad \mathbf{T} = \mathbf{X}\mathbf{W}$$

T are scores (**Y** is in column space of **T**).
W are weights for PLS and **loadings** for PCR
(**B** is in the column space of **W**).

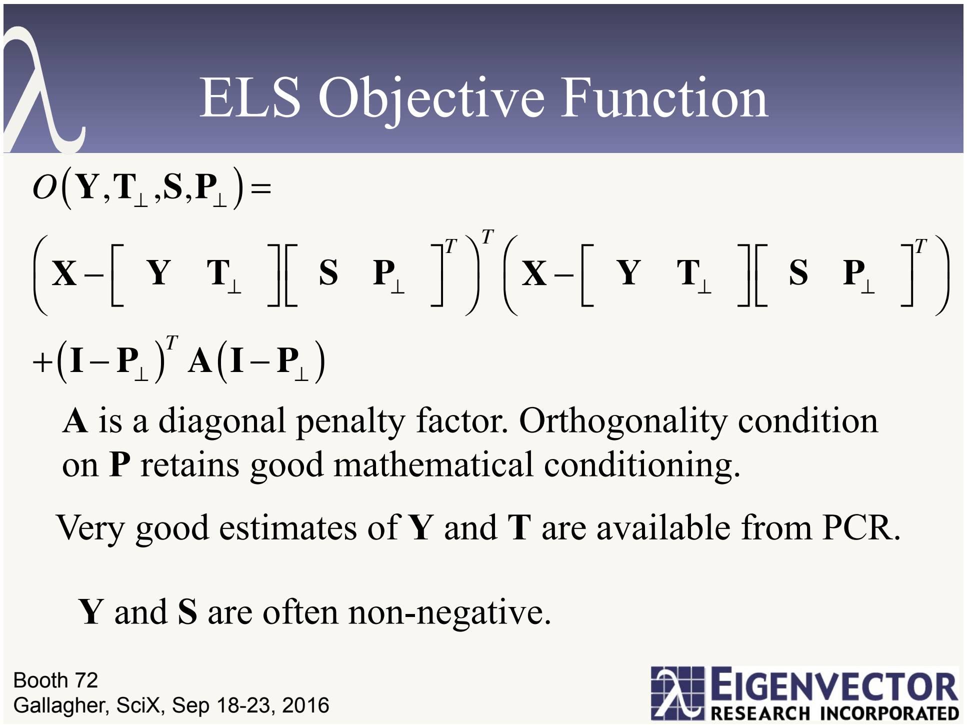
… rearrange the variance...

$$\begin{bmatrix} \mathbf{Y} & \mathbf{T}_\perp \end{bmatrix} = \mathbf{X}\begin{bmatrix} \mathbf{B} & \mathbf{W}_\perp \end{bmatrix}$$ OPLS

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} & \mathbf{T} \end{bmatrix}\begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix}^T$$ ELS

Booth 72
Gallagher, SciX, Sep 18-23, 2016

EIGENVECTOR RESEARCH INCORPORATED



ELS Objective Function

$$O\left(\mathbf{Y},\mathbf{T}_\perp,\mathbf{S},\mathbf{P}_\perp\right) =$$

$$\left(\mathbf{X} - \begin{bmatrix} \mathbf{Y} & \mathbf{T}_\perp \end{bmatrix}\begin{bmatrix} \mathbf{S} & \mathbf{P}_\perp \end{bmatrix}^T\right)^T\left(\mathbf{X} - \begin{bmatrix} \mathbf{Y} & \mathbf{T}_\perp \end{bmatrix}\begin{bmatrix} \mathbf{S} & \mathbf{P}_\perp \end{bmatrix}^T\right)$$

$$+\left(\mathbf{I} - \mathbf{P}_\perp\right)^T \mathbf{A}\left(\mathbf{I} - \mathbf{P}_\perp\right)$$

