Modeling Response Curves

Silvio Miccio

Trainer and Consultant for Empirical Modelling and Optimisation

Procter & Gamble

About me

- Silvio Miccio
- Digitization, Empirical Modeling & Optimization
- Started with JMP 3
- Areas of interest:
  - DOE
  - Data Mining/Predictive Analytics
  - Generalized Linear Mixed Models
  - Multivariate methods PLS and SPC
  - Time Series Analysis
  - Data Management

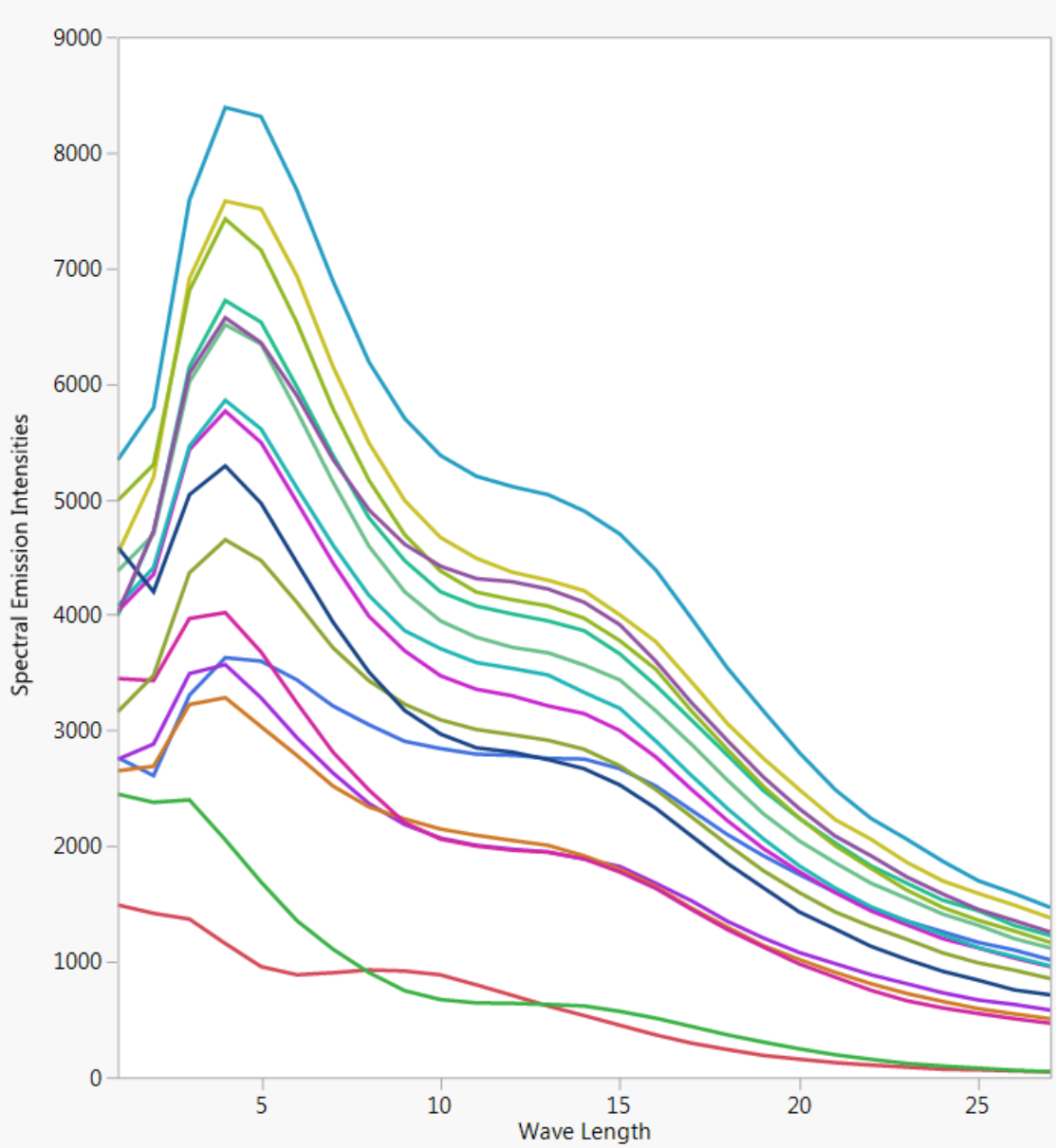
Introduction

- Not all results are based on a few continuous or discrete responses
- Sometimes the response is a curve e.g. describing the tensile strength of a product or spectra providing insights on chemical composition or quantities
- Instead of just modelling specific points of these curves it is possible to model the entire curves/spectra, enabling a more comprehensive understanding of the system
- Please notice, although not covered in this presentation, pre-processing of spectra like data (e.g. base line correction, smoothing, derivatives, signal correction, wavelet transformation…), is essential

