FOPAM process data analytics workshop

Instructor: Leo Chiang
Guest speaker: Ivan Castillo
In collaboration with Richard Braatz and Joe Qin

August 6, 2019

Contributions from team Dow including Bea Braun, Swee-Teng Chin, Lloyd Colegrove, Mark Joswiak, Yoyo Peng, Ricardo Rendall, Alix Schmidt, Mary Beth Seasholtz, Monica Trevino, James Wade, Zhenyu Wang, and Mark Webb

Dow.com

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization
3.2 Outlier detection and data preprocessing
3.3 Method selection
3.4 How good is good enough? Industrial tips and tricks of the trade
3.5 Industrial case studies
Visualization: Data in context

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.76</td>
<td>5.14</td>
</tr>
<tr>
<td>1.46</td>
<td>5.34</td>
</tr>
<tr>
<td>1.63</td>
<td>5.47</td>
</tr>
<tr>
<td>0.81</td>
<td>1.23</td>
</tr>
<tr>
<td>1.17</td>
<td>5.45</td>
</tr>
<tr>
<td>0.25</td>
<td>1.89</td>
</tr>
<tr>
<td>1.93</td>
<td>5.25</td>
</tr>
<tr>
<td>0.02</td>
<td>1.40</td>
</tr>
<tr>
<td>3.33</td>
<td>1.19</td>
</tr>
<tr>
<td>0.17</td>
<td>1.16</td>
</tr>
</tbody>
</table>

... ... vs.

Visualization: Data in context

Correlation coefficient → caution
Example: Anscombe’s quartet

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of x</td>
<td>9</td>
<td>Exact</td>
</tr>
<tr>
<td>Variance of x</td>
<td>11</td>
<td>Exact</td>
</tr>
<tr>
<td>Mean of y</td>
<td>7.50</td>
<td>To 2 decimal places</td>
</tr>
<tr>
<td>Variance of y</td>
<td>4.125</td>
<td>+/- 0.003</td>
</tr>
<tr>
<td>Correlation between x and y</td>
<td>0.816</td>
<td>To 3 decimal places</td>
</tr>
<tr>
<td>Linear regression</td>
<td>y = 3.00 + 0.500*x</td>
<td>To 2, 3 decimal places</td>
</tr>
<tr>
<td>Coefficient of determination of linear regression</td>
<td>0.67</td>
<td>To 2 decimal places</td>
</tr>
</tbody>
</table>
How to visualize a dataset: Direct method

**Direct**

- Correlation
- Parallel coordinates

How to visualize a dataset: Indirect method

**Indirect**

Dimensionality reduction techniques (such as PCA) to preserve characteristics of original dataset

Principal component analysis is the industry workhorse
- Linear combinations of original variables
- Identifies correlations (structured variation)
- Leaves noise behind (unstructured variation)
- No parameters to tune
- Fast

**Analogy**
Interpreting the shadows of a complicated 3D-geometry after projection onto a 2D surface...
**Besides PCA, what are the other options?**

A variety of techniques exist
- Global/local
- Linear/non-linear
- Parametric/non-parametric
- Manifold learners
- Tunable parameters

**t-SNE is the current gold standard**

MNIST dataset (handwritten digits 0-9, colored by number) van der Maaten and Hinton *J. MLR 2009*

How it works:
- Point-point similarity is a probability, which is sought to be preserved in reduction.
- Create latent graph most similar to original graph. Minimize reconstruction error of weights in graph edges

\[
\min \sum_{e \in G} w_h \log \frac{w_h}{w_l} \tag{get clusters right}
\]

Identifies natural clusters
Solves crowding problem

**t-SNE** Local focus
t-SNE is the current gold standard…but here comes UMAP

**Uniform Manifold Approximation and Projection**


**Case study 1: What occurs prior to an unplanned event?**

UMAP and t-SNE immediately provide new insights

**Labels from plant engineers**

Joswiak et al., *Control Engr Practice* 2019
**Dimension reduction quality is a local/global tradeoff**

- **Local Quality**
  - UMAP & t-SNE are the best for local quality (even for only 2 dim.)