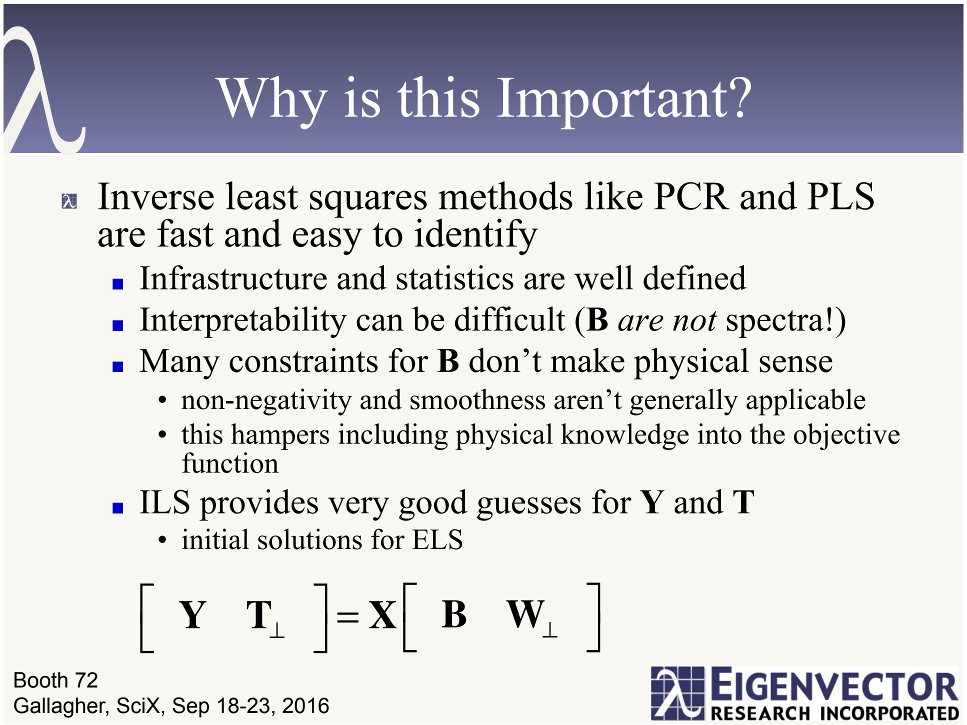
A is a diagonal penalty factor. Orthogonality condition on **P** retains good mathematical conditioning.

Very good estimates of **Y** and **T** are available from PCR.

Y and **S** are often non-negative.

Booth 72
Gallagher, SciX, Sep 18-23, 2016

EIGENVECTOR RESEARCH INCORPORATED

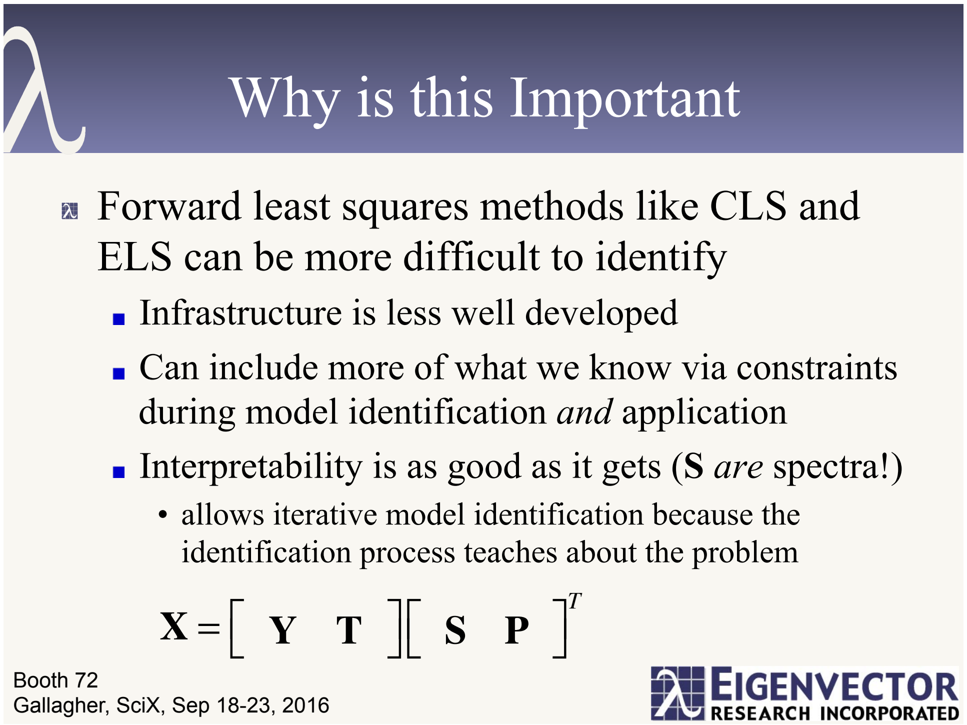


Why is this Important?

- Inverse least squares methods like PCR and PLS are fast and easy to identify
  - Infrastructure and statistics are well defined
  - Interpretability can be difficult (**B** *are not* spectra!)
  - Many constraints for **B** don't make physical sense
    - non-negativity and smoothness aren't generally applicable
    - this hampers including physical knowledge into the objective function
  - ILS provides very good guesses for **Y** and **T**
    - initial solutions for ELS

$$\begin{bmatrix} \mathbf{Y} & \mathbf{T}_\perp \end{bmatrix} = \mathbf{X}\begin{bmatrix} \mathbf{B} & \mathbf{W}_\perp \end{bmatrix}$$

