A Practical Strategy for Spectral Library Partitioning and Least-Squares Identification

CASIS 2015

Shawn Higbee, PhD
Motivation

- This problem has no name:
  - En-masse use of least squares (LS) methods
  - Continuous arrival of large volumes of data
  - Results that need human interpretation
  - Highly correlated physical processes
  - Rank deficient spectral libraries
  - “standard” strategies aren’t helpful
  - Quantitative methods end up being surprisingly subjective

Goal: provide a practical strategy for navigating this situation
“Ordinary” Least Squares

- $Y = XB + \epsilon$
- $B_{\text{hat}} = (X'X)^{-1}X'Y$
- $\text{Cov}(B_{\text{hat}}) = \sigma^2(X'X)^{-1}$
- Most physical scientists pre-occupy with $\sigma^2$ and not $(X'X)^{-1}$
- Some useful tools in LS
  - $\text{SVD}(A) = UWV'$
  - $\text{cond}(A) \equiv \frac{\lambda_n}{\lambda_1}$
  - $VIF = \text{diag}(\text{Cor}(X)^{-1})$
- Fundament accuracy limit of LS: $\text{cond}(X) = 10^c \rightarrow \text{accuracy}(B_{\text{hat}}) \approx r - c$

A spectral library’s properties can dominate uncertainty in spectral ID
Standard Strategies

- **Regularization**
  \[ B_{rr} = (X'X + kI)^{-1}X'Y \]
  Note: \( \text{cond}(X'X + kI) \geq 2e9 \ \forall \ k < 1 \)

- **Principle Components Analysis**
  - Data lack an exploitable structure

Summary box is now has a full-width bleed
Alternative Strategies

- Library thinning
  - It’s not always practical to get rid of spectra

- Library partitioning
  - How many partitions?
  - Where to start assignment?
  - Criteria for each assignment?

- Criteria:
  - For any subset of the library – optimal partitions will have: $\lambda_i \times m/\text{trace}(W)$ above 1
  - SVD based assignment – maximize marginal condition number
  - VIF-based – minimize top three VIF values

Thinning and partitioning strategies leverage basic measurands of LS process
## Results

<table>
<thead>
<tr>
<th>Full Library</th>
<th>Seed Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random</td>
</tr>
<tr>
<td>condition number</td>
<td>47761</td>
</tr>
<tr>
<td>top 3 VIFs</td>
<td>201560</td>
</tr>
<tr>
<td></td>
<td>205120</td>
</tr>
<tr>
<td>Mean Cond</td>
<td>13114</td>
</tr>
<tr>
<td>Mean VIF</td>
<td>918</td>
</tr>
<tr>
<td>Max partition nth Singular Value</td>
<td>5899</td>
</tr>
<tr>
<td>Mean Cond</td>
<td>5899</td>
</tr>
<tr>
<td>Mean VIF</td>
<td>147</td>
</tr>
<tr>
<td>Min partition VIF</td>
<td>7430</td>
</tr>
<tr>
<td>Mean Cond</td>
<td>7430</td>
</tr>
<tr>
<td>Mean VIF</td>
<td>209</td>
</tr>
<tr>
<td>mean VIF</td>
<td>425</td>
</tr>
<tr>
<td>mean Cond</td>
<td>8814</td>
</tr>
</tbody>
</table>

## Conclusions

- Sizable reduction in error of both point and interval estimates is possible
- Significant tunability exists for specific CONOPS

Partition seed strategy is far less important than assignment criteria