- **Global Quality**
  - PCA shows overlap of A and B
  - UMAP & t-SNE show outlier cluster
  - Quick visual analysis in just 2D
  - Near t-SNE local quality with ~better than average global quality
  - Clear (main) cluster separation even if data was not labeled
  - Comparing UMAP clusters yields 7 more variables of interest over PCA-based analysis

**Case Study 2: Different performance of two identical plants, A and B**

- PCA (and many other techniques) shows overlap of A and B
- UMAP & t-SNE show outlier cluster
- Quick visual analysis in just 2D
- Near t-SNE local quality with ~better than average global quality
- Clear (main) cluster separation even if data was not labeled
- Comparing UMAP clusters yields 7 more variables of interest over PCA-based analysis
Dimension Reduction Quality Is a Local/Global Tradeoff

Local Quality
Has my neighborhood changed?

Global Quality
Has my distance to others changed?

UMAP & t-SNE are the best for local quality (even for only 2 dim.)
UMAP outperforms t-SNE, especially at 2 dimensions

References

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization
3.2 Outlier detection and data preprocessing
3.3 Method selection
3.4 How good is good enough? Industrial tips and tricks of the trade
3.5 Industrial case studies

Some Opinions on Deep Learning vs Machine Learning
Some Opinions on Deep Learning vs Machine Learning

Data: Is it any good? How do you know?

Data must be analyzed in context.
Do You Know Your Data?

What is accuracy?

What is precision?

A Quick Scenario
A Quick Scenario

Diagram 1

A Quick Scenario

Diagram 2
A Quick Scenario

Variable transformation

Raw Y variable

Log(Y)

The model is predicting negative concentration values
How to incorporate measurement uncertainty?

Measurement uncertainty near physical limits (absolute std. dev.)

Truncation of the interval
- First computing expanded uncertainty intervals using the current statistical methods (frequencist approach).
- Then, the interval is truncated in order to cover only the values in the feasible region.

Bayesian intervals
Introduces the requirement of respecting the feasibility domain, through a prior distribution presenting finite density only in the feasible region.

Missing value estimation

Random missing values are manageable: the NIPALS (PLS) algorithm works well.

Systematic missing value are problematic and there is no one-size-fit-all solution:
Start with:
1) Remove these variables
2) Use only data points with non-missing values

Outliers

Pop quiz

What is the first thing you do when you see an outlier?
A. Eliminate the point
B. Investigation
C. Assume it is not an issue
D. Correct the number
Pre-processing methods that are less sensitive to outliers

Auto scaling (gold standard)
- Data are mean centered and then divided by the standard deviation
- With the presence of outliers, mean and standard deviation are biased

Robust scaling
- Replace mean with median
- Replace standard deviation with MAD (median absolute deviation from median)

Dow modified scaling
- For each variable, find the \( n/2 \) observations that are closest to the median
- Use these \( n/2 \) observations to determine median and standard deviation

Robust outlier detection algorithm

- RHM (observations with small vector lengths after resampling)
- SHV (observations that are close together)
- CDC (observations that are close to the mean)
- MVT (observations with small $T^2$ stat. after iterations)
- CDC/MVT (use CDC in the initial step for MVT)
Outlier detection and data preprocessing summary

Outlier exclusion → invalid values, e.g. clamp transform for valid outliers
Missing data → NIPAS (PLS) algorithm works well for random missing values
Variable transformation → log
Normalization → mean centering, variance standardization
Robust statistics → less sensitive to outliers

Use your domain knowledge (e.g., incorporate uncertainty into decision making)

References

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization
3.2 Outlier detection and data preprocessing
3.3 Method selection
3.4 How good is good enough? Industrial tips and tricks of the trade
3.5 Industrial case studies

Challenge of method selection?
That’s why data scientist has a job!
Motivation: Why Utilize Feature Selection?

