ILS to CLS: Synergistic Regression Modeling for Improved Control and Interpretability

Neal B Gallagher
Eigenvecter Research, Inc.


Inverse least squares (ILS) models such as partial least squares and principle components regression are popular regression tools for chemometrics modeling. A major reason for this popularity is that extensive infrastructure has been developed to make model identification fast and easy. Additionally, statistical diagnostics provide tools to develop useful models for exploratory analysis and quantification tasks. However, although much work has gone into developing tools for interpretation of ILS models, classical least squares (CLS) models are superior for interpretation. CLS also provides more control during model identification and application because the model form is amenable to constraints that incorporate known physics and chemistry. Unfortunately, identification of CLS models is often more difficult than for ILS models – a property often attributed to interference signal present in measurements. This talk will show that the advantages of ILS and CLS can be used synergistically resulting in models that provide enhanced diagnostics and interpretability. Two examples typically modeled using ILS will be shown.

Booth 72
Summary

- Demonstrate the theory and practice of using inverse least squares (PLS and PCR) with classical least squares methods (CLS and ELS)
- PLS & PCR are fast and easy to identify
  - difficult to control, difficult to interpret and can be confused by coincidental correlation
- CLS & ELS can be difficult to identify
  - Extended Least Squares (extended mixture model)
  - easy to add what we know via constraints, easy to interpret shows if coincidental correlation is present

Inverse Least Squares

\[
y = Xb
\]

Partial Least Squares (PLS) and Principal Components Regression (PCR)

- \( X \) is the predictor (e.g., measured spectra)
- \( y \) is the predictand (e.g., concentration, univariate)
- \( b \) is the regression vector (univariate)

\[
Y = XB
\]

Multivariate \( Y \) is the more general case.

\[
B = W (T^T T)^{-1} T^T Y
\]
**ILS to ELS**

\[ Y = XB \quad T = XW \]

- **T** are scores (\( Y \) is in column space of **T**).
- **W** are weights for PLS and loadings for PCR (\( B \) is in the column space of **W**).

… rearrange the variance...

\[
\begin{bmatrix}
Y & T_{\perp}
\end{bmatrix} = X \begin{bmatrix}
B & W_{\perp}
\end{bmatrix}
\]

**OPLS**

\[
X = \begin{bmatrix}
Y & T
\end{bmatrix} \begin{bmatrix}
S & P_{\perp}
\end{bmatrix}^T
\]

**ELS**

\[
\begin{bmatrix}
Y, T_{\perp}, S, P_{\perp}
\end{bmatrix}
\]

**ELS Objective Function**

\[
O(Y, T_{\perp}, S, P_{\perp}) = \left( X - \begin{bmatrix}
Y & T_{\perp}
\end{bmatrix} \begin{bmatrix}
S & P_{\perp}
\end{bmatrix}^T \right)^T \left( X - \begin{bmatrix}
Y & T_{\perp}
\end{bmatrix} \begin{bmatrix}
S & P_{\perp}
\end{bmatrix}^T \right) + \left( I - P_{\perp} \right)^T A (I - P_{\perp})
\]

- **A** is a diagonal penalty factor. Orthogonality condition on **P** retains good mathematical conditioning.

Very good estimates of **Y** and **T** are available from PCR.

**Y** and **S** are often non-negative.
Why is this Important?

Inverse least squares methods like PCR and PLS are fast and easy to identify:
- Infrastructure and statistics are well defined
- Interpretability can be difficult (\(B\) are not spectra!)
- Many constraints for \(B\) don’t make physical sense
  - non-negativity and smoothness aren’t generally applicable
  - this hampers including physical knowledge into the objective function
- ILS provides very good guesses for \(Y\) and \(T\)
  - initial solutions for ELS

\[
\begin{bmatrix}
  Y & T \perp \\
\end{bmatrix}
= X \begin{bmatrix}
  B & W \perp \\
\end{bmatrix}
\]

---

Why is this Important

Forward least squares methods like CLS and ELS can be more difficult to identify:
- Infrastructure is less well developed
- Can include more of what we know via constraints during model identification and application
- Interpretability is as good as it gets (\(S\) are spectra!)
  - allows iterative model identification because the identification process teaches about the problem

\[
X = \begin{bmatrix}
  Y & T \\
\end{bmatrix} \begin{bmatrix}
  S & P \\
\end{bmatrix}^T
\]
NIR: Wheat Protein and Moisture

- Hard red winter wheat ground
- Calibration and validation sets measured at different times
- Cary-14 spectrometer system, 1000 – 2598.4 nm at 1.6 nm intervals, 3 nm resolution
- Protein by Kjeldahl, each sample measured 16 times, averaged
  - Estimated Standard Error of Laboratory, 0.14% protein for the averaged results

Data courtesy P. Williams and K. Norris

Error in the Reference Method

- Protein RMSEC = 0.13
- Every measurement has an error. The reference method has an error. How big is it?
Scores and Weights

The regression vector, $b$, is a linear combination of the weights, $W_k$.

Performance Comparison

<table>
<thead>
<tr>
<th></th>
<th>Protein (%)</th>
<th>Moisture (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model</td>
<td>RMSEC</td>
</tr>
<tr>
<td>ELS$^1$</td>
<td>0.13</td>
<td>NA</td>
</tr>
<tr>
<td>PCR$^2$</td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td>PLS1$^2$</td>
<td>0.14</td>
<td>0.16</td>
</tr>
<tr>
<td>PLS2$^2$</td>
<td>0.14</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Outliers (two) not removed, no serious attempt at model optimization.

Are the results significantly different based on what is known about the reference error?

$^1$used MSC only (it was a force fit through zero)

$^2$used MSC+mean-centering

Booth 72
MCR Estimates for WHELGA

Protein
RMSEC = 0.13

Moisture
RMSEC = 0.07

Interpretation

PLS regression vector for protein

water factor

protein factor

first derivative
**Interpretation**

ELS spectra for protein and moisture:

- Water factor
- H-bonded overtones
- CH overtones
- Protein factor

ELS spectra identified using multivariate curve resolution with soft equality constraints

Booth 72

---

**Interferences**

Interferences can be interpreted too... and the estimate of the factors might be tuned up with additional constraints

Factor 3 is a major interference

Factor 5 shows sharp bands

Factor 4 correlates with protein (color by protein)
SFCM Data Example

Estimate level in a slurry fed ceramic melter*

- measurements are not spectra
- measured 20 temperatures (thermocouples) in two vertical thermal wells
- thermocouples near the surface vary with level


Performance Comparison

Demonstration of ELS/CLS for engineering variables.
Mean-Centering for both PLS and ELS
3 factors for PLS and ELS

<table>
<thead>
<tr>
<th>Level (in)</th>
<th>PLS</th>
<th>ELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSEC</td>
<td>0.106</td>
<td>0.114</td>
</tr>
<tr>
<td>RMSECV</td>
<td>0.113</td>
<td>0.118</td>
</tr>
<tr>
<td>RMSEP</td>
<td>0.138</td>
<td>0.145</td>
</tr>
</tbody>
</table>
Constraints

- Constraints can be employed on both C and S during ELS model identification
  - e.g., non-negativity, smoothness, priors, time-series lagging, etc…
- … and on C during model application
- Allows imposing chemical and physical knowledge into the model identification
Non-Continuous Banding for Smoothness Constraint

Conclusions

ILS or CLS?  ILS + CLS!

- Inverse least squares methods like PCR and PLS are fast and easy to identify
  - Interpretability can be difficult (B are not spectra!)
- Forward least squares methods like CLS and ELS allow more control over model identification
  - Interpretability is as good as it gets (S are spectra!)