Symbolic Regression in Multicollinearity Problems

Flor Castillo, Carlos Villa

The Dow Chemical Company
Outline

• Why we need Symbolic Regression in multicollinearity
• A case study
• The proposed approach using GP
• Results
• Conclusions
GP in Multiple Linear Regression (MLR) Models

• GP has been used in two situations
  – design of experiments (DOE) scheme to solve lack of fit situations (LOF)

  – MLR with historical (plant data) to minimize multicollinearity (strong relationship among inputs)
Box-Behnken Experimental Design

Response (Output):
Particle size distribution of a chemical compound

Inputs:
• $X_1$, $X_2$, $X_3$, $X_4$

What if LOF is statistically significant?

$$S_k = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i<j} \beta_{ij} X_i X_j + \sum \beta_{ii} X_i^2$$
LOF: Model does not properly fit the data

Statistical test can detect LOF

p value for LOF < 0.05: Significant LOF
Possible LOF Solutions

• Ignore it
  – Possible limitations on conclusions
• Collect more data
  – Induce correlation
  – Cost of additional sampling, etc.
• Try a different more complex model
  – Current data may not support new model
• Try a different transformed model
  – Transformation to try not obvious (Genetic Programming (GP) can help)
Box-Behnken Data Analysis

Full Model
\[ R^2 = 0.88 \]

\[ S_k = \beta_0 + \sum_{i=1}^k \beta_i X_i + \sum_{i<j} \beta_{ij} X_i X_j + \sum_{i} \beta_{i1} X_i^2 \]

Reduced model (without \( X_1 \) terms)
\[ R^2 = 0.85 \]

Significant Lack-of-fit in full model

All Terms involving \( X_1 \) are not significant

Still, significant Lack-of-fit in reduced model
GP Generated transformations

Fit model in transformed variables

\[ S_k = \beta_0 + \sum_{i=2}^{4} \beta_i Z_i + \sum_{i<j} \beta_{ij} Z_i Z_j + \sum_{i=2}^{4} \beta_{ii} Z_i^2 \]

\[ y = \frac{x_2^{0.54528}}{\sqrt{\ln(x_3 x_2 + x_3)} \cdot x_2 x_4} \]

Variable transformations suggested by GP model

<table>
<thead>
<tr>
<th>Original Variable</th>
<th>Transformed Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_2)</td>
<td>(Z_2 = x_2^{0.5})</td>
</tr>
<tr>
<td>(x_3)</td>
<td>(Z_3 = [\log(x_3)]^{0.5})</td>
</tr>
<tr>
<td>(x_4)</td>
<td>(Z_4 = x_4^{\text{-1}})</td>
</tr>
</tbody>
</table>

Notice no significant Lack-of-fit p>0.05
Symbolic Regression in Multicollinearity Problems

- Plant data often becomes the focus of a modeling exercise.
- Initial model consider: multiple regression model (MLR)
- Issues with plant data
  - Data collinearity: relationship between inputs
  - Severe Multicollinearity:
    - Affects the precision of the estimated regression coefficients.
    - Can cause real concerns with the stability, validity, and usefulness of the resulting model
Possible Multicollinearity Solutions

- Use PCA, PLS to create independent meta-variables (linear combinations of inputs)
- Meta-variables are independent of each other however variable interpretation is a challenge (plant people)
- Collect additional data (not always feasible)
- Try a different transformed model
- GP can help minimize multicollinearity in MLR models.
Proposed Approach Using GP to Minimize Multicollinearity

1. Select models from GP-generated symbolic regression with $R^2 > 0.9$

2. Transform input variables according to the GP model

3. Fit Transformed First Order Model
   
   \[ y = \beta_0 + \sum_{i} \beta_i z_i + \sum_{j} \sum_{k} \beta_{jk} z_j z_k \]
   
   with transformed variables (TFOM)

4. Check error structure and perform variance stabilizing transformation if required

5. Error normally and independently distributed with mean zero and constant variance

6. Check Multicollinearity

7. VIF < 10

8. Stable polynomial model in original variables

1. Generate several GP models

2. Generate non-linear input transforms according to GP model

3. Fit MLR model in transformed variables

4. Perform statistical analysis and check Multicollinearity (check error structure, residuals, correlations (VIF))

5. Repeat steps 2-4 until a stable MLR model is obtained (multicollinearity is minimized)
Case study with small data set

The data set consisted of three inputs variables (x1-x3) and one response (y) from a chemical process

First order polynomial considered by MLR

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \]

| Term      | Estimate | t Ratio | Prob>|t| | VIF |
|-----------|----------|---------|-----|-----|-----|
| Intercept | -0.879   | -7.145  | <.0001 | .   |
| x1        | 0.265    | 1.526   | 0.137 | 5.46|
| x2        | -4.246   | -8.679  | <.0001 | 77.58|
| x1*x2     | 0.537    | 2.701   | 0.011 | 3.00|
| x3        | 2.549    | 5.518   | <.0001 | 68.20|
| x2*x3     | 0.891    | 4.318   | <.0001 | 1.69|

large Multicollinearity observed VIF>10
STEP 1. Generate GP models

$X_2$ was included most often

\[ f(x) = 0.38275 + 13196 \times 10^{42} \times \left( \frac{p(x_3, 5.121)}{p(x_2, 14.1394)} \right) \]

Error Statistics
- Corr. Coeff.: 0.98906
- Std. Dev.: 0.26842
- Rel. Error: 0.14753
- R²-Statistic: 0.97824
- RMSEP: 0.26842
- Ratio Nodes: 11

\[ y = \frac{0.38275 + 1.3196 \times 10^{42} \times x_3^{1.121}}{x_2^{14.1394}} \]
STEP 2. Generate input transforms according to GP models

