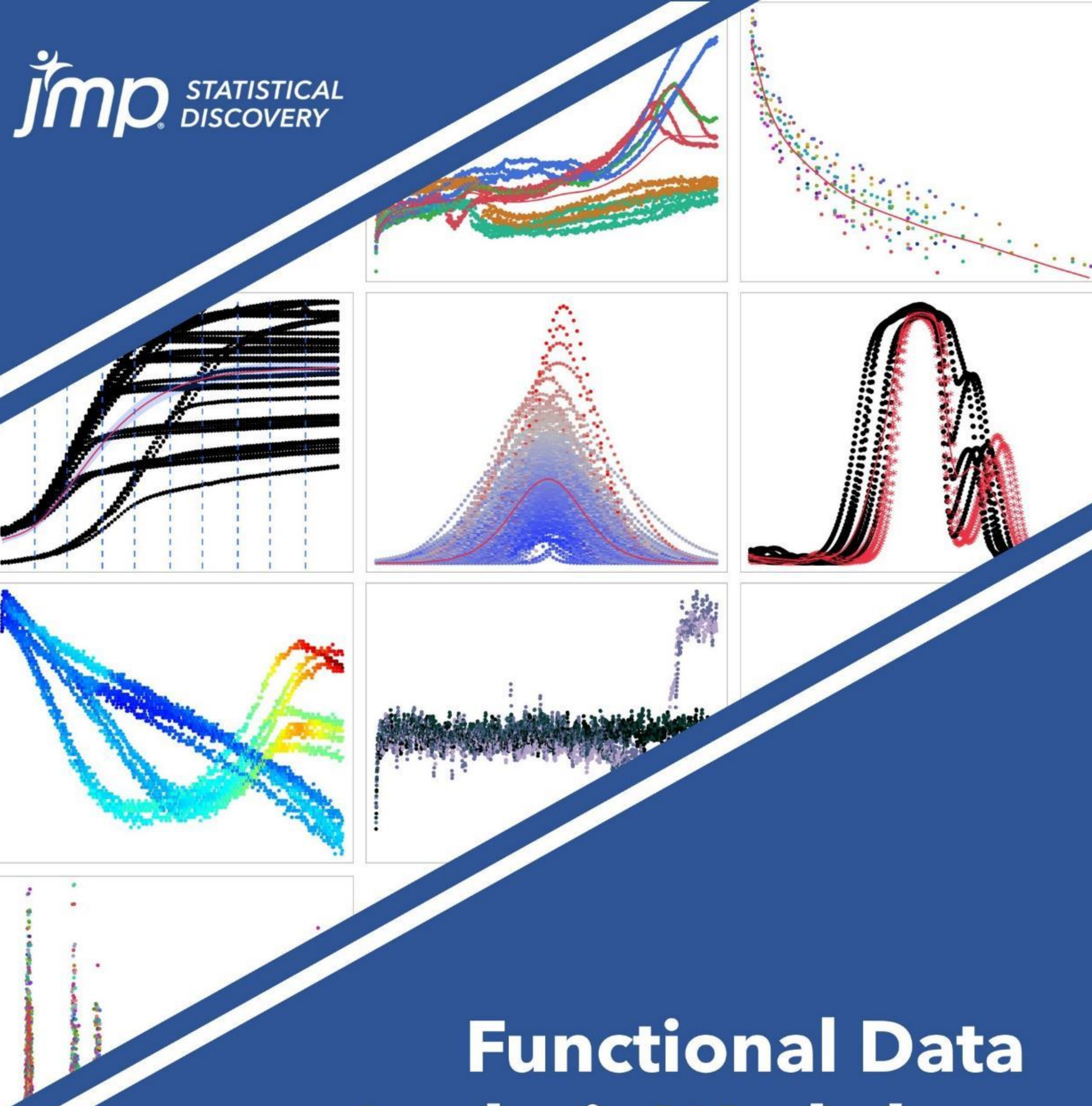
jmp STATISTICAL DISCOVERY



Functional Data Analysis Workshop

Chris Gotwalt

JMP Chief Data Scientist

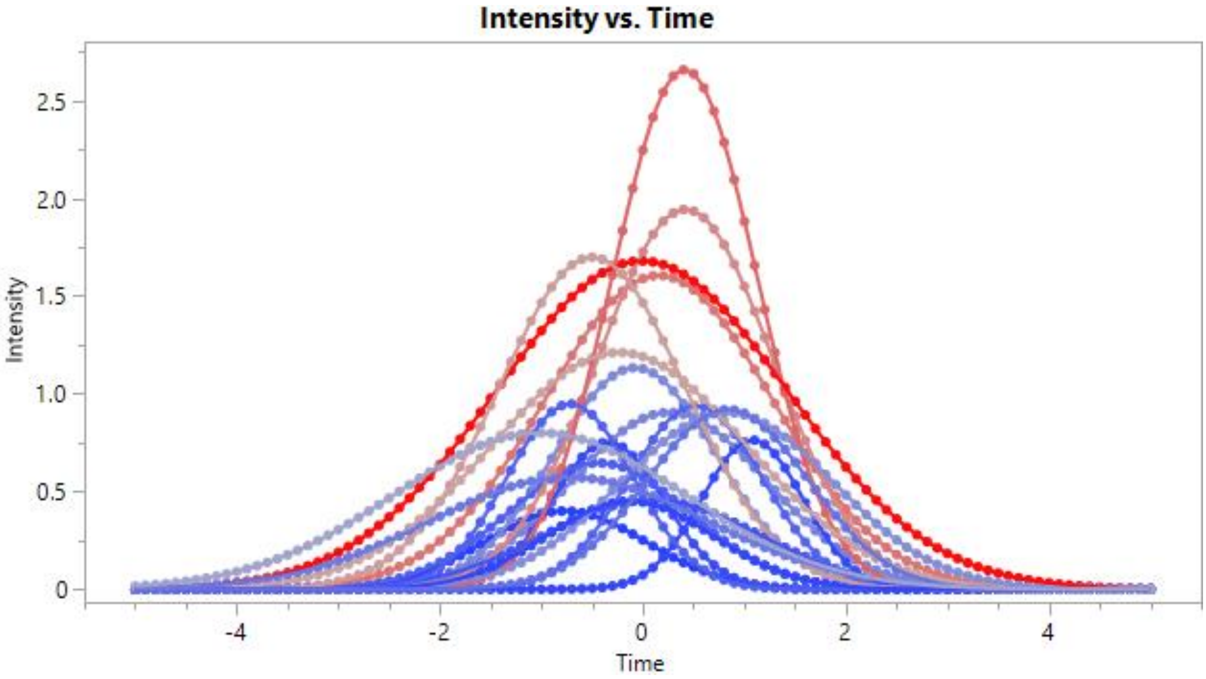
3rd March 2023

Case Study Exercise #1

A simple example of using functional data

Background

This simulated example shows a set of peak curves of varying intensities and width. We want to find out what these peaks have in common and where they differ – identifying the key components of their shape and distilling them into single parameters that we can use to compare, analyse and recreate.



Goals

This is an example to familiarise you with the Functional Data Explorer platform. You will:

- Convert discrete data into continuous functions.
- Extract characteristic shape components with Functional Principal Component (FPCA) analysis.
- Identify and analyse the differences between the curves.

