# Optimizing Models: Preprocessing and Variable Selection

EIGENVECTOR RESEARCH INCORPORATED

## What is Preprocessing?

- Preprocessing includes any procedure done to transform data *before* it is modeled (with PCA, PLS, MCR, ANNs, etc.)
- Preprocessing is often the critical difference between adequate and inadequate models
- There are many ways to preprocess data, many of them particular to spectroscopy and spectrometry



EIGENVECTOR RESEARCH INCORPORATED

# Goals of Preprocessing

- The goal of preprocessing is to remove/reduce variation you don't care about and emphasize variation you do care about
  - Don't care about variance that is not related to the problem you are trying to solve
  - Do care about variation related to the thing you are looking for
- This lets the analysis focus on the variation due to the problem of interest

71

EIGENVECTOR RESEARCH INCORPORATED

# Sources of Variation

- Variation *not* related to the problem of interest
  - Measurement noise
  - Baseline effects
  - Non-linearities (detector, sample matrix effects, etc.)
  - Other analytes
  - Physical effects (scatter, sample temperature or pressure)
  - Intra-class variation (and sometimes inter-class variation)

Clutter

- Variation of interest
  - Signal from analytes or property of interest
  - Intra-class variation

72

EIGENVECTOR RESEARCH INCORPORATED

# Background/Baseline Subtraction

Removal of broad (low-frequency) interferences while retaining higher-frequency features.

- **Detrend:** fit polynomial to *entire* spectrum
- **Selected-Points baselining:** fit polynomial to selected points in spectrum
- **Weighted Least-squares (a.k.a. asymmetric) baselining:** fit to *automatically* selected points on the bottom of the spectrum
- **Windowed:** Whittaker, Rolling Ball, Median, Minimum, etc.

73

EIGENVECTOR RESEARCH INCORPORATED

# Sample Normalization Methods

How about variance due to changing magnitude?

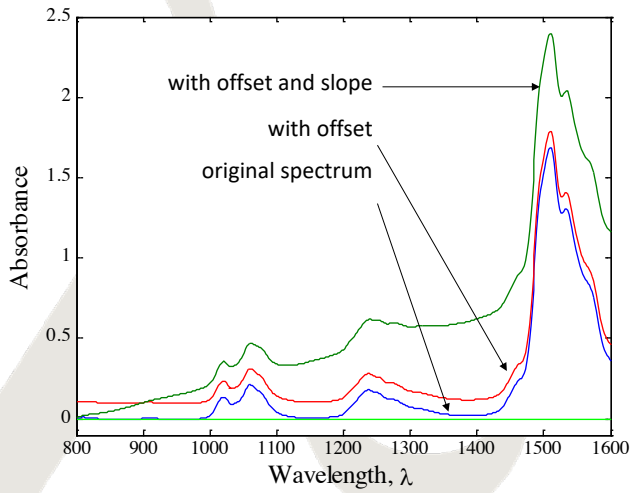
- variable source or lighting magnitude
- scattering effects and path length effects

- **Row Normalization**: removes magnitude
- **Standard Normal Variate (SNV)**: subtracts the row mean from each row and scales to unit variance
- **Multiplicative Scatter Correction:** Determines scale factor that best fits new spectrum to reference
- Be aware that these can "blow up" noisy samples to have more variance

74

EIGENVECTOR RESEARCH INCORPORATED

# Smoothing and Derivatives




- Derivatives with respect to λ can be used to remove offsets/slopes
- Savitzky-Golay
  - piece-wise fit polynomial to each spectrum
  - smooth + derivative can be boiled down to a set of coefficients
- First derivative will remove offset
- Second derivative will remove offset and slope



EIGENVECTOR RESEARCH INCORPORATED

# Orthogonalization Filters

- Develop a description of clutter, *i.e.* variation that you would like to eliminate
  - Variation in samples *not* related to analyte of interest
- Use description to develop filter to remove clutter
- External Parameter Orthogonalization (EPO)
  - Use PCA to develop variation basis
  - Filtered data is residuals on this PCA model
- Generalized Least Squares Weighting (GLSW)
  - Inverse square root of clutter covariance



