Automated Peak and Peak-Ratio Selection for Regression and Classification Models of Raman and LIBS Data

Jeremy M. Shaver
Eigenvector Research, Inc.

Brian Marquardt, Tom Dearing, Sergey Mozharov, MarqMetrix

EIGENVECTOR RESEARCH INCORPORATED

NIR Shootout 2002

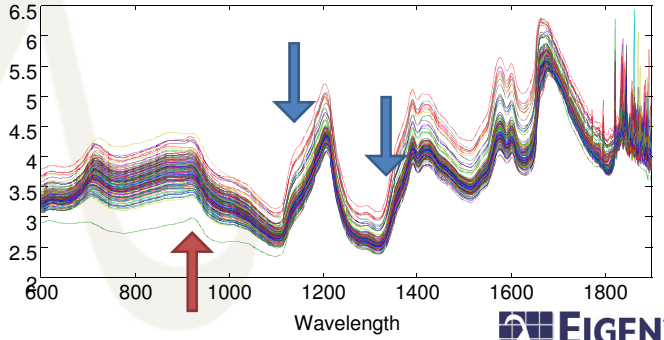
- 2002 International Diffuse Reflectance Conference (IDRC) "Shootout" data
  - NIR spectra
  - 654 pharmaceutical tablets
  - Calibration Set, Validation Set, Test Set
  - Two spectrometers
  - **Goal: best model with calibration transfer**
- Won by Karl Norris using "Norris Regression" – selected peaks and peak ratios including gap-segment derivative

EIGENVECTOR RESEARCH INCORPORATED

Norris' "Winning" Model

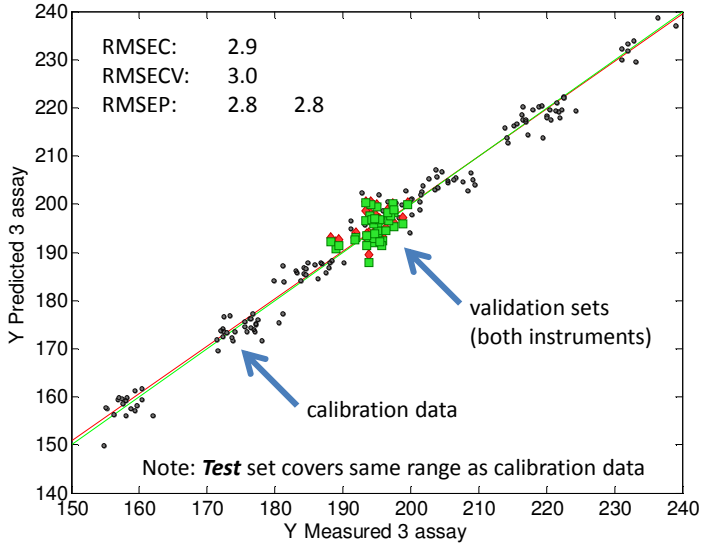
	Term 1			Term 2		
	Wavelength	Smooth	Gap	Wavelength	Smooth	Gap
Numerator	1142 nm	10 nm	26 nm	1338 nm	0 nm	22 nm
Denominator	920 nm	0 nm	30 nm			