<table>
<thead>
<tr>
<th>Original Variable</th>
<th>Transformed Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$Z_1$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$Z_2 = 1/x_2^{14}$</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$Z_3 = x_3^4$</td>
</tr>
</tbody>
</table>

Relationship revealed by GP model

STEP 3. Fit MLR model in transformed variables

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{23} z_2 z_3$$

STEP 4. Perform statistical analysis and check (check error structure, residuals, correlations (VIF))

Error structure shows departure from constant variance assumption
Variance stabilizing transformation needed:

• Box and Cox Transformation:

\[ y = y^\lambda \]

\[ \lambda = \text{Value that minimizes the SSE} \]

MLR model with variance stabilizing transformation:

\[ y^\lambda = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{23} z_2 z_3 \]
MLR model with variance stabilizing transformation

| Term                | Estimate | t Ratio | Prob>|t| | VIF |
|---------------------|----------|---------|------|-----|-----|
| Intercept           | 1.370    | 12.89   | <.0001 | .   |
| x1                  | -0.476   | -4.92   | <.0001 | 3.60 |
| 1/x2^14             | 0.493    | 4.52    | <.0001 | 5.35 |
| x1*(1/x2^14)        | -0.332   | -2.49   | 0.0180 | 3.66 |
| x3^5                | -0.241   | -3.00   | 0.0051 | 4.37 |

Improved model:
Stable polynomial model
No evidence of severe Multicollinearity
VIF<10

adequate error structure:
Normally and independently distributed errors with mean zero and constant variance
Case study with larger data set

• In another chemical process, data obtained from 3-month process history was used in empirical modeling effort
• A (detrimental) bi-product concentration was response (output) of interest
• All other variables considered potential inputs
• Can a reasonable empirical model be developed to predict how this bi-product output can be minimized?
Case study with larger data set

The data set consisted of thirteen inputs variables \((x_1-x_{13})\) and one response \((y)\) from a chemical process

First order polynomial considered by MLR

\[
y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \beta_{23}x_2x_3
\]

| Term | \(\beta\) Estimate | \(t\) Ratio | Prob>|t| | VIF |
|------|--------------------|-------------|--------|-----|
| Intercept | 230.70902 | 0.33 | 0.7432 |  |
| x1 | 0.9406677 | 19.31 | <0.0001 | 3.84056 |
| x2 | -2.428614 | -22.97 | <0.0001 | 7.05279 |
| x3 | 0.4005954 | 2.97 | 0.0041 | 9.42801 |
| x4 | -10.17105 | -0.36 | 0.7217 | 861.2503 |
| x5 | 2.956458 | 0.20 | 0.8385 | 343.7906 |
| x6 | 10.223555 | 0.36 | 0.7164 | 918.9986 |
| x7 | -31.91927 | -0.57 | 0.5686 | 3431.5002 |
| x8 | 14.871442 | 0.35 | 0.7257 | 1976.0583 |
| x9 | -135.1481 | -0.69 | 0.4919 | 1000231.8 |
| x10 | 117.8077 | 0.68 | 0.4967 | 964097.17 |
| x11 | 16.152238 | 0.40 | 0.6930 | 70850.669 |
| x12 | 14.186557 | 0.89 | 0.3750 | 77.489476 |
| x13 | -19.53814 | -0.67 | 0.5023 | 19404.123 |

Undesigned data will often be too unbalanced for standard modeling techniques

large Multicollinearity observed VIF>10
STEP 1. Generate GP models

Pareto front optimization used to select model with “best” balance between performance & complexity

\[ y = 10275 - 16078 \frac{x_6(x_2 + x_{11})}{x_1 + x_{13}} \]
STEP 2. Generate input transforms according to GP models

<table>
<thead>
<tr>
<th>Original Variable</th>
<th>Transformed Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2, x_{11}$</td>
<td>$Z_1 = (x_2 + x_{11})$</td>
</tr>
<tr>
<td>$x_1, x_{13}$</td>
<td>$Z_2 = 1/(x_1 + x_{13})$</td>
</tr>
<tr>
<td>$x_6$</td>
<td>$Z_3 = x_6$</td>
</tr>
</tbody>
</table>

- **STEP 3.** Fit MLR model in transformed variables

\[ y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{23} z_2 z_3 \]

| Term               | $\beta$ Estimate | t Ratio | Prob>|t| | VIF   |
|--------------------|-------------------|---------|------|------|
| Intercept          | 2955.597          | 16.616  | <0.0001 |      |
| $Z_1 = x_6$        | -7.265            | -5.812  | <0.0001 | 1.496 |
| $Z_1 = x_2 + x_{11}$ | -2.148           | -32.646 | <0.0001 | 2.504 |
| $Z_2 = 1/x_1 + x_{13}$ | -908023.43        | -21.148 | <0.0001 | 2.392 |

No multicollinearity problems
Conclusions

Approach using GP to minimized multicollinearity has been applied successfully in the Dow Chemical Company.

**Unique features of the proposed approach**
- Combine linear regression models (designed experiments, undesigned data) with GP generated models
- Uses the unique potential of GP generated models for suggesting variable transforms that minimized multicollinearity
- Maximizes the use of available data when model extrapolation is required

**Advantages of the approach**
- Produces stable polynomial (MLR model) with adequate error structure
- Provides a simple model which is easily understood by engineers and process people and offers
- Statistical analysis: outlier detection on the input space, influential observations and confidence band of the parameters can be applied offering additional assurance on the capabilities of the obtained model
- Improves model validation (alternative models)