Feature Selection is the elimination of irrelevant variables from the model.

1) Feature selection is relevant because it helps us address the “curse of dimensionality”:
   - Decreases the risk of overfitting
   - Improves prediction accuracy
   - Eliminates irrelevant and redundant features
   - Improves interpretability
   - Decreases computational time

2) Drastically simplify the complexity of implementation

Current Feature Selection Approaches

Feature Selection Methods

- Filter Methods: Use variable ranking methods. A threshold is used to remove variables.
- Wrapper Methods: Subset of variables is generated before and selection is based on the model performance.
- Embedded Methods: Subset of variables is proposed and evaluated during model development.
- Hybrid Methods: Combination of filter, wrapper and embedded approaches.
- Other Methods: Unsupervised approaches.

How can these methods be applied in Big Chemical Data?
Wide spectrum feature selection (WiSe) Approach

1. Remove irrelevant or noisy features
2. Wrappers and/or embedded methods for further feature selection

Stage 1: Ranking Relevance to Response Variable

Ranking methods are helpful to understand the relevance of a feature

Most utilized filtering methods are entropy-based and statistical

- Information Gain (IG)
- Gain Ratio (GR)
- Symmetrical Uncertainty (SU)
- Mutual Information (MI)
- Pearson's correlation
- Chi-square test
- Spearman's Correlation

Filtering methods can be applied without significant computational burden (Thousands of variables)
Stage 1: Selected Filters

Univariate filters to efficiently remove noisy features
- Pearson correlation for linear relationships
- Spearman correlation for monotonic relationships
- Symmetrical Uncertainty (SU) for non-linear relationships
- Combinations of the aforementioned methods

![Graph showing correlation types: Linear, Monotonic, Non-linear]

Stage 1: Determining the Threshold for Removal

Noise levels in the data are estimated utilizing random permutations
- Compute correlation metric for 100 random shuffles, using all features
- Estimate the p-value for each feature
- Select feature if p-value is below 0.2

![Histogram showing correlation metric distribution with a threshold of 0.133]
Stage 2: Model Building

The second step of feature selection is based on wrapper and embedded methods
- Forward Stepwise Regression
- LASSO
- Partial Least Squares

Dow data set

Industrial batch process of a functionalized silicone polymer
- Continuous quality parameter as a Y variable
- 29 batch conditions, 7 un-aligned batch trajectories of ~100 points
- >600 batches
Our classic data-driven analysis problems

- **Continuous Process**
  - Variables, p
  - Many samples, n, (time)
  - \( n >> p \)
  - Steady-state continuous processes
  - Dynamic behavior
  - Correlated variables
  - Usually for quality prediction

- **Batch Process**
  - Variables
  - End Properties
  - Initial Conditions
  - Variable Trajectories
  - Usually transient and dynamic
  - Data variety / volume challenge, need to match context, multi-dimensional data
  - High frequency, large volume datasets


---

**Generated Features for industrial data set**

Features: 29 batch conditions + SPA features for 7 trajectories

The following features are computed using Statistical Pattern Analysis (SPA):

- Means of all process variables
- Variance of all process variables
- Skewness
- Kurtosis
- Covariance between variables

Results for Case 2 – Industrial Dataset

In this data, the filter combinations eliminated the most features and also resulted in the best model performance on test data.

Conclusions

1. Filters efficiently reduced dimensionality in the first stage
   - Although performance is dependent on the dataset, most of the important predictors were selected
   - Many irrelevant variables were removed

2. Usefulness of eliminating features was demonstrated on an industrial dataset
   - The adoption of filters led to improved prediction performance across the three regression methods
   - Interpretation of which features are selected can bring insights
Machine Learning Algorithms


References

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization
3.2 Outlier detection and data preprocessing
3.3 Method selection
3.4 How good is good enough? Industrial tips and tricks of the trade
3.5 Industrial case studies

"All Models are wrong, but some are useful"
George Box

So, how good is good enough?
How to avoid overfitting

- Tune model parameters
- Tune hyper-parameters
- True Evaluation

Training | Validation | Testing

*Do not use the testing data AT ALL when you are developing the model

Bias / Variance Trade-Off

- Include more variables
- Transform/generate variables
- Add more components
- Remove outliers / ill-fitting data
- Use nonlinear techniques
Another Dow example

New data
New data means new insight

One of the worst models I've seen at Dow
Traditional Approach

Who contributes more to the variability?