Interactive manual selection of regions and smooth/gap parameters



EIGENVECTOR RESEARCH INCORPORATED

Results using (Approx.) Norris Regression

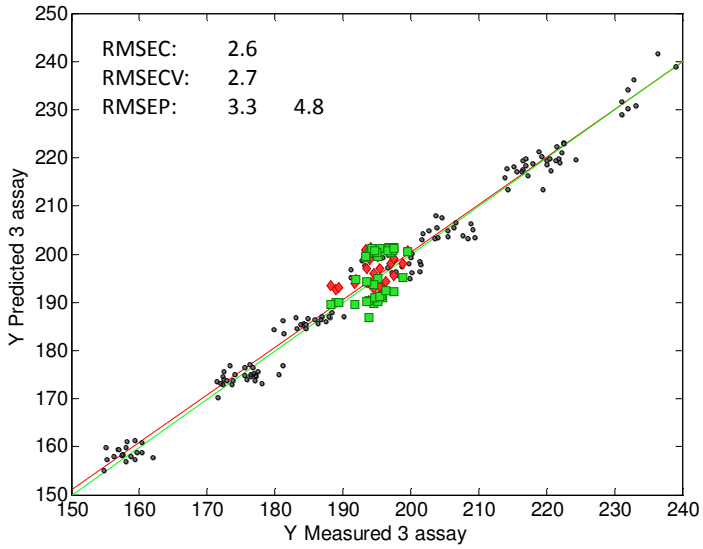


Preprocessing: 2nd Derivative (gap: 18 nm, segment: 6 nm) + Integrate + Autoscale

EIGENVECTOR RESEARCH INCORPORATED

EMSC + 1st Derivative + Mean Centering

EMSC = Extended Multiplicative Scatter Correction



EIGENVECTOR RESEARCH INCORPORATED

Tabulated Results

	RMSEC	RMSECV	Val 1	Val 2	Test 1	Test 2
Norris Regression	2.7	2.7	2.8	2.8	3.0	3.3
Expert-Selected Preprocessing	2.6	2.7	3.3	4.8	2.8	4.2

Good Model… but Bad Transfer

EIGENVECTOR RESEARCH INCORPORATED

Norris Regression – Generically

Non-linear Regression

$$y = b_1 (x_1)$$

$$y = b_1 (x_1 - x_2)$$ (Gap-Segment 1st Derivative)

$$y = b_1 \frac{x_1}{x_3}$$ (Peak Normalization)

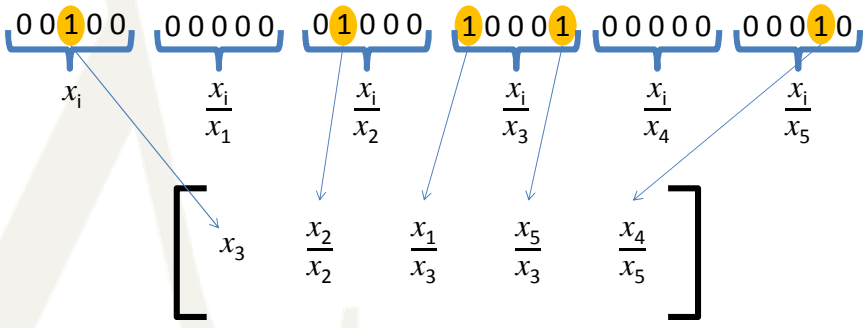
$$y = b_1 \frac{x_1 - x_2}{x_3 - x_4}$$ (Peak Normalization with variable- gap 1st derivative)

$$y = b_1 \frac{x_1 - x_2}{x_3 - x_4} + b_2 (x_5 - x_6) + b_3 x_7 + \ldots$$

EIGENVECTOR RESEARCH INCORPORATED

Binary Encoding of Norris Equations

- Example for 5 variables: $[\, x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \,]$

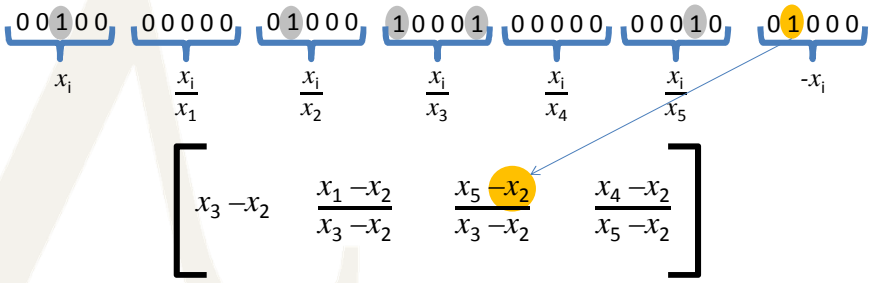


This much could be done by pre-computing…
but at a big memory cost
(525MB for shootout data)