Booth 72
Gallagher, SciX, Sep 18-23, 2016

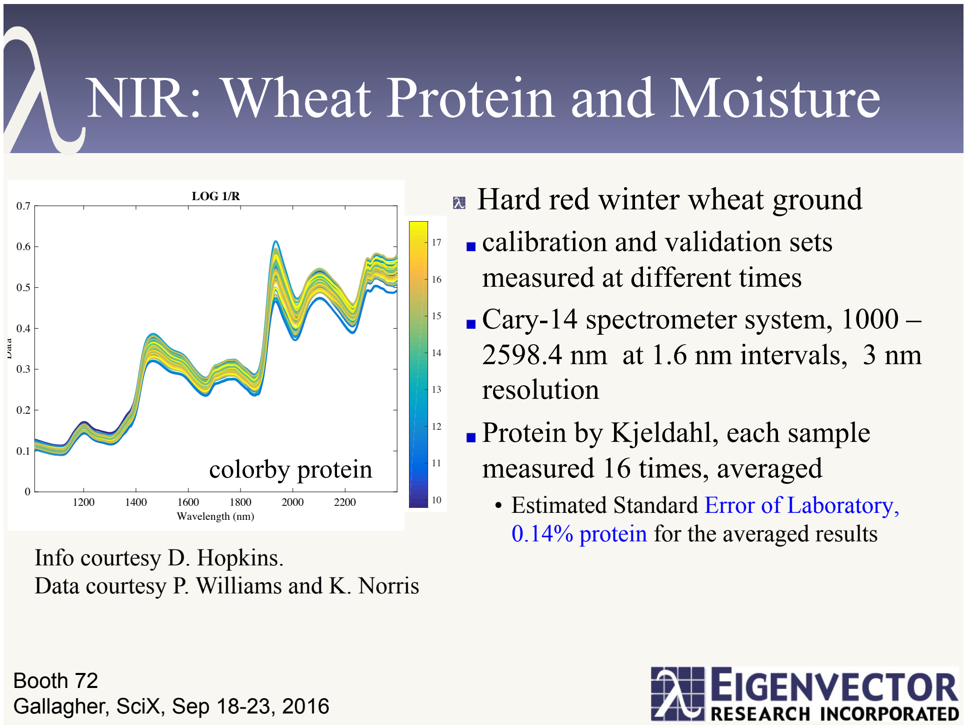


Why is this Important

- Forward least squares methods like CLS and ELS can be more difficult to identify
  - Infrastructure is less well developed
  - Can include more of what we know via constraints during model identification *and* application
  - Interpretability is as good as it gets (**S** *are* spectra!)
    - allows iterative model identification because the identification process teaches about the problem

$$\mathbf{X} = \begin{bmatrix} \mathbf{Y} & \mathbf{T} \end{bmatrix}\begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix}^T$$

Booth 72
Gallagher, SciX, Sep 18-23, 2016

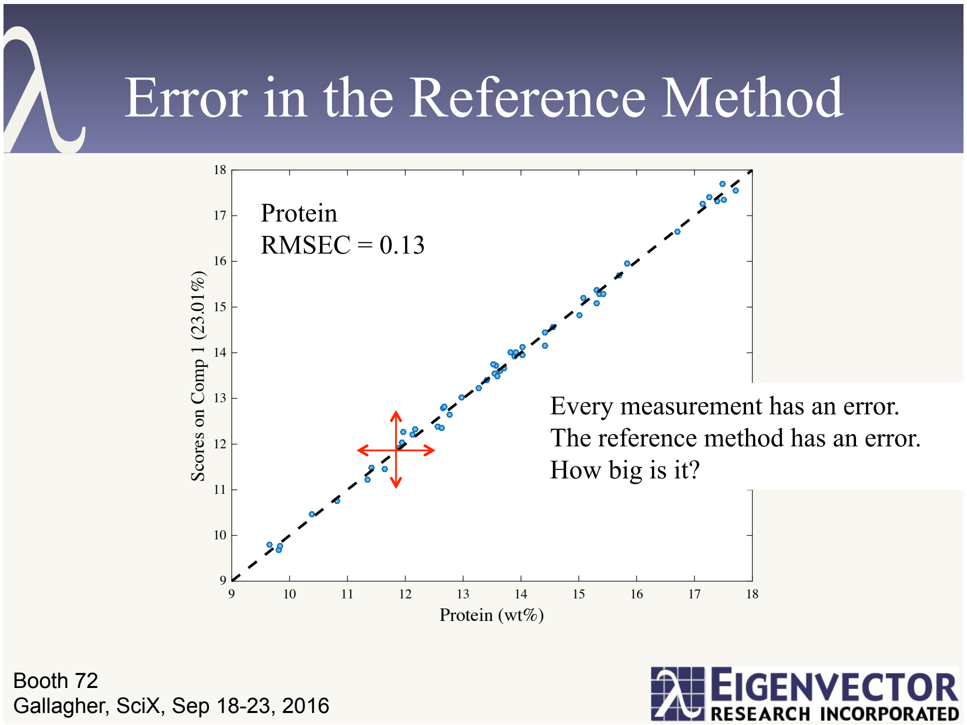


NIR: Wheat Protein and Moisture

- Hard red winter wheat ground
  - calibration and validation sets measured at different times
  - Cary-14 spectrometer system, 1000 – 2598.4 nm at 1.6 nm intervals, 3 nm resolution
  - Protein by Kjeldahl, each sample measured 16 times, averaged
    - Estimated Standard Error of Laboratory, 0.14% protein for the averaged results

