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

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
Norris' "Winning" Model

<table>
<thead>
<tr>
<th></th>
<th>Term 1</th>
<th>Term 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Numerator</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength</td>
<td>1142 nm</td>
<td>1338 nm</td>
</tr>
<tr>
<td>Smooth</td>
<td>10 nm</td>
<td>0 nm</td>
</tr>
<tr>
<td>Gap</td>
<td>26 nm</td>
<td>22 nm</td>
</tr>
<tr>
<td><strong>Denominator</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength</td>
<td>920 nm</td>
<td>1338 nm</td>
</tr>
<tr>
<td>Smooth</td>
<td>0 nm</td>
<td>0 nm</td>
</tr>
<tr>
<td>Gap</td>
<td>30 nm</td>
<td>22 nm</td>
</tr>
</tbody>
</table>

Interactive manual selection of regions and smooth/gap parameters

Results using (Approx.) Norris Regression

- **RMSECV**: 3.0
- **RMSEP**: 2.8

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

Note: **Test** set covers same range as calibration data
EMSC + 1\textsuperscript{st} Derivative + Mean Centering

EMSC = Extended Multiplicative Scatter Correction

\begin{center}
\begin{tabular}{c c c c}
\multicolumn{2}{c}{Tabulated Results} & & \\
\hline
 & 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 \\
\hline
\end{tabular}
\end{center}

Good Model... but Bad Transfer
Norris Regression – Generically
Non-linear Regression

\[
y = b_1 \left( x_1 \right)
\]

\[
y = b_1 \left( x_1 - x_2 \right)
\] (Gap-Segment 1\textsuperscript{st} 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 1\textsuperscript{st} derivative)

\[
y = b_1 \frac{x_1 - x_2}{x_3 - x_4} + b_2 \left( x_5 - x_6 \right) + b_3 x_7 + ...
\]

---

**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)
+ Allow Subtraction...

- Example for 5 variables: \[ \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} \]
- One additional group to identify "baseline"

\[
\begin{bmatrix}
0.1 & 0.0 & 1.0 & 0.0 & 0.0
\end{bmatrix}
\]

Pre-computation would now require
2 \times 10^{201} variables (for the shootout data)
variables = 2^n (n^2+n)

+ Binning to Reduce Dimensionality

650 Variables = 422,500 possible ratios (many quite boring)
130 variables = 16,900 possible ratios

5-fold variable bin
Similar effect to use of smoothing in derivatives
Gene 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?

---

**Tabulated Results**

<table>
<thead>
<tr>
<th></th>
<th>RMSEC</th>
<th>RMSECV</th>
<th>Val 1</th>
<th>Val 2</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norris Regression</td>
<td>2.7</td>
<td>2.7</td>
<td>2.8</td>
<td>2.8</td>
<td>3.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Expert-Selected Preprocessing</td>
<td>2.6</td>
<td>2.7</td>
<td>3.3</td>
<td>4.8</td>
<td>2.8</td>
<td>4.2</td>
</tr>
<tr>
<td>GA Norris (Cal 1 only)</td>
<td>2.4</td>
<td>2.5</td>
<td>3.9</td>
<td>5.0</td>
<td>2.8</td>
<td>3.7</td>
</tr>
<tr>
<td>GA Norris (Cal 1 &amp; 2)</td>
<td>2.8</td>
<td>2.9</td>
<td>3.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Simple GA (Cal 1 &amp; 2)</td>
<td>2.6</td>
<td>2.7</td>
<td>3.7</td>
<td>3.8</td>
<td>3.3</td>
<td>3.5</td>
</tr>
</tbody>
</table>

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

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"
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
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
- **1\(^{st}\) Derivative**: Savitzky-Golay 1\(^{st}\) Derivative (15 point)
- **Whittaker Baseline**: Automatic baseline subtraction
- **GA Norris**: Binning + GA Norris Variable Selection + Ratios

(All methods also include mean centering)

RMSEP

0

0.05

0.1

0.15

0.2

* Throughput Errors + Baseline Errors

**Graph**: Prediction Error Vs. Interferences

- As Measured
- w/ 0.1 cm\(^{-1}\) Axis Shift
- * Throughput Errors
- + Baseline Errors

**Axis**: RMSEP (0 to 0.15)

**X-axis**: Normalize (1000 cm\(^{-1}\))

**Y-axis**: Predicted values
Prediction Error Vs. Interferences

- RMSEP
- Autoscale
- Normalize
- Normalize (1000 cm⁻¹)
- 1st Derivative
- Whitaker Baseline
- Baseline + Normalize
- GA Norris

Outlier Detection: Baseline+Norm

- Good predictions but Bad Outlier Status
- Q Residuals
Outlier Detection: GA Norris

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
Example GA Norris Results

Class #1

Fewer Selected Peaks

Worse than using all variables

Better than using all variables

1000 Selection Results

All Classes

Class #1

All = 17%
Best = 8%

Class #2

All = 9%
Best = 2%

Class #3

All = 19%
Best = 0%

Class #4

All = 8%
Best = 3%
Non-linear model
+ variable selection
+ large domain
= large chance of over-fit
= use caution & permutation tests
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!!