EIGENVECTOR RESEARCH INCORPORATED

+ Allow Subtraction…

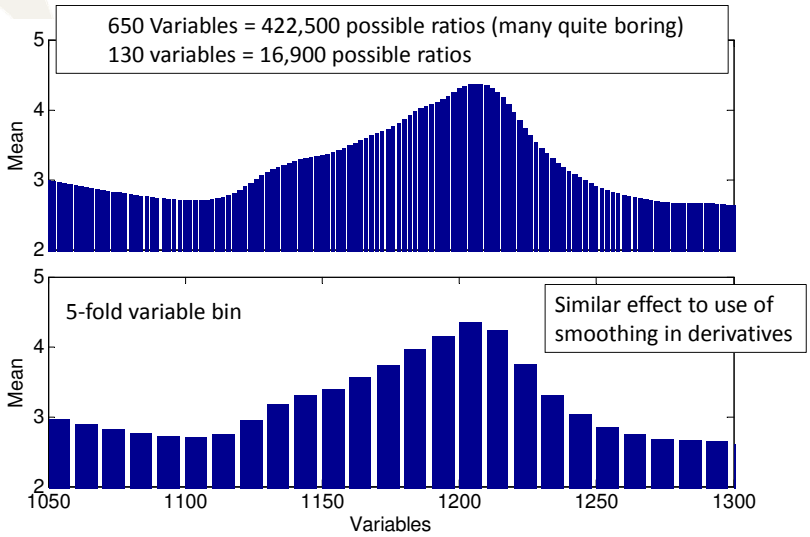
- Example for 5 variables: $[\,x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5\,]$
- One additional group to identify "baseline"



Pre-computation would now require
2 x10^{201} *variables* (for the shootout data)
variables = 2^n (n^2+n)

EIGENVECTOR RESEARCH INCORPORATED

+ Binning to Reduce Dimensionality



EIGENVECTOR RESEARCH INCORPORATED

+ Genetic Algorithm to Select Terms

- Try lots of combinations (Calculate variable ratios and offsets on-the-fly)
- Choose best cross-validated results
- Breed (intermix terms) and repeat
- Will refer to this as "GA-Norris"

- **Question:** Can this approach approximate what the interactive Norris approach does?

EIGENVECTOR RESEARCH INCORPORATED

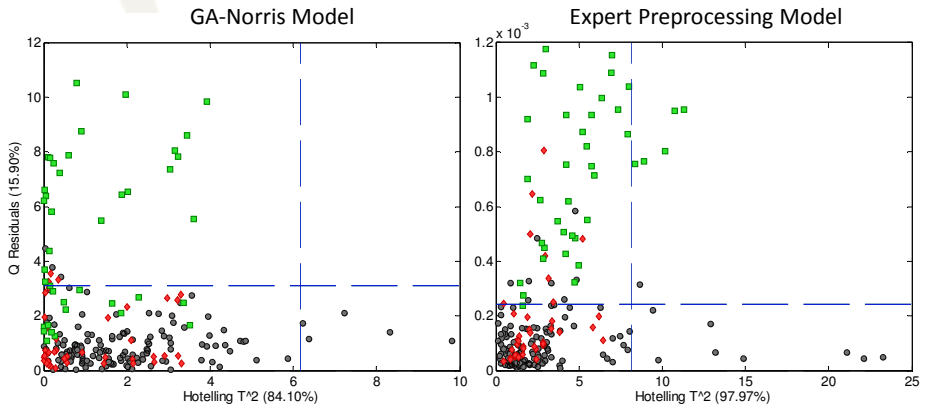
Tabulated Results

	RMSEC	RMSECV	Val 1	Val 2	Test 1	Test 2
Norris Regression	2.7	2.7	2.8	2.8	3.0	3.3
Expert-Selected Preprocessing	2.6	2.7	3.3	4.8	2.8	4.2
GA Norris (Cal 1 only)	2.4	2.5	3.9	5.0	2.8	3.7
GA Norris (Cal 1 & 2)	2.8	2.9	3.0	3.0	3.0	3.3
Simple GA (Cal 1 & 2)	2.6	2.7	3.7	3.8	3.3	3.5

Selecting Variables based on both instruments (building model from ONE) yields GA Norris preprocessing which closely approximates what Karl Norris did.

