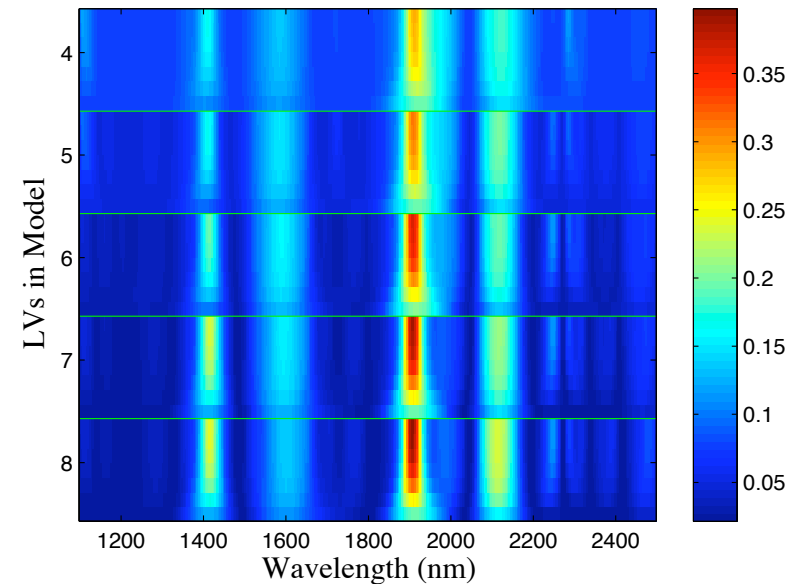


Prediction Error for Corn Moisture with 4–8 LVs



Tools for Multivariate Calibration Robustness Testing with Observations on Effects of Data Preprocessing

Barry M. Wise



Introduction

- When developing calibration models focus is generally on improving prediction error
- Models often developed with small amount of data taken over relatively short time
- Prediction errors over long term often dominated by artifacts not represented in calibration data
 - Spectrometer
 - Sample

What Constitutes a Good Model?

- Acceptable prediction error
 - Note: *not* best achievable
- Longevity, *i.e.* robustness to minor changes

Possible Changes to System

- Sample
 - New analyte(s)
 - Changes in physical properties (e.g. scattering)
 - Temperature
 - Pressure
- Instrument
 - Wavelength registration shift
 - Stray light
 - Resolution
 - Noise

Robustness Tests

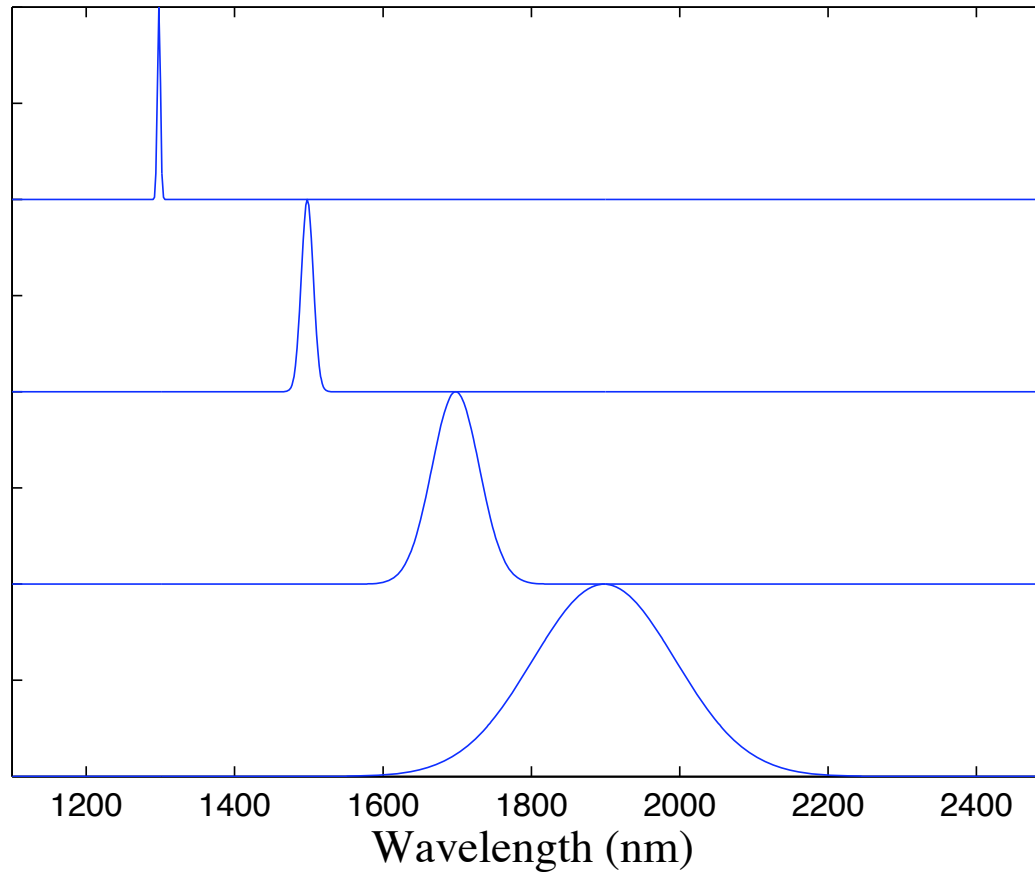
- Series of functions developed to test model against system changes
 - Develop model with desired preprocessing, #LVs, etc.
 - “Perturb” test data set
 - Apply calibration model to “perturbed” data
 - Look at prediction error as function of perturbations
 - Test and compare multiple models