Lab  Process

How to Improve Data Quality?

Design of Experiments
**Design Of Experiments (DOE)**

- Batches: 1 to 6
- Unit: A, B
- Unit Batch: 2
- Location: Top-center, Middle-side
- Sample: 1, 2
- Operator: 1, 2
- Analyzer: 1, 2

**Systematic Statistical Approach**

- Historical data
- DOE plant data
- Model prediction
- Actual value
Results: Process Variables that cause the difference performances in one major unit

Speed of propeller in Unit

![Graph showing Speed of propeller in Unit A and Unit B over time]

Image Classification at Dow

Good vs Bad

The shape of plastic pellets is a major quality factor

Deep Neural Networks for Image Classification

A deep neural network contains many layers:
- Multiple layers allow the learning of high level features
- Each layer computes a specific function

Interesting Points of DNN:
- Convolutional layers and other type of layers
- Better optimization tools and other developments (ReLU activation, batch normalization, transfer learning, dropout, etc.)

Convolutional layers:
- Contain filters that convolve with the input, outputting a matrix
- The parameters of the filter are optimized with training data
Deep Neural Networks for Image Classification

Transfer Learning:
- A pre-trained network is modified to a different classification task
- Relevant features in the original domain tend to be useful in the target domain

VGG-16 (original) → VGG-16 (Transfer Learning)

Deep Neural Networks for Image Classification

Two deep neural networks were tested

### Simpler Deep Neural Network (SDNN)

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output Shape</th>
<th>Param. #</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv2d_1 (Conv2D)</td>
<td>(None, 46, 46, 33)</td>
<td>320</td>
</tr>
<tr>
<td>conv2d_2 (Conv2D)</td>
<td>(None, 44, 44, 44)</td>
<td>10444</td>
</tr>
<tr>
<td>max_pooling2d_1 (MaxPooling2D)</td>
<td>(None, 22, 22, 44)</td>
<td>0</td>
</tr>
<tr>
<td>dropout_1 (Dropout)</td>
<td>(None, 22, 22, 44)</td>
<td>0</td>
</tr>
<tr>
<td>flatten_1 (Flatten)</td>
<td>(None, 1024)</td>
<td>0</td>
</tr>
<tr>
<td>dense_1 (Dense)</td>
<td>(None, 128)</td>
<td>1045056</td>
</tr>
<tr>
<td>dropout_2 (Dropout)</td>
<td>(None, 128)</td>
<td>0</td>
</tr>
<tr>
<td>dense_2 (Dense)</td>
<td>(None, 2)</td>
<td>208</td>
</tr>
</tbody>
</table>

Total parameters 3,384,130
Non-trainable parameters: 6

### VGG-16 with Transfer Learning

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output Shape</th>
<th>Param. #</th>
</tr>
</thead>
<tbody>
<tr>
<td>input_1 (InputLayer)</td>
<td>(None, 96, 96, 3)</td>
<td>0</td>
</tr>
<tr>
<td>block1_conv1 (Conv2D)</td>
<td>(None, 96, 96, 64)</td>
<td>1792</td>
</tr>
<tr>
<td>block1_conv2 (Conv2D)</td>
<td>(None, 96, 96, 64)</td>
<td>24960</td>
</tr>
<tr>
<td>block1_pool (MaxPooling2D)</td>
<td>(None, 48, 48, 64)</td>
<td>0</td>
</tr>
<tr>
<td>block2_conv1 (Conv2D)</td>
<td>(None, 48, 48, 128)</td>
<td>73454</td>
</tr>
<tr>
<td>block2_conv2 (Conv2D)</td>
<td>(None, 48, 48, 128)</td>
<td>147704</td>
</tr>
<tr>
<td>block2_pool (MaxPooling2D)</td>
<td>(None, 24, 24, 128)</td>
<td>0</td>
</tr>
<tr>
<td>dense_1 (Dense)</td>
<td>(None, 1024)</td>
<td>524288</td>
</tr>
<tr>
<td>dense_2 (Dense)</td>
<td>(None, 2)</td>
<td>2050</td>
</tr>
</tbody>
</table>