Info courtesy D. Hopkins.
Data courtesy P. Williams and K. Norris

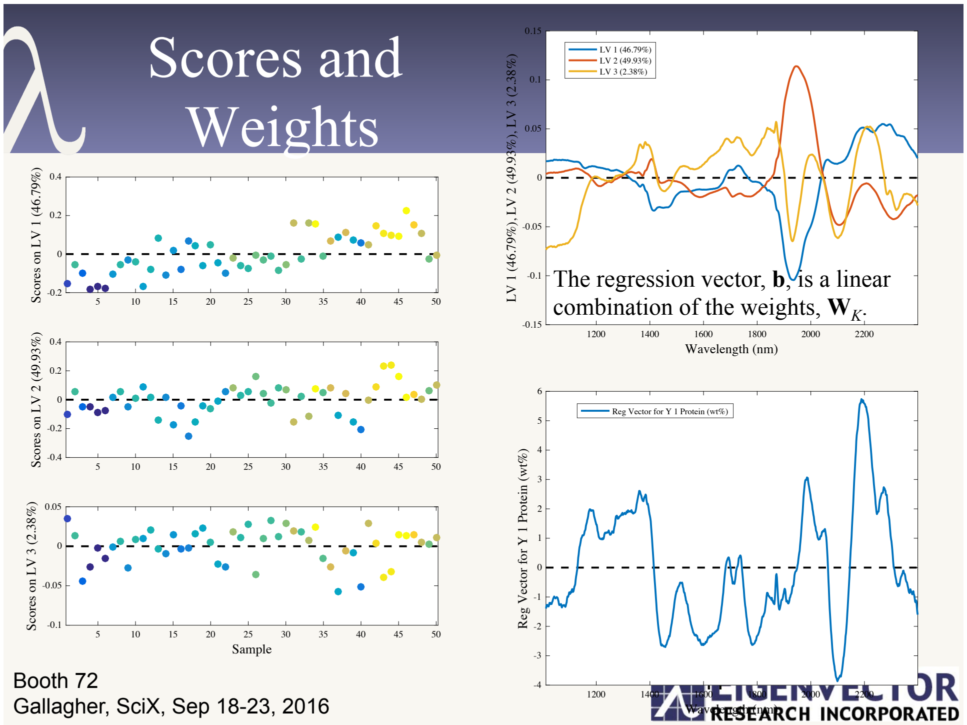
Booth 72
Gallagher, SciX, Sep 18-23, 2016



Error in the Reference Method

Every measurement has an error.
The reference method has an error.
How big is it?

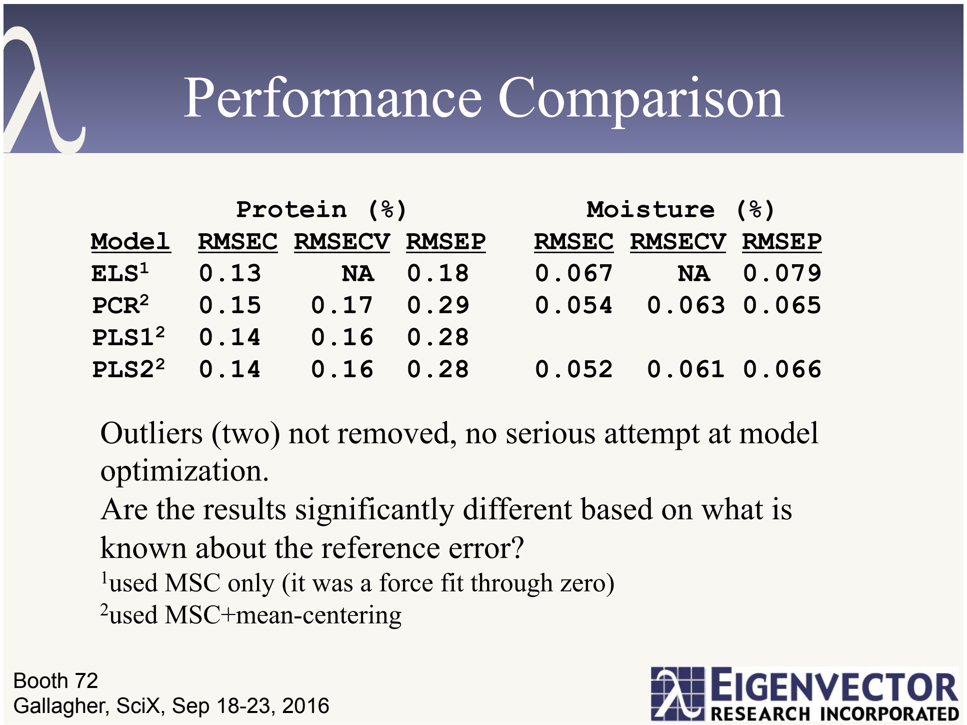
Booth 72
Gallagher, SciX, Sep 18-23, 2016



Scores and Weights

The regression vector, **b**, is a linear combination of the weights, $\mathbf{W}_K$.

