PREAMBLE

USING FUNCTIONAL DATA ANALYSIS FOR TIME-DEPENDENT OPTIMIZATION OF BATCH PROCESSES

• Wolfgang Ketterle, Nobel Lecture (2001)
  • “Imagine how many aspects of nature we would miss if we lived on the surface of the sun…without refrigerators.”

• If we can create conditions that haven’t been created before, then we’ll make new discoveries.
• If we look at things in a way that hasn’t been done before, then we’ll see new things.
• If we analyze data using new methods, then we’ll gain new insights.
FUNCTIONAL DATA ANALYSIS

• Very often, data will be “telemetric” in nature - many repeated measures of several metrics through time.
  • This is true of data from many sources.
    • Machines output
    • Traditional time series data
    • Sensor data
    • Vibration signals
• A wide variety of specific tools and methods have been created to deal with this type of data:
  • Signal Processing
  • ARIMA Time Series
  • Partial Least Squares
  • Growth Curves via SEM
  • Mixed Models
FUNCTIONAL DATA ANALYSIS

- Many products are made in batches by machines that now have many sensors embedded in them
- Sensors record things like temperature, pressure, feed rate, chemical content (ammonia, CO2, ethanol, sugar), vibration, etc.
- Companies care about end results:
  - Yield: the quantity of product created (yield)
  - Quality: Measurable properties of the product (flavor, room temperature viscosity, shear strength, chemical composition)
- They want to understand how the sensor readings relate to the end results
  - To fix ‘bad batches’, or terminate their production early
  - Reduce occurrence of bad batches (process improvement)
- This is not a new problem - Due to the explosion of data access a lot more people want to take advantage of functional data
FUNCTIONAL DATA ANALYSIS

- Traditional approaches are too often inadequate and overcomplicated
  - Converting data to wide format (one input variable per time period) and using PLS.
    - Data cleaning step can be very time-consuming.
    - Sparse table if time-points not aligned, lost data if sizes not equal.
    - Difficult to interpret results for optimization (may be possible for early flagging of batches)
- Least Squares modeling of summary statistics (mean, min, max, etc.).
  - Too simple, all time-dependent information is lost.
  - Model says nothing about the shape of the functions.
- Fitting logistic curves, using parameter estimates as features.
  - Very limited in the set of shapes of curves that can be fit.
  - More flexibility needed than simple logistic curves and Gaussian peak models (again, too simple).
DEALING WITH TELEMETRIC DATA (TRADITIONAL APPROACH)

Goldrick et al.: MVDA of Trisulfide Bond Formation, Biotechnology and Bioengineering, Vol. 114, No. 10, October, 2017
USES OF FUNCTIONAL DATA ANALYSIS

1. Functional factors, constant responses.
2. Constant factors, functional responses.

Example: Maximizing the yield of human insulin produced by modified yeast cells.

Ethanol ☐
Temperature ☐
Molasses Feed ☐
NH3 Feed ☐
Air ☐
Tank Level ☐
pH ☐

Yield (%)
There are 100 batches in total
100 time measurements per batch
Measurements were taken at fixed time intervals
This isn’t always the case!
INVESTIGATING INPUT FACTORS

Plotting the factors over time for each batch reveals the complexity of the problem.
INVESTIGATING INPUT FACTORS

Exploring batches individually further emphasizes the challenge ahead.
GOOD BATCHES VS BAD BATCHES

What we’ll find:
1) We can achieve a 74% Yield.
2) Ethanol, Molasses, NH3 and Air are significant.
3) These are their ideal profiles:

We could try to look for characteristics of good vs bad batches, but what defines “good?”