EIGENVECTOR RESEARCH INCORPORATED

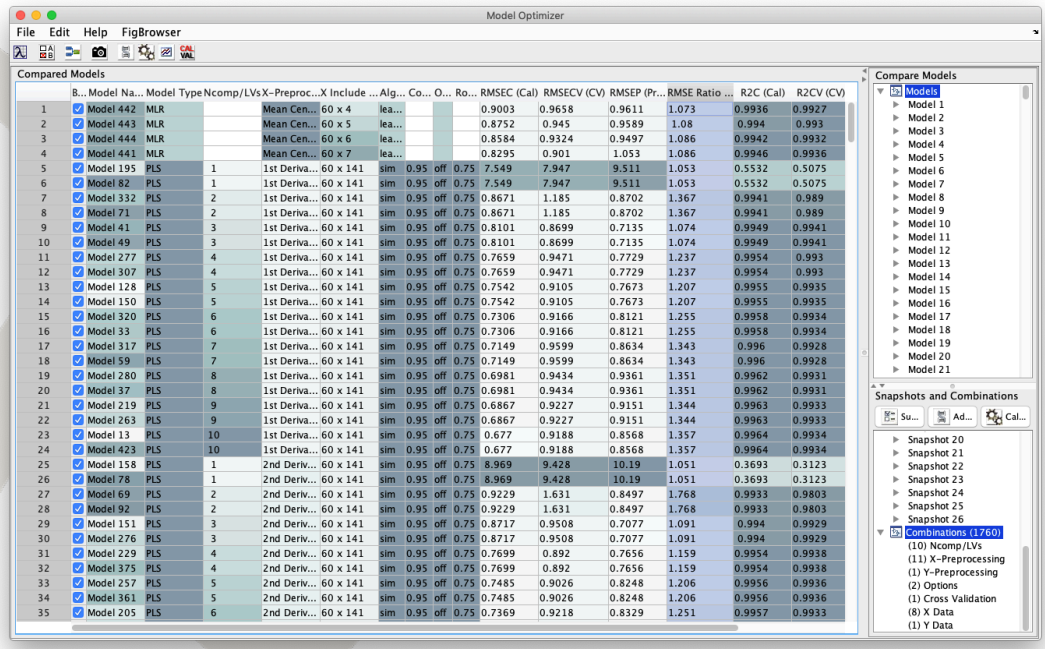
# Variable Selection

- Irrelevant and noisy variables degrade a model's predictive performance
- Many methods available to select "best" variables
  - Step-wise (forward or backward)
  - interval PLS (i-PLS)
  - Variable Influence on Prediction (VIP)
  - Selectivity Ratio (SR)
  - Genetic Algorithm
  - Many many others