Booth 72
Gallagher, SciX, Sep 18-23, 2016



Performance Comparison

	Protein (%)			Moisture (%)		
Model	**RMSEC**	**RMSECV**	**RMSEP**	**RMSEC**	**RMSECV**	**RMSEP**
ELS[1]	0.13	NA	0.18	0.067	NA	0.079
PCR[2]	0.15	0.17	0.29	0.054	0.063	0.065
PLS1[2]	0.14	0.16	0.28			
PLS2[2]	0.14	0.16	0.28	0.052	0.061	0.066

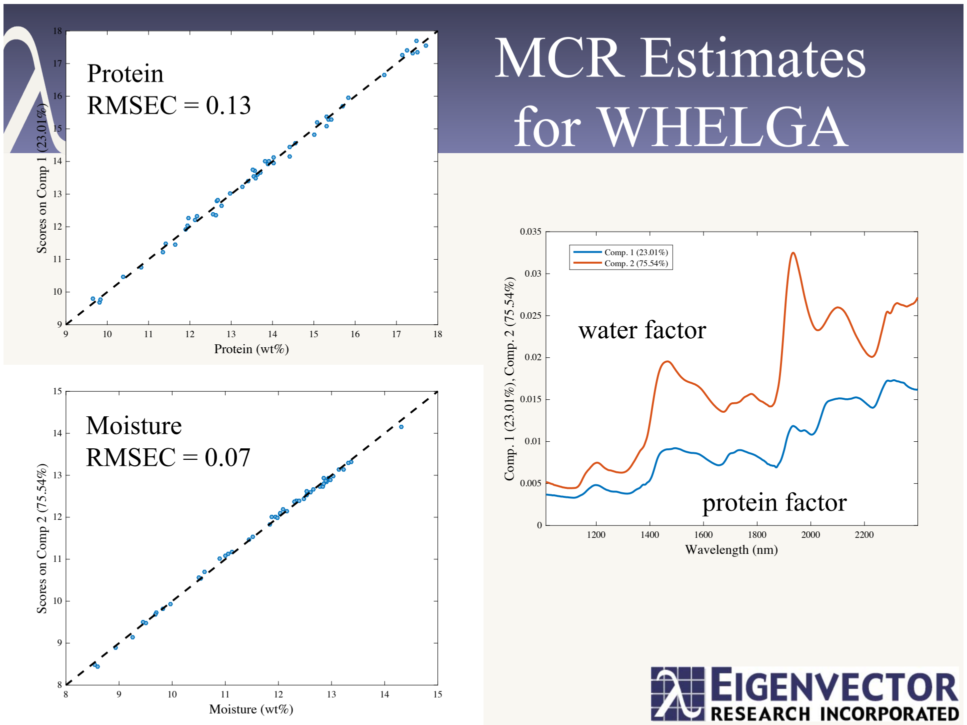
Outliers (two) not removed, no serious attempt at model optimization.

Are the results significantly different based on what is known about the reference error?

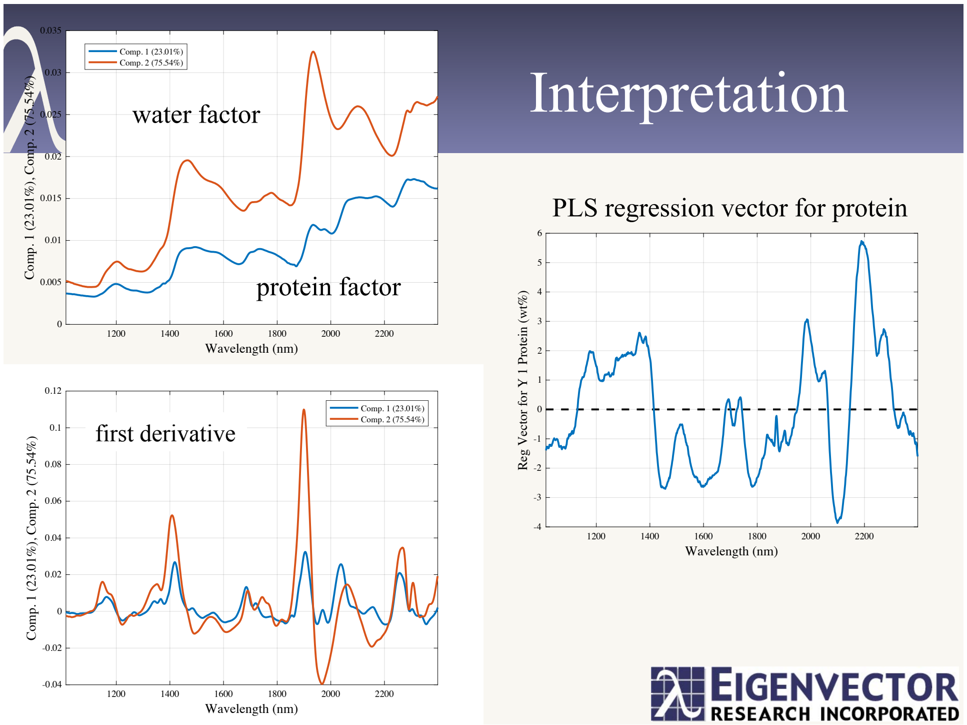
[1] used MSC only (it was a force fit through zero)