Data

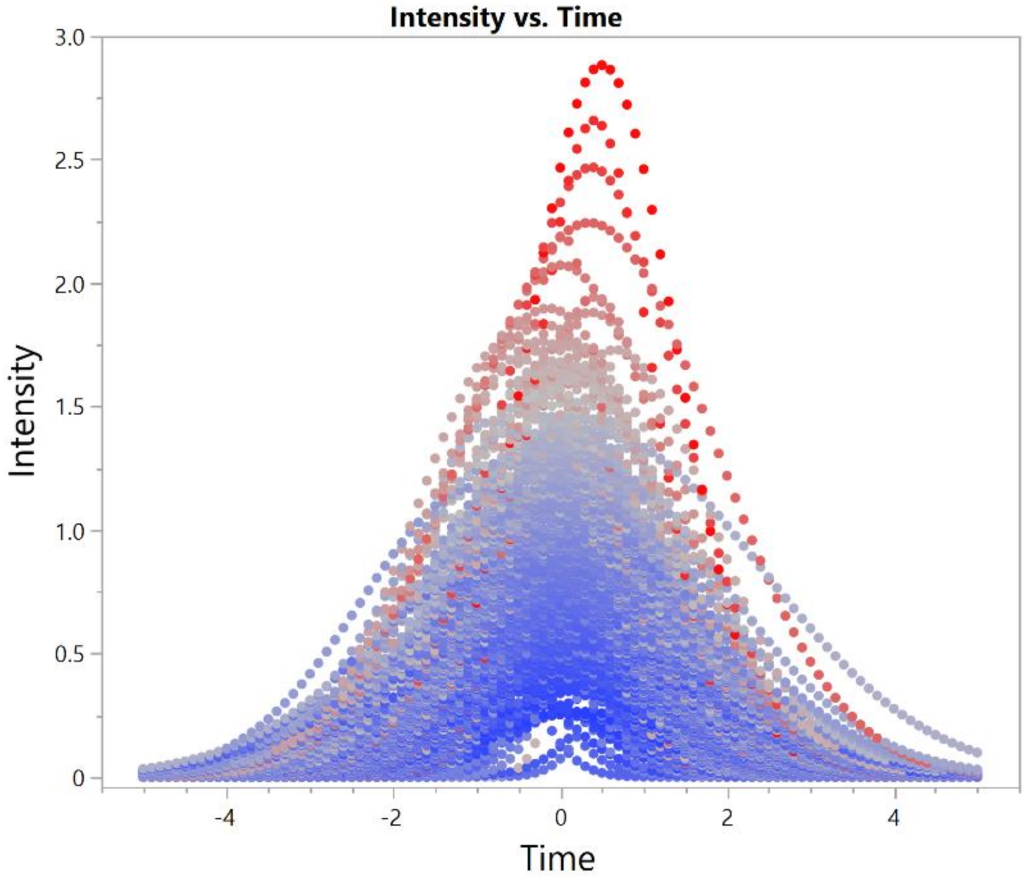
Open 'Peaked Data FDE.jmp' to access the dataset.

This is a simulated data set for single peaks of varying position and intensity. The data table contains:

- The ID for each peak (Batch)
- The measured value (Intensity) over the interval (Time [mins]).

Analysis

Graph Building



Use Graph Builder to drag and drop variables and plot the data for each batch.

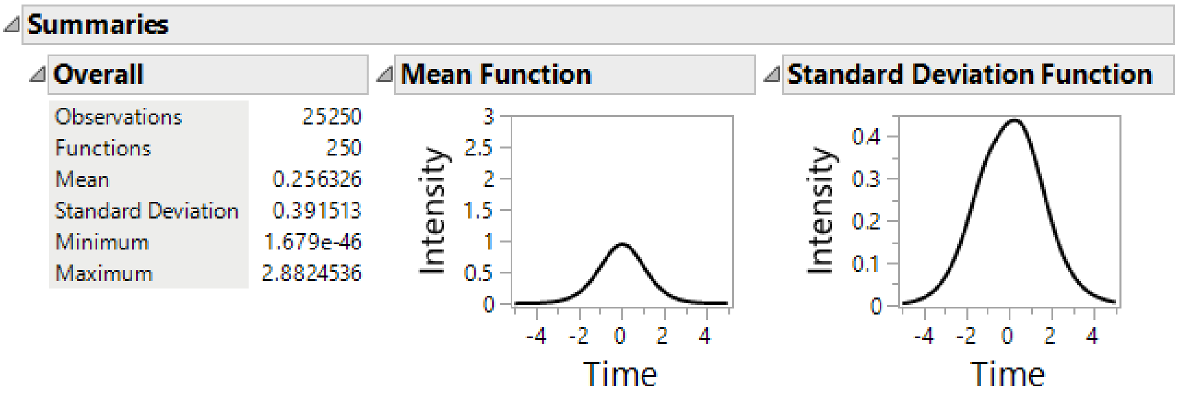
1. Graph > Graph Builder. Y = Intensity, X = Time and click Done.

Here you can see there are >250 individual peak shapes to assess. We can see some differences from a glance. For example, in the width of peaks and the time that they peak. However, the data set is large and it quickly becomes difficult to discern the individual differences between them by eye. We can apply functional data analysis to assess the differences between these curves.