EIGENVECTOR RESEARCH INCORPORATED

Outlier Detection Achieved…



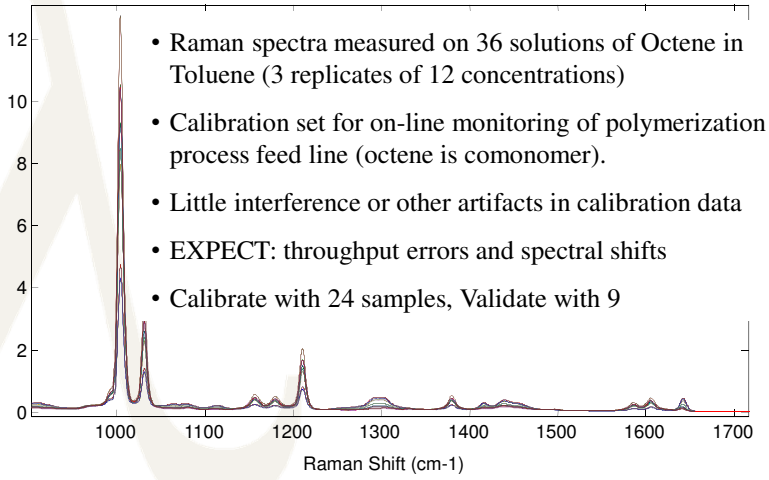
EIGENVECTOR RESEARCH INCORPORATED

Where Else Would Ratios Help?

- Raman – correcting for throughput differences and offsets
- LIBS – correcting for throughput differences and for emphasizing the importance of "relative abundance"

EIGENVECTOR RESEARCH INCORPORATED

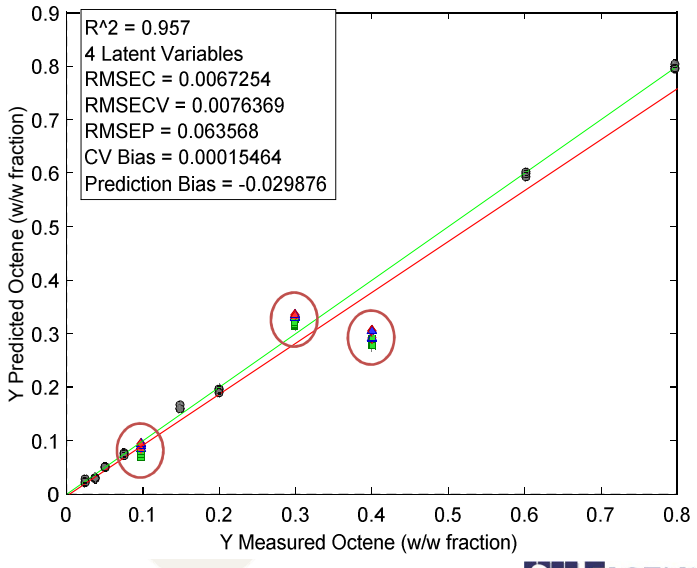
Raman of Octene in Toluene



- Raman spectra measured on 36 solutions of Octene in Toluene (3 replicates of 12 concentrations)
- Calibration set for on-line monitoring of polymerization process feed line (octene is comonomer).
- Little interference or other artifacts in calibration data
- EXPECT: throughput errors and spectral shifts
- Calibrate with 24 samples, Validate with 9