Perturbations

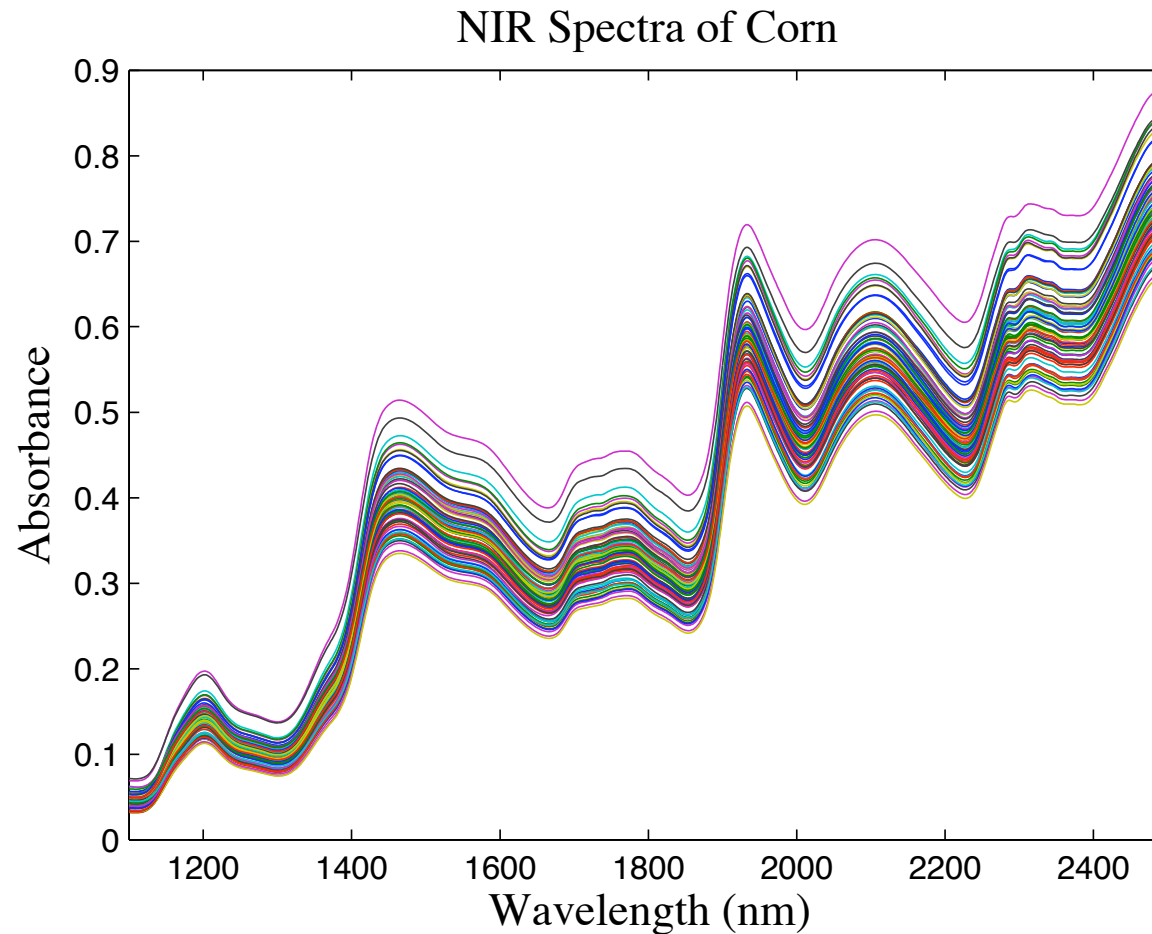
- New analyte – add Gaussian peak of variable width across wavelength range
- Wavelength registration shift – shift spectra left-right as well as expand and contract
- Baseline shift – change offset and slope
- Stray light – add fraction of signal before log transform
- Temperature – decrease resolution and vary path length
- Noise variation – add noise with varying bandwidth