Baltic Sea Example

- This example is from spectrometric calibration, which is an area where partial least squares is very effective. Suppose you are researching pollution in the Baltic Sea. You would like to use the spectra of samples of sea water to determine the amounts of three compounds that are present in these samples.
- The three compounds of interest are:
  - lignin sulfonate (ls), which is pulp industry pollution
  - humic acid (ha), which is a natural forest product
  - an optical whitener from detergent (dt)
- The amounts of these compounds in each of the samples are the responses. The predictors are spectral emission intensities measured at a range of wavelengths (v1–v27).
- For the purposes of calibrating the model, samples with known compositions are used. The calibration data consist of 16 samples of known concentrations of lignin sulfonate, humic acid, and detergent. Emission intensities are recorded at 27 equidistant wavelengths. Use the Partial Least Squares platform to build a model for predicting the amount of the compounds from the spectral emission intensities

(taken from JMP Help)



Multivariate Projections Methods

- Multivariate projection methods like PLS or PCA often provide good solution for modeling response curves
- PCA
  - Compress information from response curves (dimension reduction)
  - Helps to identify clusters
  - Centering (useful when being focused on differences in the samples rather than in the absolute value of the signal)
- PLS (will be demonstrated)
  - Model spectra as function of variables
  - Model responses based on spectra

PLS (Projection to Latent Structures or Partial Least Square Regression)

- PLS is “similar” to principal component regression (PCR)
- But unlike PCR, PLS extracts so called latent factors from the X matrix in a way that the covariance between the X and the Y matrix is maximized
- Since this factors are determined based on their relevance to Y, PLS usually needs less factors than PCR for getting a model with same predictive power
- PLS uses the correlation structure in X and Y to model the response, therefore it is especially efficient for modeling systems with highly correlated responses like response curves
- For more information on PLS please see:
  - Cox, Ian and Gaudard, Marie. 2013.Discorvering Partial Least Square with JMP. Cary, NC: SAS Institute Inc. (excellent book, highly recommended)
  - PLS-regression: a basic tool of chemometrics; Chemometrics and Intelligent Laboratory Systems 58 Ž2001. 109–130

Functional Data Analysis

- **Functional data analysis** (FDA) is a branch of statistics that analyses data providing information about curves, surfaces or anything else varying over a continuum. In its most general form, under an FDA framework each sample element is considered to be a function. The physical continuum over which these functions are defined is often time, but may also be spatial location, wavelength, probability, etc. (taken from Wikipedia)

Possible Approach (will be demonstrated)

- Reduce dimensions of response curves as parameters of non-linear functions (by sample/experimental condition)
- Model the parameters of the non-linear function (obtained per sample/experimental condition) via regression

Something More on Functional Data Analysis

- Response curves have to be “smooth” or can be smoothed (signal pre-processing can be as simple as just averaging the samples)
- Usually based on non-linear/complex functions
- Measurements do not need to be taken at the same point in time, wavelength….
- For more information
  - Slides: Functional Data Analysis, A Short Course, Giles Hooker, International Workshop on Statistical Modeling, Glasgow University, July 4, 2010
  - Books: e.g. Ramsay & Silverman, “Functional Data Analysis”

Case Study 1 – Modeling a Force Curve via PLS

- Model the effect of raw material, equipment and process set-up on the force required to push a product through a packing machine
- Force over the entire distance recorded
- Objective: Identify system set-up for minimizing the force
- JMP Tools
  - Overlay Plot
  - Validation Column
  - PLS
  - Spectral Profiler
  - Model Comparison

Case Study 2 – Modeling a Response Curve via FDA

- Investigate effect of product formulation and process set-up on product performance based on a Designed Experiment (combination of mixture components and process settings)
- Response is not constant over time, but describes a smooth curve
- Objective is overall system optimization
- JMP Tools
  - Overlay Plot
  - Non-linear Fit
  - PLS (other regression methods also work, but in this particular case best option is again PLS)
  - Make a constrained Spectral Profiler