77

EIGENVECTOR RESEARCH INCORPORATED

# Model Optimizer




| | B... | Model Na... | Model Type | Ncomp/LVs | X-Preproc... | X Include ... | Alg... | Co... | O... | Ro... | RMSEC (Cal) | RMSECV (CV) | RMSEP (Pr... | RMSE Ratio ... | R2C (Cal) | R2CV (CV) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ☑ | Model 442 | MLR | | Mean Cen... | 60 x 4 | lea... | | | | 0.9003 | 0.9658 | 0.9611 | 1.073 | 0.9936 | 0.9927 |
| 2 | ☑ | Model 443 | MLR | | Mean Cen... | 60 x 5 | lea... | | | | 0.8752 | 0.945 | 0.9589 | 1.08 | 0.994 | 0.993 |
| 3 | ☑ | Model 444 | MLR | | Mean Cen... | 60 x 6 | lea... | | | | 0.8584 | 0.9324 | 0.9497 | 1.086 | 0.9942 | 0.9932 |
| 4 | ☑ | Model 441 | MLR | | Mean Cen... | 60 x 7 | lea... | | | | 0.8295 | 0.901 | 1.053 | 1.086 | 0.9946 | 0.9936 |
| 5 | ☑ | Model 195 | PLS | 1 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 7.549 | 7.947 | 9.511 | 1.053 | 0.5532 | 0.5075 |
| 6 | ☑ | Model 82 | PLS | 1 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 7.549 | 7.947 | 9.511 | 1.053 | 0.5532 | 0.5075 |
| 7 | ☑ | Model 332 | PLS | 2 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.8671 | 1.185 | 0.8702 | 1.367 | 0.9941 | 0.989 |
| 8 | ☑ | Model 71 | PLS | 2 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.8671 | 1.185 | 0.8702 | 1.367 | 0.9941 | 0.989 |
| 9 | ☑ | Model 41 | PLS | 3 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.8101 | 0.8699 | 0.7135 | 1.074 | 0.9949 | 0.9941 |
| 10 | ☑ | Model 49 | PLS | 3 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.8101 | 0.8699 | 0.7135 | 1.074 | 0.9949 | 0.9941 |
| 11 | ☑ | Model 277 | PLS | 4 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7659 | 0.9471 | 0.7729 | 1.237 | 0.9954 | 0.993 |
| 12 | ☑ | Model 307 | PLS | 4 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7659 | 0.9471 | 0.7729 | 1.237 | 0.9954 | 0.993 |
| 13 | ☑ | Model 128 | PLS | 5 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7542 | 0.9105 | 0.7673 | 1.207 | 0.9955 | 0.9935 |
| 14 | ☑ | Model 150 | PLS | 5 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7542 | 0.9105 | 0.7673 | 1.207 | 0.9955 | 0.9935 |
| 15 | ☑ | Model 320 | PLS | 6 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7306 | 0.9166 | 0.8121 | 1.255 | 0.9958 | 0.9934 |
| 16 | ☑ | Model 33 | PLS | 6 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7306 | 0.9166 | 0.8121 | 1.255 | 0.9958 | 0.9934 |
| 17 | ☑ | Model 317 | PLS | 7 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7149 | 0.9599 | 0.8634 | 1.343 | 0.996 | 0.9928 |
| 18 | ☑ | Model 59 | PLS | 7 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7149 | 0.9599 | 0.8634 | 1.343 | 0.996 | 0.9928 |
| 19 | ☑ | Model 280 | PLS | 8 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.6981 | 0.9434 | 0.9361 | 1.351 | 0.9962 | 0.9931 |
| 20 | ☑ | Model 37 | PLS | 8 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.6981 | 0.9434 | 0.9361 | 1.351 | 0.9962 | 0.9931 |
| 21 | ☑ | Model 219 | PLS | 9 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.6867 | 0.9227 | 0.9151 | 1.344 | 0.9963 | 0.9933 |
| 22 | ☑ | Model 263 | PLS | 9 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.6867 | 0.9227 | 0.9151 | 1.344 | 0.9963 | 0.9933 |
| 23 | ☑ | Model 13 | PLS | 10 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.677 | 0.9188 | 0.8568 | 1.357 | 0.9964 | 0.9934 |
| 24 | ☑ | Model 423 | PLS | 10 | 1st Deriva... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.677 | 0.9188 | 0.8568 | 1.357 | 0.9964 | 0.9934 |
| 25 | ☑ | Model 158 | PLS | 1 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 8.969 | 9.428 | 10.19 | 1.051 | 0.3693 | 0.3123 |
| 26 | ☑ | Model 78 | PLS | 1 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 8.969 | 9.428 | 10.19 | 1.051 | 0.3693 | 0.3123 |
| 27 | ☑ | Model 69 | PLS | 2 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.9229 | 1.631 | 0.8497 | 1.768 | 0.9933 | 0.9803 |
| 28 | ☑ | Model 92 | PLS | 2 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.9229 | 1.631 | 0.8497 | 1.768 | 0.9933 | 0.9803 |
| 29 | ☑ | Model 151 | PLS | 3 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.8717 | 0.9508 | 0.7077 | 1.091 | 0.994 | 0.9929 |
| 30 | ☑ | Model 276 | PLS | 3 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.8717 | 0.9508 | 0.7077 | 1.091 | 0.994 | 0.9929 |
| 31 | ☑ | Model 229 | PLS | 4 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7699 | 0.892 | 0.7656 | 1.159 | 0.9954 | 0.9938 |
| 32 | ☑ | Model 375 | PLS | 4 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7699 | 0.892 | 0.7656 | 1.159 | 0.9954 | 0.9938 |
| 33 | ☑ | Model 257 | PLS | 5 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7485 | 0.9026 | 0.8248 | 1.206 | 0.9956 | 0.9936 |
| 34 | ☑ | Model 361 | PLS | 5 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7485 | 0.9026 | 0.8248 | 1.206 | 0.9956 | 0.9936 |
| 35 | ☑ | Model 205 | PLS | 6 | 2nd Deriv... | 60 x 141 | sim | 0.95 | off | 0.75 | 0.7369 | 0.9218 | 0.8329 | 1.251 | 0.9957 | 0.9933 |