New Analyte

Example Peak Shapes for Testing Robustness to New Analytes

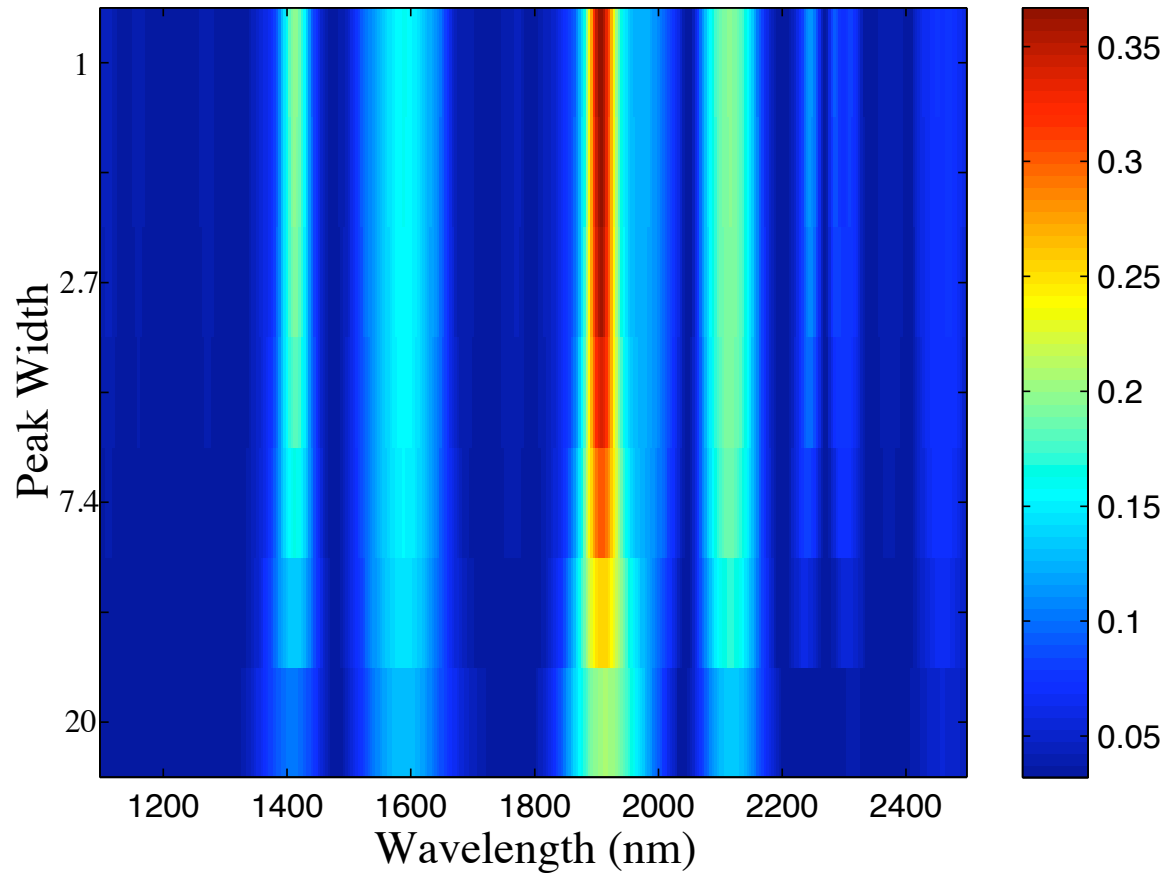


Example Data: Corn NIR



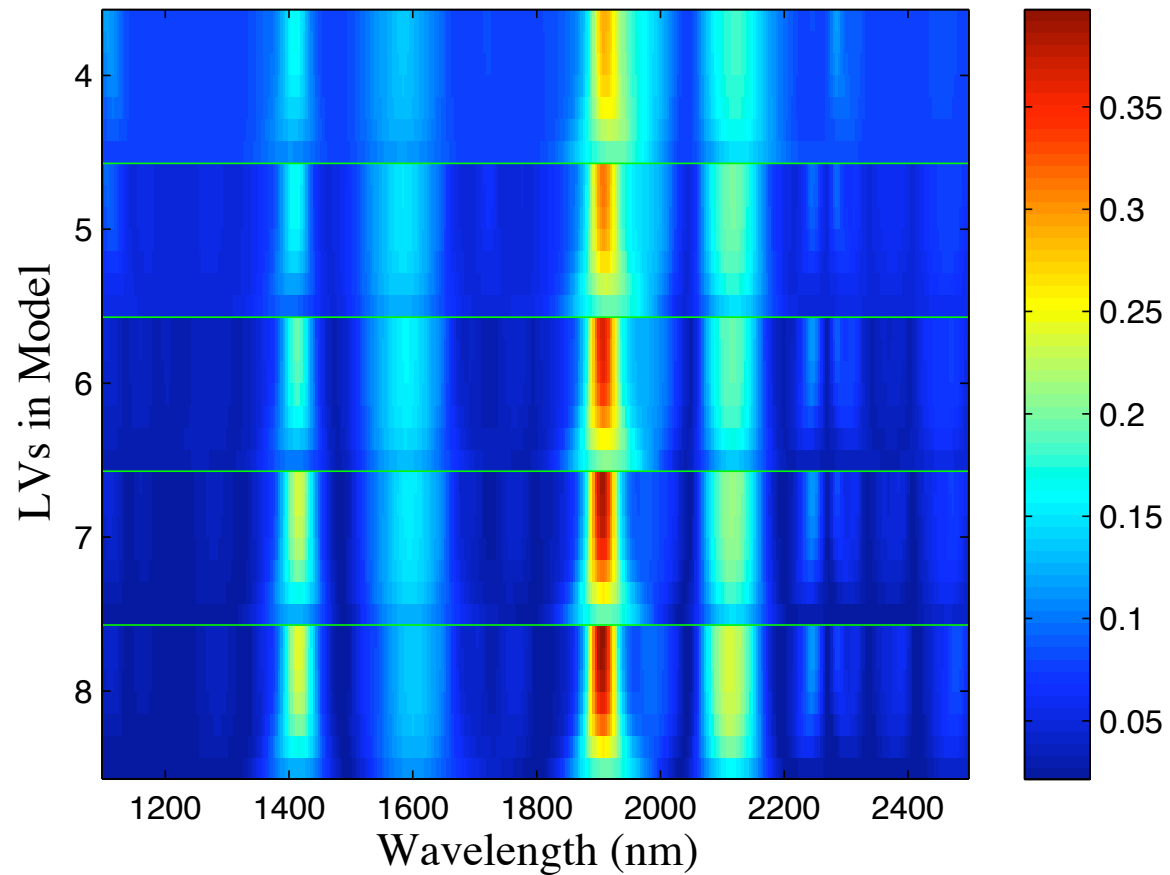
Test Corn Model

Prediction Error for Corn Moisture with 6 LVs



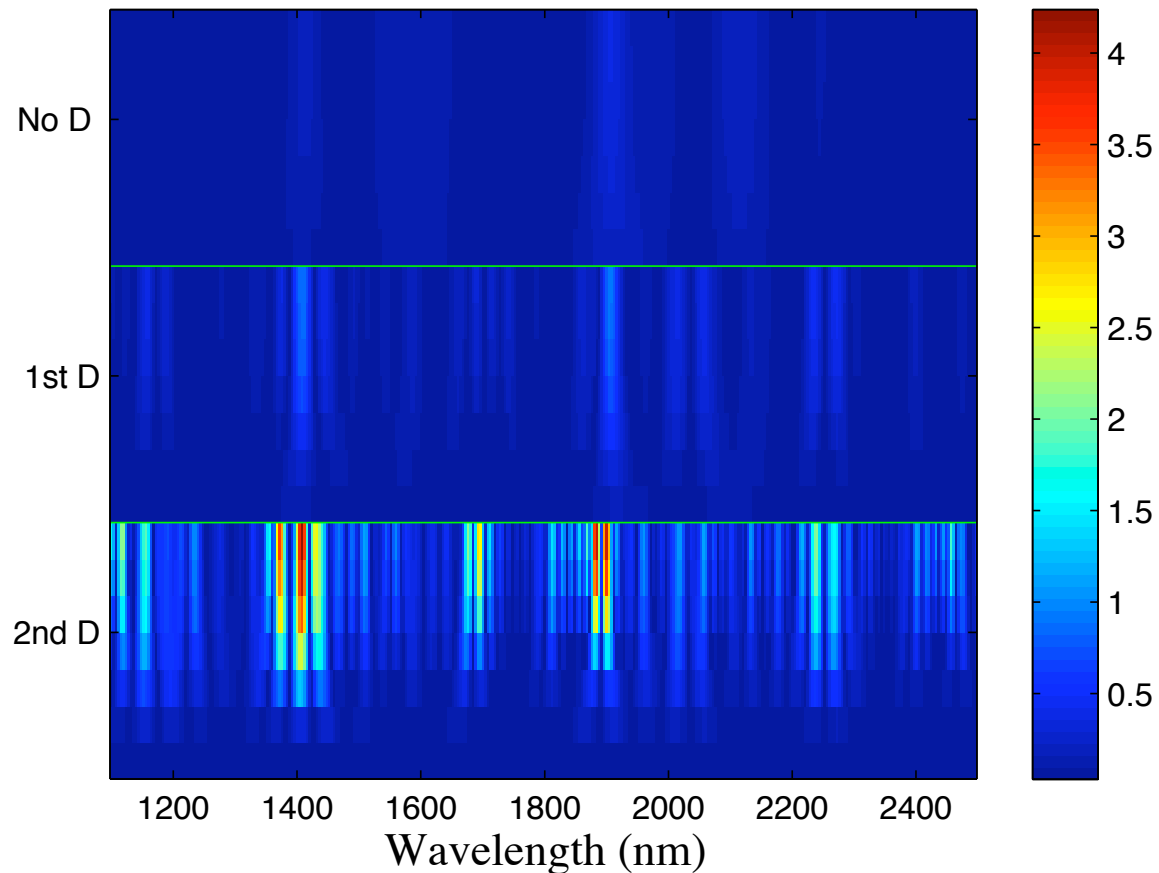
Compare Models- #LVs

Prediction Error for Corn Moisture with 4–8 LVs

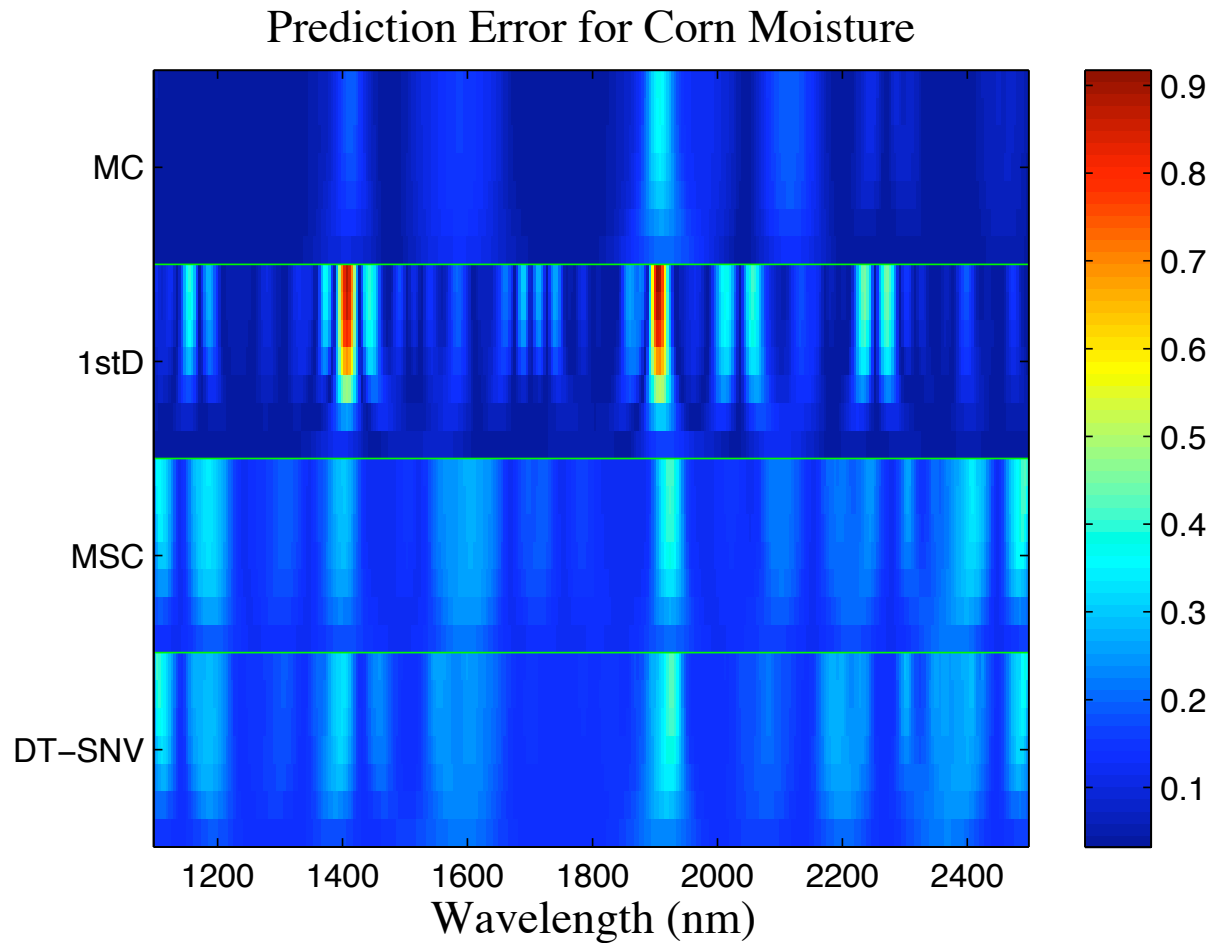


Compare Models-Derivative

Prediction Error for Corn Moisture with no, 1st and 2nd Derivative

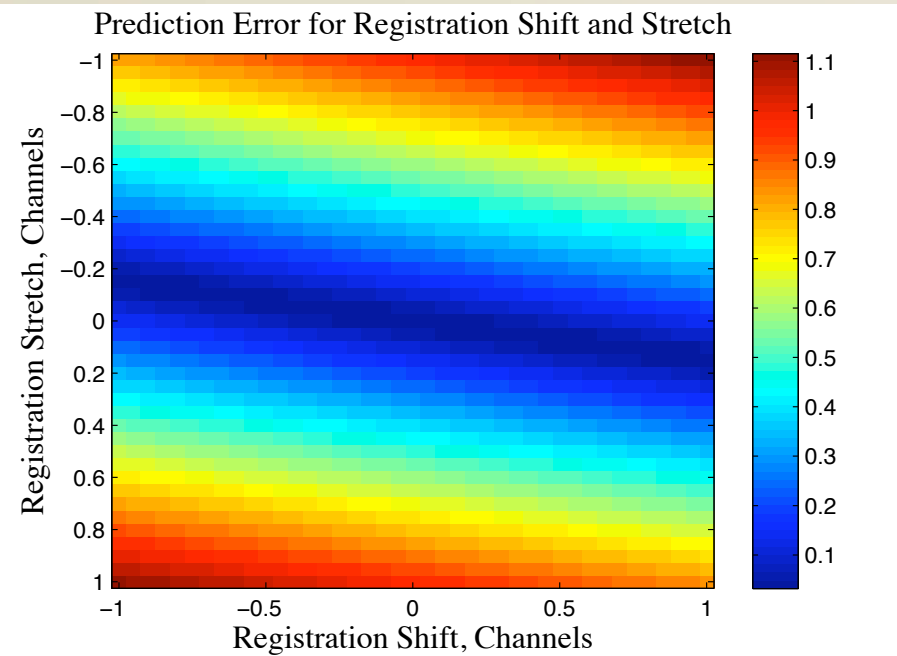


Other Preprocessing