Total parameters 15,282,590.0
Non-trainable parameters 317,362.0
Classification Methods Tested

Process experts manually labelled ~6000 images:
- These were split in training, validation and test sets
- Different classifiers were tested
  - PSSD
  - Random Forests
  - Deep Neural Networks

Classification Results

Good vs Bad Pellets
- Training: 2961 samples
- Validation: 1777 samples
- Test: 1185 samples

<table>
<thead>
<tr>
<th>Set</th>
<th>PSSD</th>
<th>Random Forests</th>
<th>SDNN</th>
<th>SDNN</th>
<th>Transfer Learning (VGG-16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0.816</td>
<td>1</td>
<td>0.98</td>
<td>0.964</td>
<td>0.971</td>
</tr>
<tr>
<td>Validation</td>
<td>0.817</td>
<td>0.941</td>
<td>0.913</td>
<td>0.956</td>
<td>0.966</td>
</tr>
<tr>
<td>Test</td>
<td>0.805</td>
<td>0.937</td>
<td>0.917</td>
<td>0.957</td>
<td>0.967</td>
</tr>
</tbody>
</table>

1 Approaches based on features
2 Uses sample augmentation techniques
Complex interaction between analytics and expert knowledge

Neither one alone is sufficient

Analytics Model

Expert Domain Knowledge

Both components are needed to improve model quality

Analytics culture change

• Innovation
• Advanced analytics and programming tools

Chemometric modelers

• Collaboration
• Data acumen (Special analytics and programming tools)

400+ Data scientists

• Foundation
• Data literacy/acuity (Practitioner analytics tools)

35,000+ Dow people

• Art of the possible
• Integrate data analytics into ChE curriculum

500,000+ US STEM graduates

Qin and Chiang, 2019
Chiang et al., 2017
References


3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization

3.2 Outlier detection and data preprocessing

3.3 Method selection

3.4 How good is good enough? Industrial tips and tricks of the trade

3.5 Industrial case studies
3.5 Industrial Case Studies

- Case Study 1: Impurity Estimation
- Case Study 2: Quality Classification

Case Study 1: Impurity Estimation

Impurity levels are constantly increasing, affecting production rate and the catalyst life of the reactor.
Objective: Identify key variables that affect impurity levels