EIGENVECTOR RESEARCH INCORPORATED

15



EIGENVECTOR RESEARCH INCORPORATED

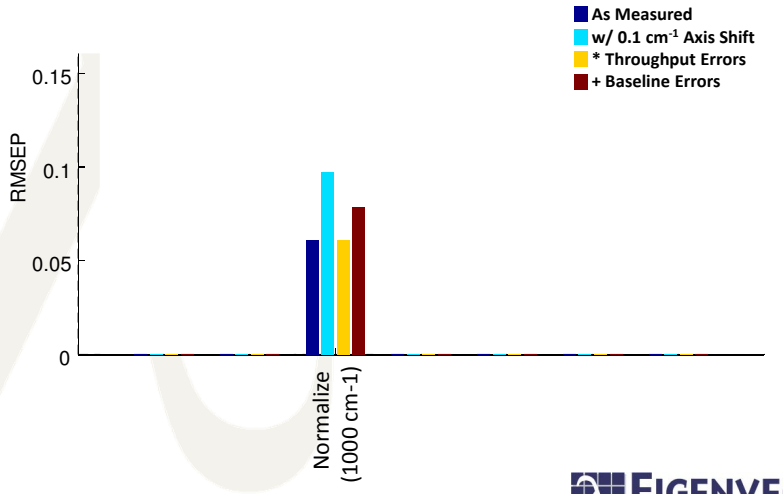
Prediction Error Vs. Interferences

Autoscale	Scale variables to unit standard deviation
Normalize	Divide by total intensity
Normalize (1000 cm^{-1})	Divide by intensity at 1000 cm^{-1} peak
1st Derivative	Savitzky-Golay 1st Derivative (15 point)
Whittaker Baseline	Automatic baseline subtraction
GA Norris	Binning + GA Norris Variable Selection + Ratios

(All methods also include mean centering)