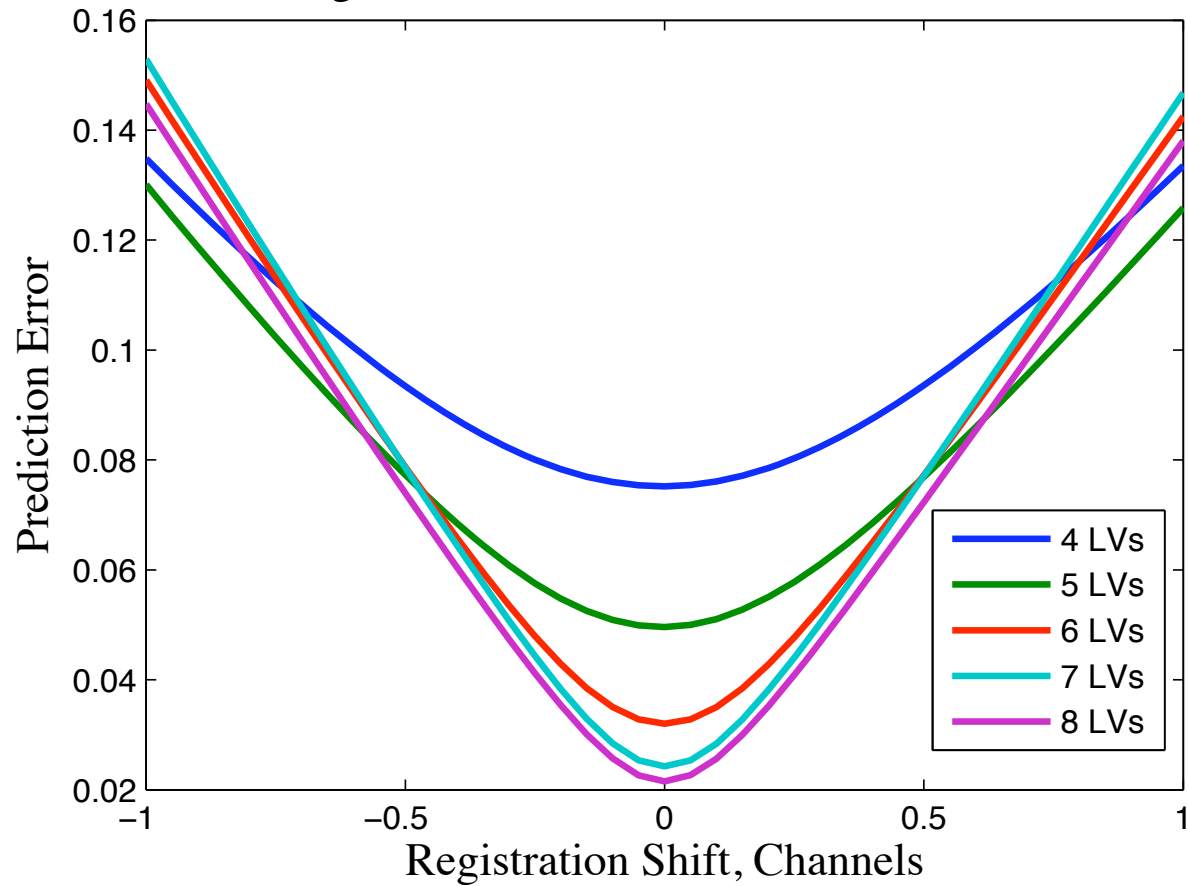
Registration Shift

- Registration shift
 - Left-right
 - Expand and Contract

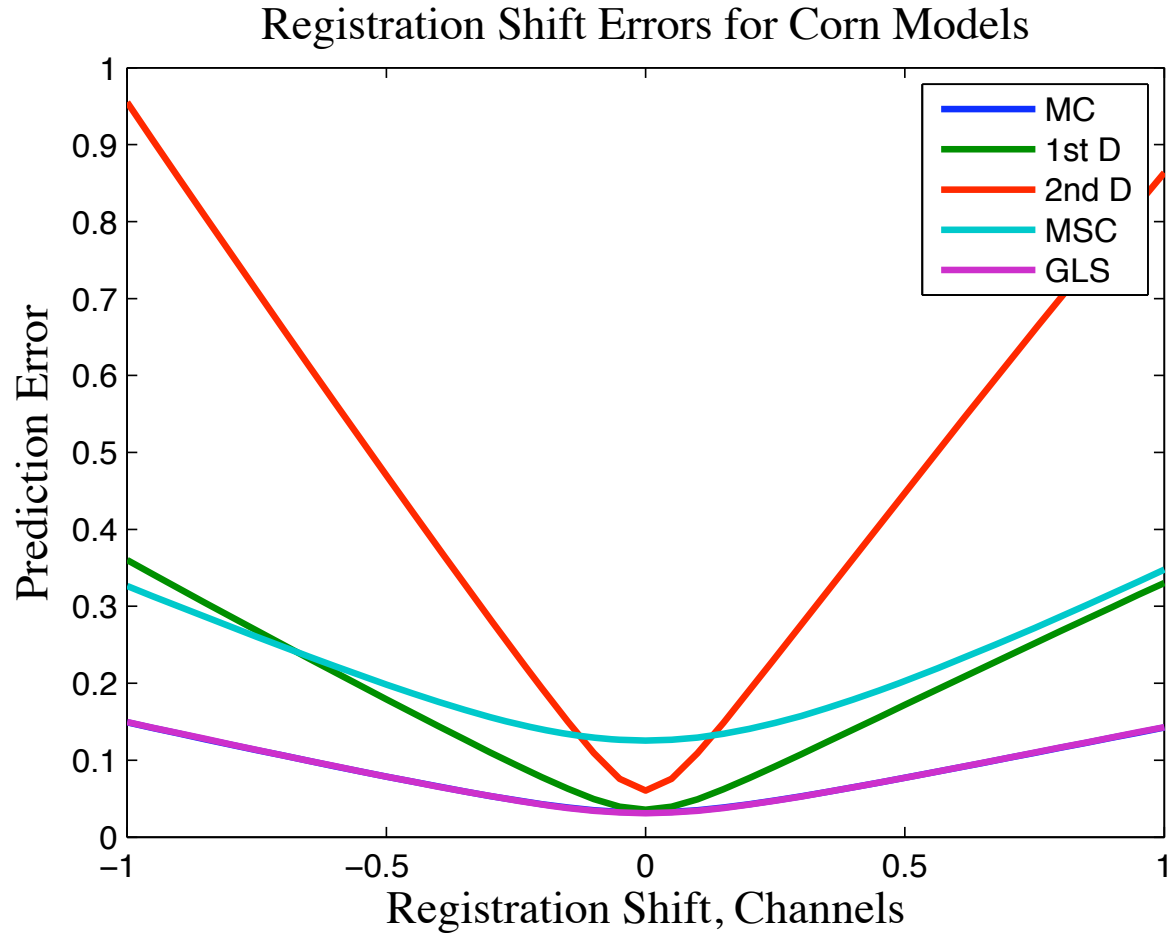


Shift with #LVs

Registration Shift Errors for Corn Models



Shift with Preprocessing



Implementation

The screenshot displays the Eigenvector software interface. The main window title is "Analysis - PLS 6 LVs - m5spec, propvals". The menu bar includes File, Edit, Preprocess, Analysis, Tools, Help, and FigBrowser. The interface is divided into several sections:

- Data:** modeled (calibration set), Var: m5spec, propvals, Size: 80 x 700, 80 x 4, Samp Lbls: , Var Lbls: .