Case Study Variables

<table>
<thead>
<tr>
<th>Column Variables</th>
<th>Column1 Variables</th>
<th>Column2 Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1: Column Reflux Flow</td>
<td>x22: Column1 Base Concentration</td>
<td>x36: Column2 Recycle Flow</td>
</tr>
<tr>
<td>x2: Column Tails Flow</td>
<td>x23: Flow from Input to Column1</td>
<td>x37: Column2 Tails Flow to Column</td>
</tr>
<tr>
<td>x3: Input to Column Bed 3 Flow</td>
<td>x24: Column1 Tails Flow</td>
<td>x38: Column2 Calculated DP</td>
</tr>
<tr>
<td>x4: Input to Column Bed 2 Flow</td>
<td>x25: Column1 Tray DP</td>
<td>x39: Column2 Steam Flow</td>
</tr>
<tr>
<td>x5: Column Feed Flow from Column2</td>
<td>x26: Column1 Head Pressure</td>
<td>x40: Column2 Tails Flow</td>
</tr>
<tr>
<td>x6: Column Make Flow</td>
<td>x27: Column1 Base Pressure</td>
<td>x41: Column2 Tails Flow</td>
</tr>
<tr>
<td>x7: Column Base Level</td>
<td>x28: Column1 Base Temperature</td>
<td>x42: Column2 Tails Temperature</td>
</tr>
<tr>
<td>x8: Column Reflux Drum Pressure</td>
<td>x29: Column1 Tray 3 Temperature</td>
<td>x43: Column2 Tails Temperature 1</td>
</tr>
<tr>
<td>x9: Column Condenser Reflux Drum Level</td>
<td>x30: Column1 Bed 1 Temperature</td>
<td>x44: Column2 Tails Temperature 2</td>
</tr>
<tr>
<td>x10: Column Bed 1 DP</td>
<td>x31: Column1 Bed 2 Temperature</td>
<td>x45: Column2 Tails Temperature 3</td>
</tr>
<tr>
<td>x11: Column Bed 2 DP</td>
<td>x32: Column1 Tray 2 Temperature</td>
<td>x46: Column2 Tails Temperature 4</td>
</tr>
<tr>
<td>x12: Column Bed 3 DP</td>
<td>x33: Column1 Tray 1 Temperature</td>
<td>x47: Column2 Tails Temperature 5</td>
</tr>
<tr>
<td>x13: Column Bed 4 DP</td>
<td>x34: Column1 Tails Temperature</td>
<td>x48: Column2 Tails Temperature 6</td>
</tr>
<tr>
<td>x14: Column Base Pressure</td>
<td>x35: Column1 Tails Concentration</td>
<td>x49: Column2 Tails Temperature 7</td>
</tr>
<tr>
<td>x15: Column Head Pressure</td>
<td>Avg_Outlet_Impurity</td>
<td>x50: Column2 Tails Temperature 8</td>
</tr>
<tr>
<td>x16: Column Tails Temperature</td>
<td>Avg_Delta_composition_column</td>
<td>x51: Column2 Tails Temperature 9</td>
</tr>
<tr>
<td>x17: Column Tails Temperature 1</td>
<td>y: Impurity</td>
<td>x52: Column2 Tails Temperature 10</td>
</tr>
<tr>
<td>x18: Column Bed 4 Temperature</td>
<td>Avg_Delta_composition_column</td>
<td>x53: Column2 Tails Temperature 11</td>
</tr>
<tr>
<td>x19: Column Bed 3 Temperature</td>
<td>Column reflexivity</td>
<td>x54: Column2 Tails Temperature 12</td>
</tr>
<tr>
<td>x20: Column Bed 2 Temperature</td>
<td>Column reflexivity</td>
<td>x55: Column2 Tails Temperature 13</td>
</tr>
<tr>
<td>x21: Column Bed 1 Temperature</td>
<td>Avg_Delta_composition_column</td>
<td>x56: Column2 Tails Temperature 14</td>
</tr>
<tr>
<td>Avg_Outlet_Impurity</td>
<td>x57: Column2 Make/Reflux</td>
<td>x58: Column2 Tails Temperature 15</td>
</tr>
<tr>
<td>Avg_Delta_composition_column</td>
<td>x59: Column2 Make/Reflux</td>
<td>x60: Column2 Tails Temperature 16</td>
</tr>
<tr>
<td>y: Impurity</td>
<td>ImpurityDataset_Validation.xlsx</td>
<td>ImpurityDataset_Training.xlsx</td>
</tr>
</tbody>
</table>
**Workflow**

Outlier detection and data preprocessing

Visualizing data

Model development: Variable selection

Model Validation: Is the model good enough?

---

**Missing values**

![Missing value visualization](image_url)
Outlier Detection (Visualizing Data Utilizing PCA)

By Applying PCA, outliers can be detected and removed.

Are These Outliers?
Visualizing Data After Eliminating Outliers

Are these clusters representing changes in operating conditions?