[2] used MSC+mean-centering

Booth 72
Gallagher, SciX, Sep 18-23, 2016

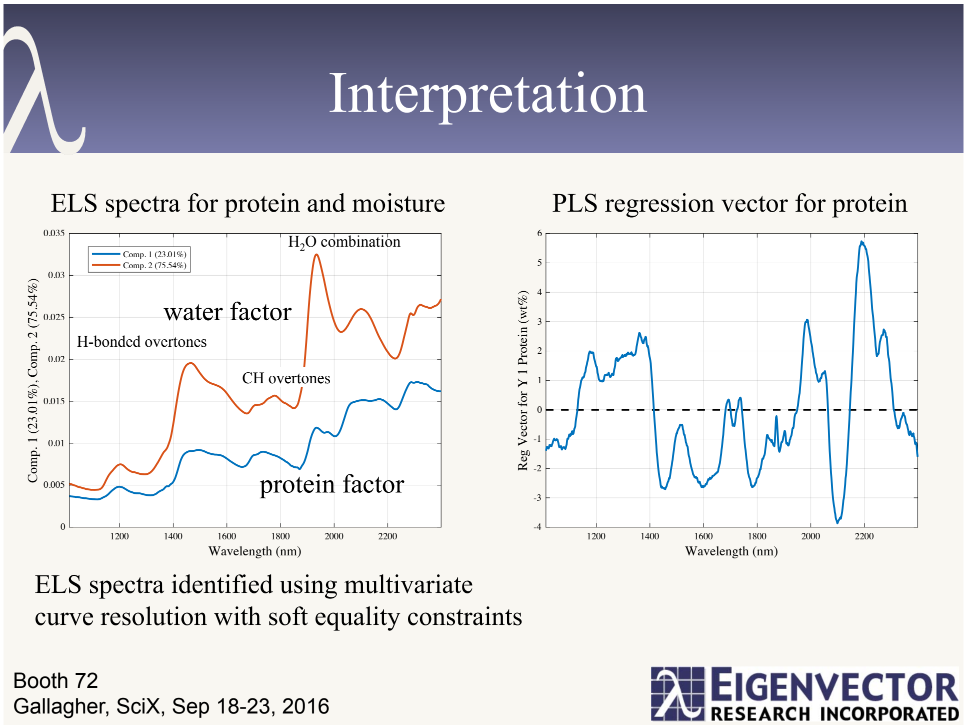


MCR Estimates for WHELGA



Interpretation

PLS regression vector for protein



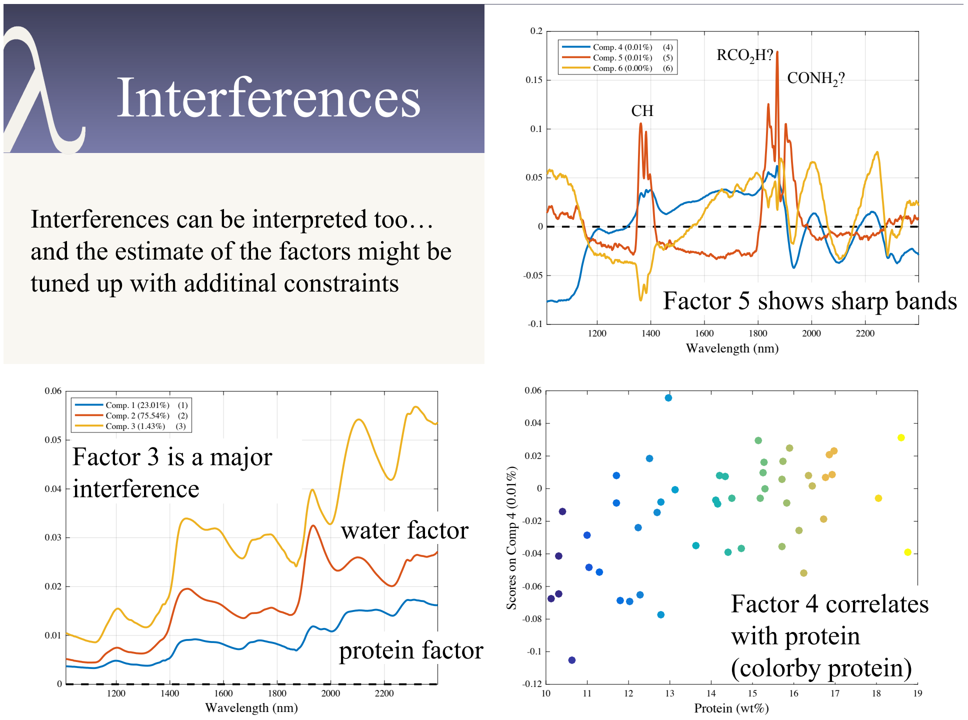
Interpretation

ELS spectra for protein and moisture

PLS regression vector for protein

ELS spectra identified using multivariate curve resolution with soft equality constraints