78

EIGENVECTOR RESEARCH INCORPORATED

# Conclusions

- Models can often be improved by optimizing preprocessing and variable selection
  - Generally easier to make a good model better than to make a bad model good
- Leads to many possible models!

79

EIGENVECTOR RESEARCH INCORPORATED

# Preprocessing and Variable Selection Resources

- Webinars
  - https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-about-preprocessing/
  - https://eigenvector.com/aiovg_videos/evrithing-you-need-to-know-about-the-model-optimizer/
- Final Model Selection conference talk
  - https://vimeo.com/chemometrics/final-model-selection
- Eigenvector University courses live and recorded
  - https://eigenvector.com/events/advanced-preprocessing/
  - https://eigenvector.com/events/variable-selection/

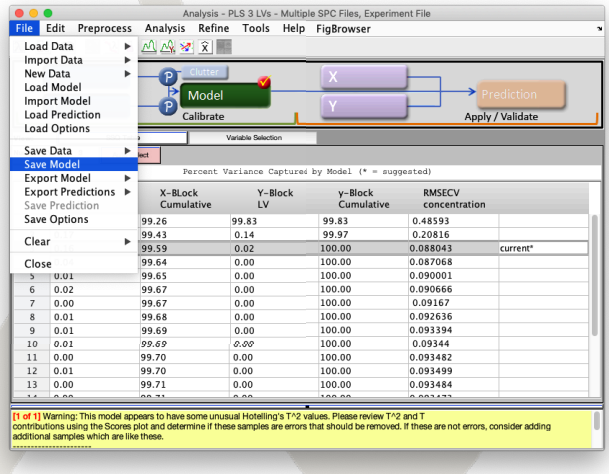
80

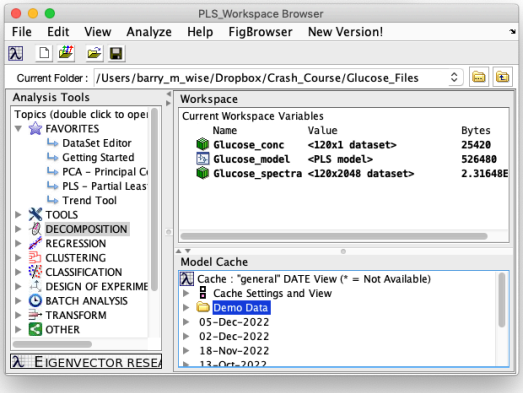
EIGENVECTOR RESEARCH INCORPORATED

# Saving and Documenting Models

EIGENVECTOR RESEARCH INCORPORATED

## Saving a Model

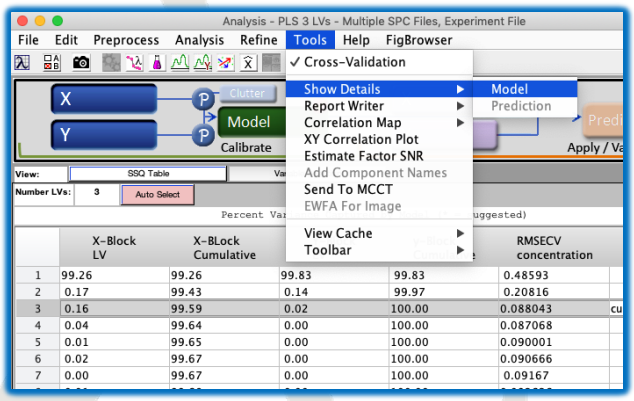


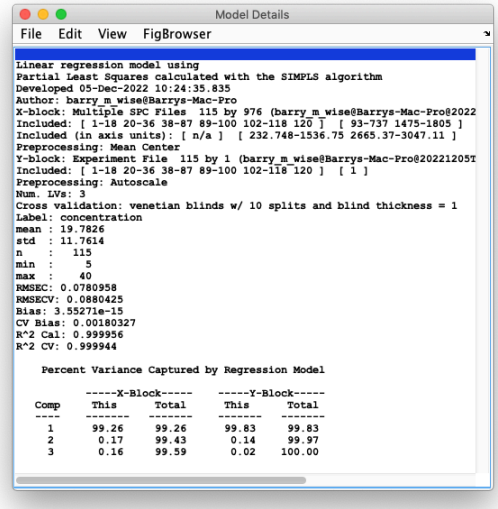




**Save model to hard drive as .mat file!**

EIGENVECTOR RESEARCH INCORPORATED

# Model Details

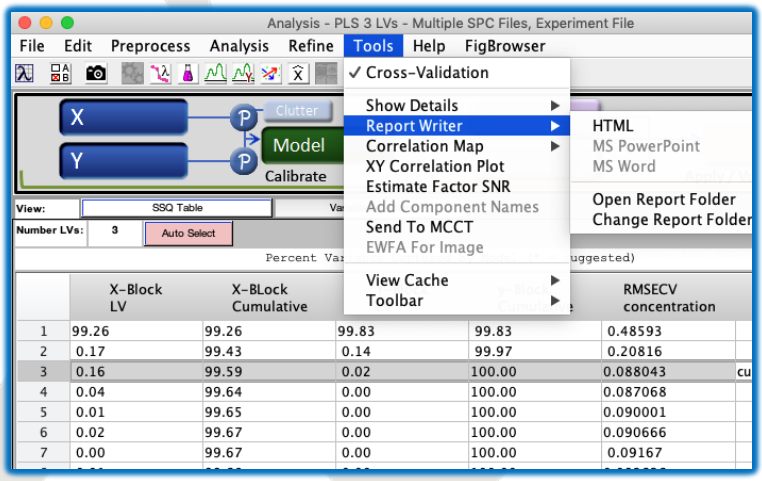



83

EIGENVECTOR RESEARCH INCORPORATED

# Report Writer




- Report Writer creates html, Word or Powerpoint files that include
  - Model Details
  - Copies of all open figure windows associated with the model
  - Autogenerated pre-set plots

84

EIGENVECTOR RESEARCH INCORPORATED

# Saving and Documenting Resources

- From our Wiki
  - https://www.wiki.eigenvector.com/index.php?title=Loading_and_Saving
  - https://www.wiki.eigenvector.com/index.php?title=Reportwriter

85

EIGENVECTOR RESEARCH INCORPORATED

# Model Maintenance

EIGENVECTOR RESEARCH INCORPORATED

## Why Model Maintenance?

- Numerous things can cause calibration models to become invalid and produce poor predictions/classifications
  - samples move to a range outside original calibration
    - analyte or interferent goes beyond calibration range or occurs in unusual combination
  - new variation is introduced into the samples
    - new interferent or variation in physical parameter, *e.g.* temperature
  - a change in the sample matrix causes the relationship between analyte and measurement to change
    - change in pH, particle size, temperature, pressure
  - a change in the hardware causes the analyte-measurement relationship to change
    - instrument maintenance, fiber optic change, source replacement, etc.