- Model:** calibrated on loaded data, Type: PLS (6 LVs), RMSEC: 0.035222, X Preprocessing: Smoothing (order: 0, wind:), Y Preprocessing: Autoscale, Data: 80 x 700, 80 x 1.
- View:** SSQ Table (selected), IPLS Variable Selection.
- Number LVs:** 6 (selected), Auto Select button.
- Percent Variance Captured by Model:** A table with columns for Latent Variable, X-Block (LV, Cum), and Y-Block (LV, Cum). Row 6 is highlighted in green.
- Eigenvector Cache by DATE (* - Not Available):** A list of models with dates. The date 2008-06-28 is selected, and a red arrow points from the highlighted row 6 in the variance table to the corresponding entry in the cache.
- Warning:** This model appears to have some unusual Hotelling's T² values. Please review T² and T contributions using the Scores plot and determine if these samples are errors that should be removed. If these are not errors, consider adding additional samples which are like these.
- Footer:** A model has been calibrated from the data. Review the model using the toolbar button(s), save

Latent Variable	X-Block		Y-Block	
	LV	Cum	LV	Cum
1	99.10	99.10	39.03	39.03
2	0.75	99.85	19.15	58.18
3	0.06	99.91	22.82	81.00
4	0.03	99.94	14.53	95.53
5	0.03	99.97	2.40	97.92
6	0.01	99.98	1.21	99.13
7	0.01	99.99	0.37	99.51
8	0.01	99.99	0.09	99.60
9	0.00	100.00	0.19	99.78
10	0.00	100.00	0.03	99.81
11	0.00	100.00	0.05	99.87

Model cache

Conclusions

- Functions developed to test model robustness in face of new data “non-idealities”
- Makes it easy to compare models
 - Across #LVs
 - Preprocessing
 - Variable selection
 - Other model types? (Functional?)
- Models more brittle with LVs (knew that)
- Some preprocessing techniques more robust (*e.g.* GLS) than others (*e.g.* 2nd derivative)

References

- S.C. Rutan, O.E. de Noord and R.R. Andréa, “Characterization of the Sources of Variation Affecting Near-Infrared Spectroscopy Using Chemometric Methods,” *Anal. Chem.* Vol. 70, 3198-3201, 1998.
- F. Estienne, F. Despagne, B. Walczak, O.E. de Noord, and D.L. Massart, “A comparison of multivariate calibration techniques applied to experimental NIR data sets Part III: Robustness against instrumental perturbation conditions,” *Chemo. Intell. Lab. Sys.* Vol. 73, 207-218, 2004.

Questions?