Booth 72
Gallagher, SciX, Sep 18-23, 2016



Interferences

Interferences can be interpreted too… and the estimate of the factors might be tuned up with additinal constraints

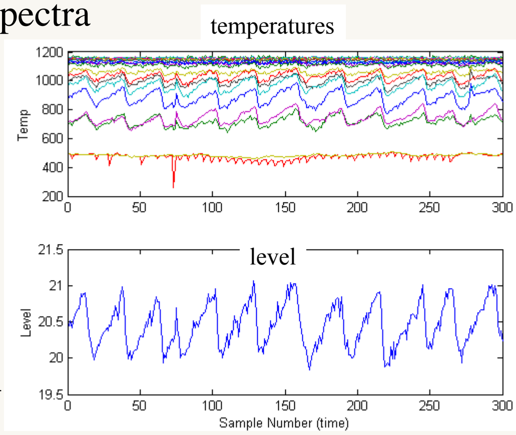
Factor 5 shows sharp bands

Factor 3 is a major interference

Factor 4 correlates with protein (colorby protein)

SFCM Data Example

- Estimate level in a slurry fed ceramic melter[*]
  - measurements are not spectra
  - measured 20 temperatures (thermocouples) in two vertical thermal wells
  - thermocouples near the surface vary with level

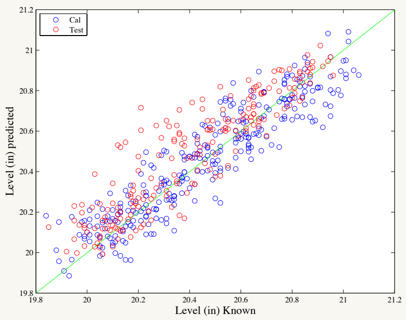


[*] Wise BM, Gallgher NB, Bro R, Shaver JM, Windig W, Koch RS, PLS_Toolbox 3.5, for Use with MATLAB™, Eigenvector Research, Inc. 2004.

Booth 72
Gallagher, SciX, Sep 18-23, 2016

Performance Comparison

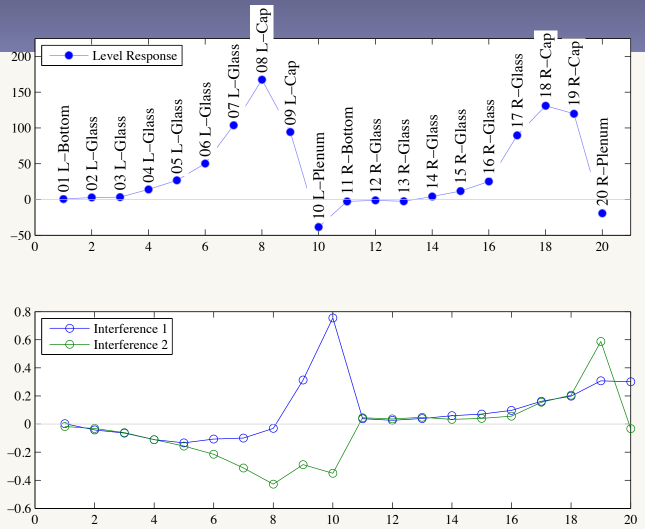
ELS Calibration and Prediction



Demonstration of ELS/CLS for engineering variables.
Mean-Centering for both PLS and ELS
3 factors for PLS and ELS