EIGENVECTOR RESEARCH INCORPORATED

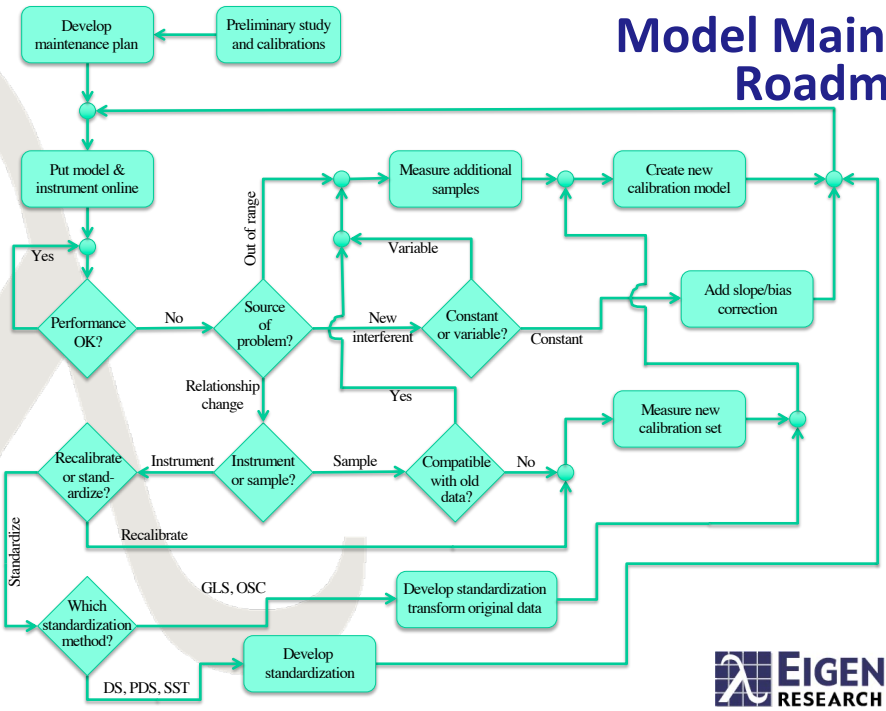
# Before Model Goes Online

- Develop a plan for maintenance
  - Assume that updated or new calibration models will eventually be required
  - Have a plan for how to detect the problem and what to do about it
  - Put it in the budget!
- Measure standard samples
  - Plan for registration and amplitude shifts
  - Characterize instrument in ranges important to model

96

EIGENVECTOR RESEARCH INCORPORATED

# Model Maintenance Roadmap




97

EIGENVECTOR RESEARCH INCORPORATED

# Model Maintenance Resources

- Webinars
  - https://eigenvector.com/aiovg_videos/evri-thing-you-need-to-know-about-how-to-build-and-deploy-instrument-standardization-models/
  - https://eigenvector.com/aiovg_videos/embed-instrument-standardization-directly-into-multivariate-models/
- Eigenvector University course live and recorded
  - https://eigenvector.com/events/calibration-model-maintenance/
- Journal article
  - B.M. Wise and R.T. Roginski, "Model Maintenance: the unrecognized cost in PAT and QbD," *Chemistry Today*, Vol. 33(2), pps. 38-43, March/April, 2015.

EIGENVECTOR RESEARCH INCORPORATED

# Overall Conclusions

EIGENVECTOR RESEARCH INCORPORATED

# Take Away Messages (1/2)

- Good calibration models start with good planning
  - Recognizing sources of variation
  - Experimental design or representative sample selection
- Review and screen data
  - Plotting and PCA
  - Find outliers
  - Identify trends and causes
- Develop regression model
  - Iterative process
  - Validate model

EIGENVECTOR RESEARCH INCORPORATED

# Take Away Messages (2/2)

- Optimize model
  - Many preprocessing methods available to handle various “problems”
  - Variable selection
  - Evaluate choices
- Save and document model
- Get model online
  - Variety of methods available
  - Selected method depends on infrastructure
- Plan for model maintenance
  - Most models don’t last forever

EIGENVECTOR RESEARCH INCORPORATED