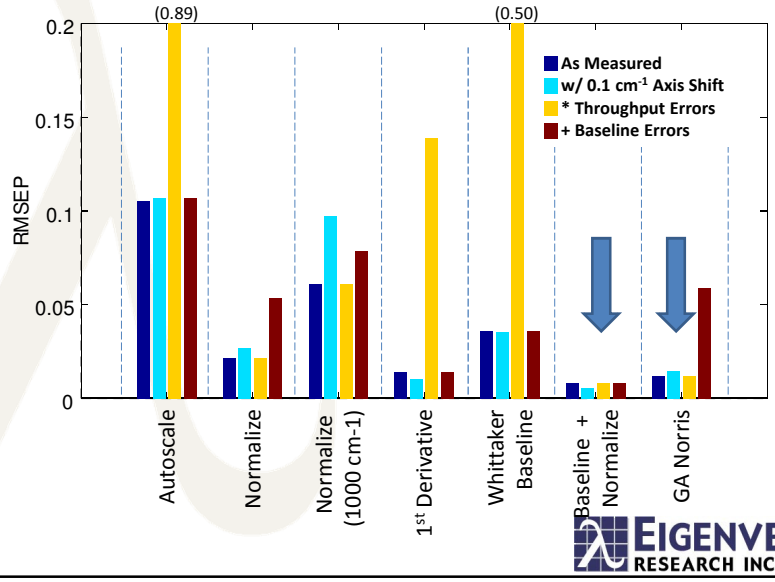
EIGENVECTOR RESEARCH INCORPORATED

Prediction Error Vs. Interferences



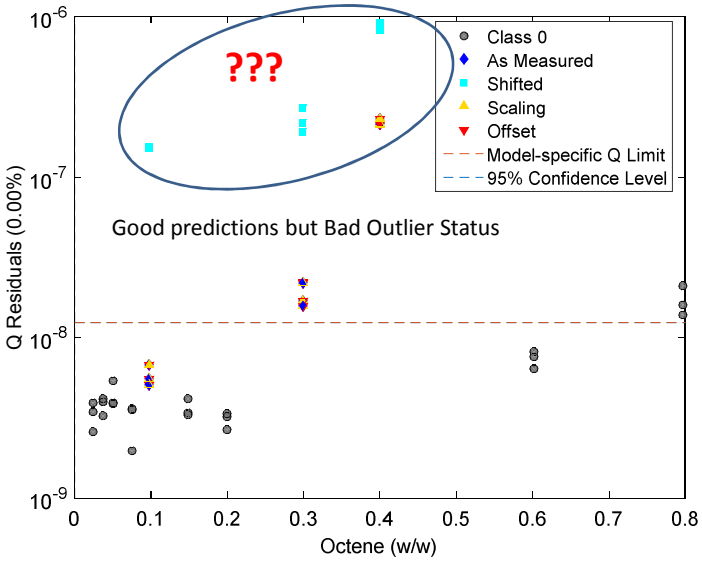
EIGENVECTOR RESEARCH INCORPORATED

Prediction Error Vs. Interferences



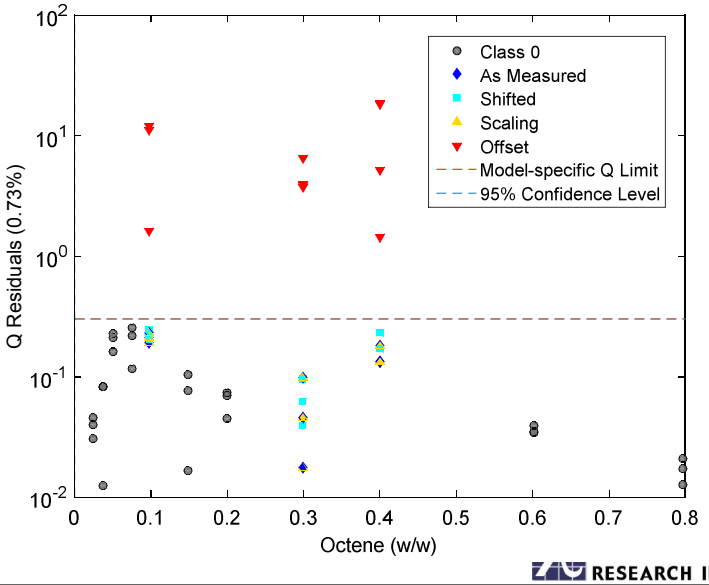
EIGENVECTOR RESEARCH INCORPORATED

Outlier Detection: Baseline+Norm



EIGENVECTOR RESEARCH INCORPORATED

Outlier Detection: GA Norris



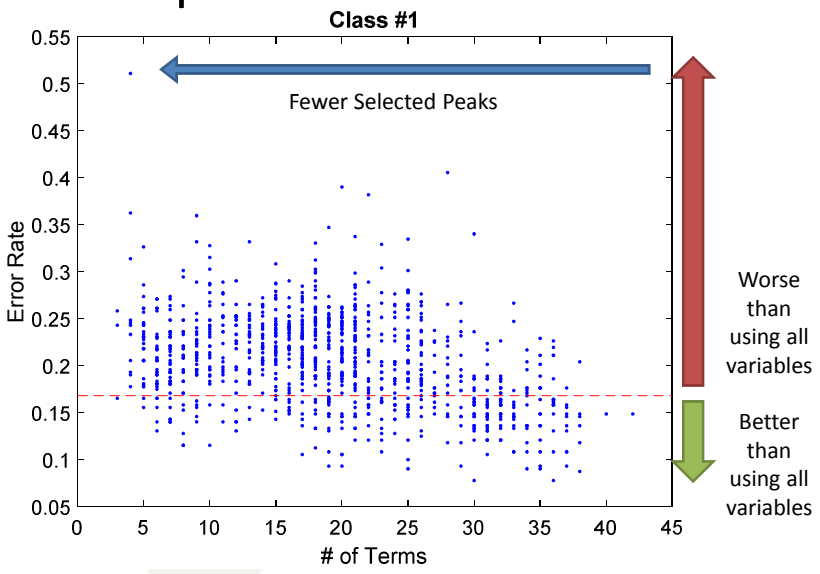
EIGENVECTOR RESEARCH INCORPORATED

LIBS / Raman Classification

- Mystery classes (natural product, difficult to separate classes)
- Raman data – not much information
- LIBS data – too much information
- Anticipate Peak Ratios should help greatly in LIBS!
- Try GA Norris on LIBS