Significant changes in operating conditions from 8/22/2016 to 12/16/2016

Business reduced column head pressure
Model Development

The model is predicting negative values in the impurity value

The model prediction when applying a log transformation of the Y

Accuracy can be improved by applying a nonlinear transformation to the output variable

Variable Selection

Process knowledge is critical to select key variables
Model Results

Is the model good enough? Best practice is to verify model accuracy by utilizing a validation dataset.

Model Results

Is the model good enough? Are the selected variables in correspondence with first principles knowledge and plant operation?
Comparing Performance Utilizing Lasso (Penalized Methods)

Model Development Results

Similar performance than PLS. More variables selected by this method, requiring further elimination of variables.

Comparing Performance with Lasso (Model Validation Results)

Between March and April 2017, the model is not capable to predict fast changes in the impurity. The PLS model has a better performance during the same time frame.
Case Study 2: Quality Classification-- Evaluating the Performance of Multiple Reactors

Batch operation (10 reactors, multiple catalyst, 38 variables per reactor)

Goal: Identify root cause of underperformance

Features:
- Mean, Stdev, Skewness, Kurtosis, Pairwise Correlation
- Max, Min, Range, Medium, Slope, Area under the curve, Begin/End Delta
- Autocorrelation (lag=1)

Supervised Case Study: Evaluating the Performance of Multiple Reactors

When applying PLS-DA and UMAP, the separation of classes is unclear
Separation Between Classes Can Be Improved by Performing Feature Selection

Objective: Find variables that separate classes and minimize number of clusters based on the following metric:

\[ f = \sum_{i=1}^{n} \sum_{j=1}^{k} \left( x_{ij} \right) \left( x_{ij} - \overline{x}_{ij} \right)^2 \]

- \( x_{ij} \) refers to cluster \( i \) (is the cardinality)
- \( n \) refers to given class/label
- \( k \) is the indicator function where if \( x \) is true else \( 0 \) = 0

\[ f_{PLS-DA} = \sum_{i=1}^{n} \sum_{j=1}^{k} \left( x_{ij} \right) \left( x_{ij} - \overline{x}_{ij} \right)^2 \]

- \( f \) refers to the number of features
- \( n \) refers to the number of neighbors (hyper-parameter used for UMAP)
- \( D \) is a constraint on the number of clusters

Selecting Relevant Features

The selected features for each method are not consistent and are located in different order. Which classifier model is better?
## Classifier Performance (Validating with new data)

Performance classifiers utilizing full feature space

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Random Forest</th>
<th>PLS-DA</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>99.4%</td>
<td>98.0%</td>
<td>93.7%</td>
</tr>
<tr>
<td>Bad</td>
<td>54.2%</td>
<td>32.2%</td>
<td>49.2%</td>
</tr>
</tbody>
</table>

Bad class performance is not ideal

Performance classifiers built upon selected features

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Top 10 Features (MI)</th>
<th>Top 10 Features (VIP)</th>
<th>Top 10 Features (RF)</th>
<th>Top 5 Features (RF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>98.8%</td>
<td>96.1%</td>
<td>98.8%</td>
<td>99.2%</td>
</tr>
<tr>
<td>Bad</td>
<td>89.8%</td>
<td>39.0%</td>
<td>79.8%</td>
<td>86.4%</td>
</tr>
</tbody>
</table>

Model validation is very helpful to identify best classifier model

## Summary – Case Studies

- Two supervised case studies were illustrated by utilizing industrial cases studies. Feature selection and model validation are key steps to evaluate model performance.
- Process knowledge is key for generating best models. Dimensionality reduction techniques are helpful to visualize high dimensional data and bring process understanding
Package Resources

Visualization: Tableau, PowerBi, Python-plotly, seaborn and matplotlib
Design of Experiment: JMP
Random Forest: Python- sklearn
Dimensionality Reduction: Python-sklearn, Matlab-toolbox by Laurens van der Maaten
PLS/PLS-DA: Sartorius-Stedim/Umetrics SIMCA, Python-pychemometrics
Deep Neural Networks: Python-keras (tensorflow)