Launching FDE and data processing

2. Select Analyze > Specialized Modelling > Functional Data Explorer.
3. Select Intensity and click Y, Output.
4. Select Time and click X, Input.
5. Select Batch and click ID, Function.
6. Click OK.

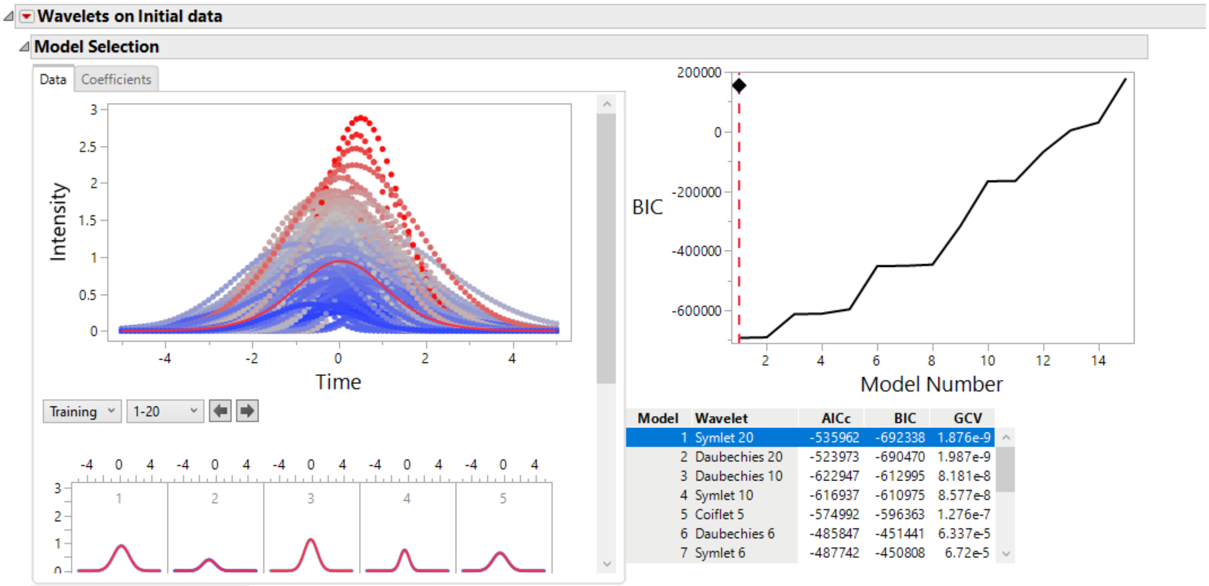
You are first given visuals like those we created in Graph Builder. You also see the Mean "trend" and how standard deviation varies with Time.



Overall	
Observations	25250
Functions	250
Mean	0.256326
Standard Deviation	0.391513
Minimum	1.679e-46
Maximum	2.8824536

There are also options to pre-process your data, available in the red triangle for 'Data Processing' or buttons under the 'Commands' section. From here you can clean up (filter values), transform and align your data sets according to your needs. In this example this is not required.

Fitting the smoothing model

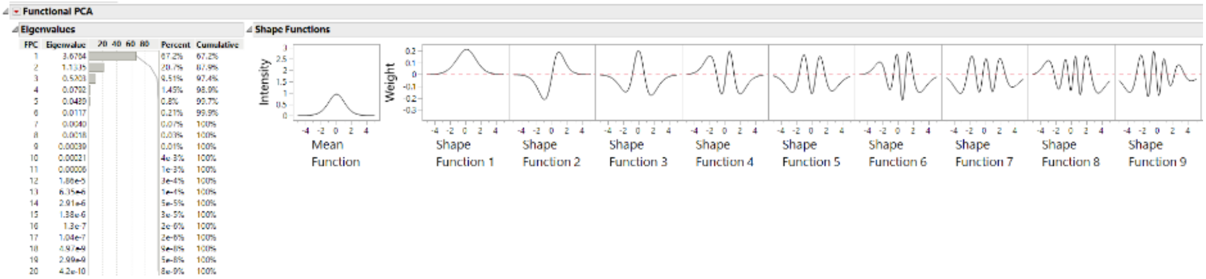


Model	Wavelet	AICc	BIC	GCV
1	Symlet 20	-535962	-692338	1.876e-9
2	Daubechies 20	-523973	-690470	1.987e-9
3	Daubechies 10	-622947	-612995	8.181e-8
4	Symlet 10	-616937	-610975	8.577e-8
5	Coiflet 5	-574992	-596363	1.276e-7
6	Daubechies 6	-485847	-451441	6.337e-5
7	Symlet 6	-487742	-450808	6.72e-5