Can “good” be “better?”
DEALING WITH TELEMETRIC DATA (USING FDA)

Functional Data Analysis

Mean

Eigenfunction 1

Eigenfunction 2

Eigenfunction n
Functional Data Analysis modeling for determining Functional Principal Components to be used in prediction.
SAVING FPC AND EIGENFUNCTIONS FOR EACH BATCH

- We now have 100 functional summaries (one for each batch)
MODELING WITH FPCS AS INPUTS

Parameter Estimates for Original Predictors

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>t-Value</th>
<th>Prob &gt;</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.53436</td>
<td>0.0020203</td>
<td>266.0</td>
<td>&lt;.0001*</td>
<td>0.530403</td>
<td>0.5383197</td>
</tr>
<tr>
<td>NH3 Feed FPC 2</td>
<td>-0.000173</td>
<td>2.4377e-5</td>
<td>&lt;.0001*</td>
<td>-0.00352</td>
<td>-0.000258</td>
<td>-0.000126</td>
</tr>
<tr>
<td>Ethanol FPC 2</td>
<td>-0.012168</td>
<td>0.0018192</td>
<td>&lt;.0001*</td>
<td>-0.03173</td>
<td>-0.02582</td>
<td>-0.00802</td>
</tr>
<tr>
<td>Molasses Feed FPC 2</td>
<td>0.0000244</td>
<td>4.0897e-6</td>
<td>&lt;.0001*</td>
<td>0.00248</td>
<td>1.6375e-5</td>
<td>0.000204</td>
</tr>
<tr>
<td>Air FPC 1</td>
<td>-2.63e-6</td>
<td>1.1457e-6</td>
<td>&lt;.0001*</td>
<td>-4.875e-6</td>
<td>-3.84e-7</td>
<td></td>
</tr>
<tr>
<td>Ethanol FPC 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ethanol FPC 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Temp FPC 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Temp FPC 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Temp FPC 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Molasses Feed FPC 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Molasses Feed FPC 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NH3 Feed FPC 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NH3 Feed FPC 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Significant factors

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Wald ChiSquare</th>
<th>Prob &gt; ChiSquare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.53436</td>
<td>0.0020203</td>
<td>69956999</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>NH3 Feed FPC 2</td>
<td>-0.000173</td>
<td>2.4377e-5</td>
<td>50.628148</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Ethanol FPC 2</td>
<td>-0.012168</td>
<td>0.0018192</td>
<td>44.736472</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Molasses Feed FPC 2</td>
<td>0.0000244</td>
<td>4.0897e-6</td>
<td>35.568946</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Air FPC 1</td>
<td>-2.63e-6</td>
<td>1.1457e-6</td>
<td>5.2684672</td>
<td>0.0217*</td>
</tr>
</tbody>
</table>

RSquare: 0.733979
RSquare Adj: 0.731265
Root Mean Square Error: 0.020353
Mean of Response: 0.53436
Observations (or Sum Wgts): 100
# Interpretation of Results

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Wald ChiSquare</th>
<th>Prob &gt; ChiSquare</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.53436</td>
<td>0.002003</td>
<td>69956.999</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>NH3 Feed FPC 2</td>
<td>-0.000173</td>
<td>2.4377e-5</td>
<td>50.628148</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Ethanol FPC 2</td>
<td>-0.012168</td>
<td>0.0018192</td>
<td>44.736472</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Molasses Feed FPC 2</td>
<td>0.0000244</td>
<td>4.0897e-6</td>
<td>35.568946</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Air FPC 1</td>
<td>-2.63e-6</td>
<td>1.1457e-6</td>
<td>5.2684672</td>
<td>0.0217*</td>
</tr>
</tbody>
</table>

![Graphs showing trends over time and eigenfunctions](image-url)
INTERPRETATION OF ONE FACTOR (ETHANOL)

Predictive formula for Ethanol, from FDA B-spline fitting

Ethanol FPC 1
+ Ethanol FPC 2
+ Ethanol FPC 3
+ Ethanol Mean Formula

These are functions of time, which can be consolidated.

The Prediction Profiler is used to visualize and explore predictive models.
Optimizing the values of the FPCs to maximize Yield, we have everything needed to find the optimal profiles.

These profiles are predicted to result in a Yield of 74%.
SUMMARY

USING FUNCTIONAL DATA ANALYSIS FOR TIME-DEPENDENT OPTIMIZATION OF BATCH PROCESSES

• Using Functional Data Analysis over traditional methods, we can:
  • Dramatically reduce total time to meaningful results,
  • Utilize all data while preserving time-dependent info,
  • Generate more interpretable knowledge output,
  • Better engage with subject-matter experts.

• Question: Can you think of opportunities within your organization where you could apply this method?

Thanks for the support of my colleagues Phil Kay, Chris Gotwalt, and Malcolm Moore