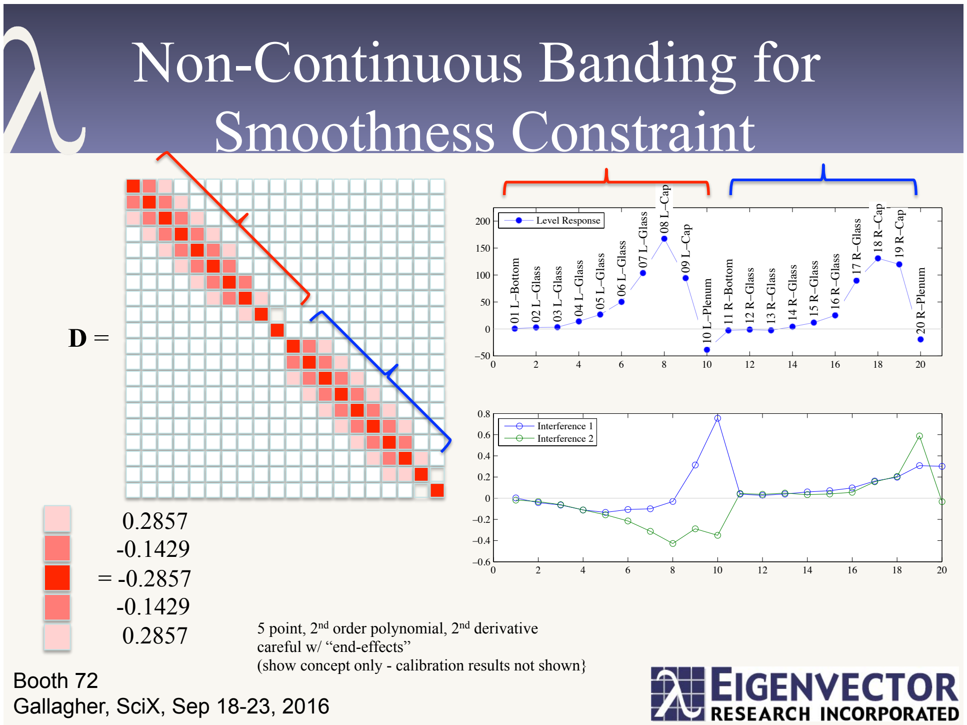
	Level (in)	
	PLS	ELS
RMSEC	0.106	0.114
RMSECV	0.113	0.118
RMSEP	0.138	0.145

Booth 72
Gallagher, SciX, Sep 18-23, 2016



Constraints

- Constraints can be employed on both **C** and **S** during ELS model identification
  - e.g., non-negativity, smoothness, priors, time-series lagging, etc…
- … and on **C** during model application
- Allows imposing chemical and physical knowledge into the model identification



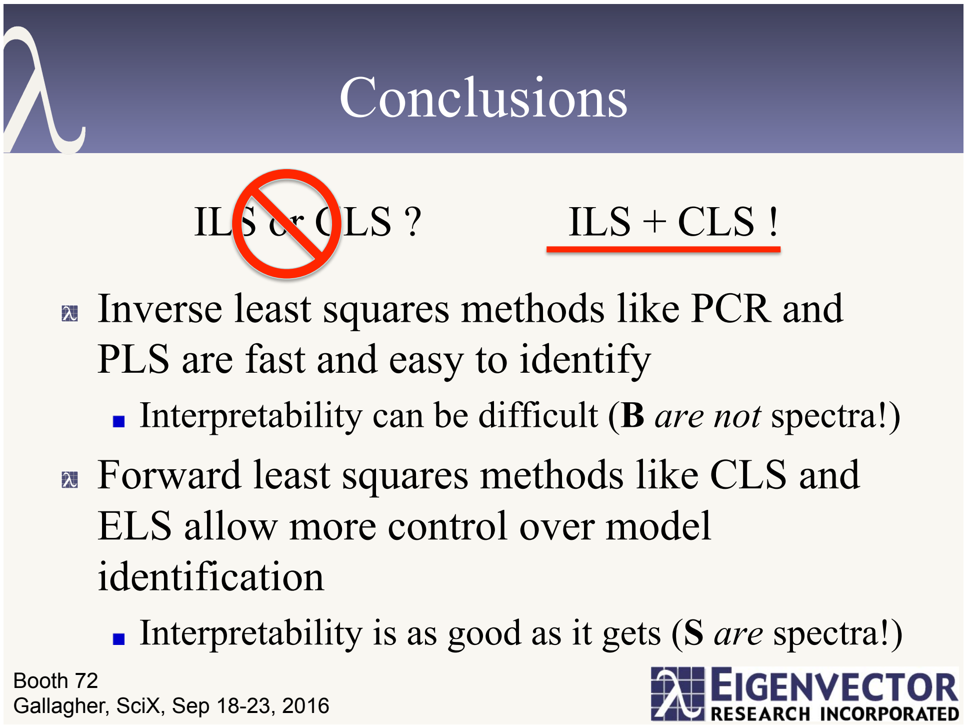
Non-Continuous Banding for Smoothness Constraint

D =

🟥	0.2857
🟥	-0.1429
🟥	= -0.2857
🟥	-0.1429
🟥	0.2857

5 point, 2nd order polynomial, 2nd derivative
careful w/ "end-effects"
(show concept only - calibration results not shown}

Booth 72
Gallagher, SciX, Sep 18-23, 2016



Conclusions

ILS or CLS ? ILS + CLS !

- Inverse least squares methods like PCR and PLS are fast and easy to identify
  - Interpretability can be difficult (**B** *are not* spectra!)
- Forward least squares methods like CLS and ELS allow more control over model identification
  - Interpretability is as good as it gets (**S** *are* spectra!)

Booth 72
Gallagher, SciX, Sep 18-23, 2016