EIGENVECTOR RESEARCH INCORPORATED

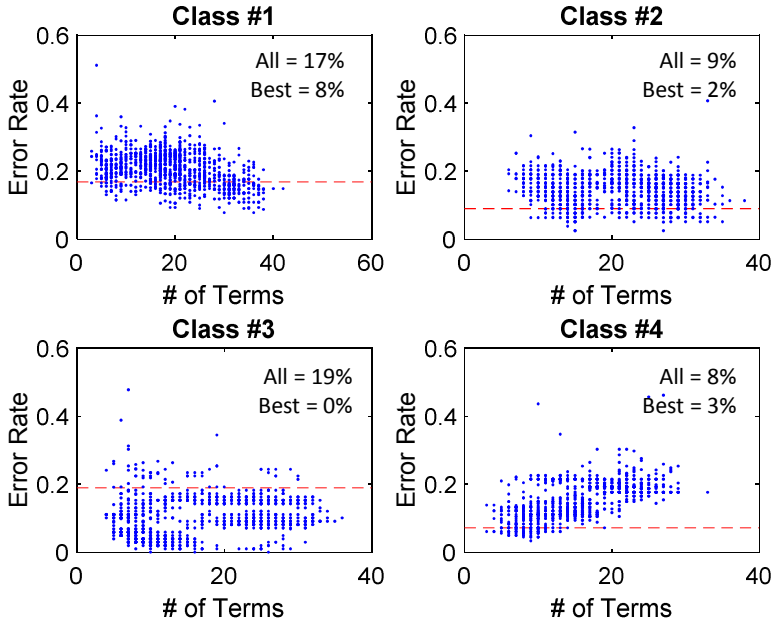
Example GA Norris Results



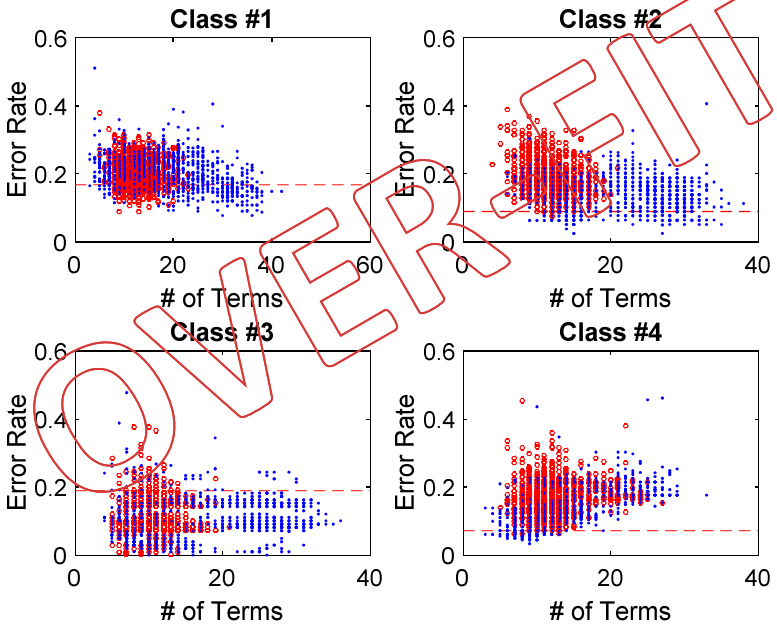
1000 Selection Results

EIGENVECTOR RESEARCH INCORPORATED

All Classes



With Randomized Classes (Red)



Non-linear model
+ variable selection
+ large domain

= large chance of over-fit

= use caution & permutation tests

EIGENVECTOR RESEARCH INCORPORATED

Conclusions

- GA Norris can reproduce Norris Regression results
- Can be used to achieve similar results to standard preprocessing (but with less sound decisions!)
- Large chance of over-fit = use caution & permutation tests, or standard methods!!