7. Click the Functional Data Explorer red triangle and select Models > Wavelets

This uses a Wavelet model that applies a smoothing fit to Intensity vs. Time for each Batch, taking the discrete numeric values and converting them to a continuous curve. FDE has automatically selected the Wavelet model with the lowest Bayesian Information Criterion (BIC) – in this case a Symlet 20 wavelet model.

8. Look at the effect of changing the wavelet model type by dragging the slider from left to right on 'Model Number', looking at the fit in the 'Model Selection' graphs. Return to Model 1 before proceeding.



Functional PCA

Eigenvalues

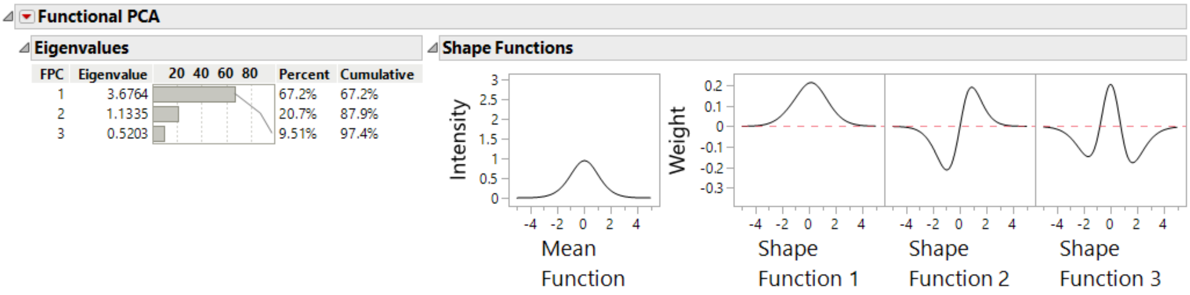
FPC	Eigenvalue	20 40 60 80	Percent	Cumulative
1	3.6764		67.2%	67.2%
2	1.1335		20.7%	87.9%
3	0.5203		9.51%	97.4%
4	0.0792		1.45%	98.9%
5	0.0439		0.8%	99.7%
6	0.0117		0.21%	99.9%
7	0.0040		0.07%	100%
8	0.0018		0.03%	100%
9	0.00039		0.01%	100%
10	0.00021		4e-3%	100%
11	0.00006		1e-3%	100%
12	1.86e-5		3e-4%	100%
13	6.35e-6		1e-4%	100%
14	2.91e-6		5e-5%	100%
15	1.38e-6		3e-5%	100%
16	1.3e-7		2e-6%	100%
17	1.04e-7		2e-6%	100%
18	4.97e-9		9e-8%	100%
19	2.99e-9		5e-8%	100%
20	4.2e-10		8e-9%	100%

Looking at the functional components

Immediately after applying the smoothing model, JMP Pro decomposes the signals into the dominant characteristic shapes it found in the data with Functional Principal Component Analysis. In mathematical language these shapes are called eigenfunctions, but a better and more approachable name would be to call them shape functions. Here we see that JMP found that there is a peaked mean function and 20 "shape components" that explain of the batch-to-batch variation in the data. The shape components are used in combination with the scores or 'weights' that are unique for each batch, representing the amount of each shape component that is added to the mean to recreate a given batch's peak shape.

For the Wavelet model, 20 shape components account for 100% of the variations in the Intensity trends between the batches, with the last 16 shape components accounting for only 1.1% of the variation.

9. See what happens as you decrease the number of shape components by dragging the slider from right to left in the 'FPCA Model selection'. Look at the fit on the 'FPCA Diagnostic Plots'.
10. Select 3 shape components.



Functional PCA

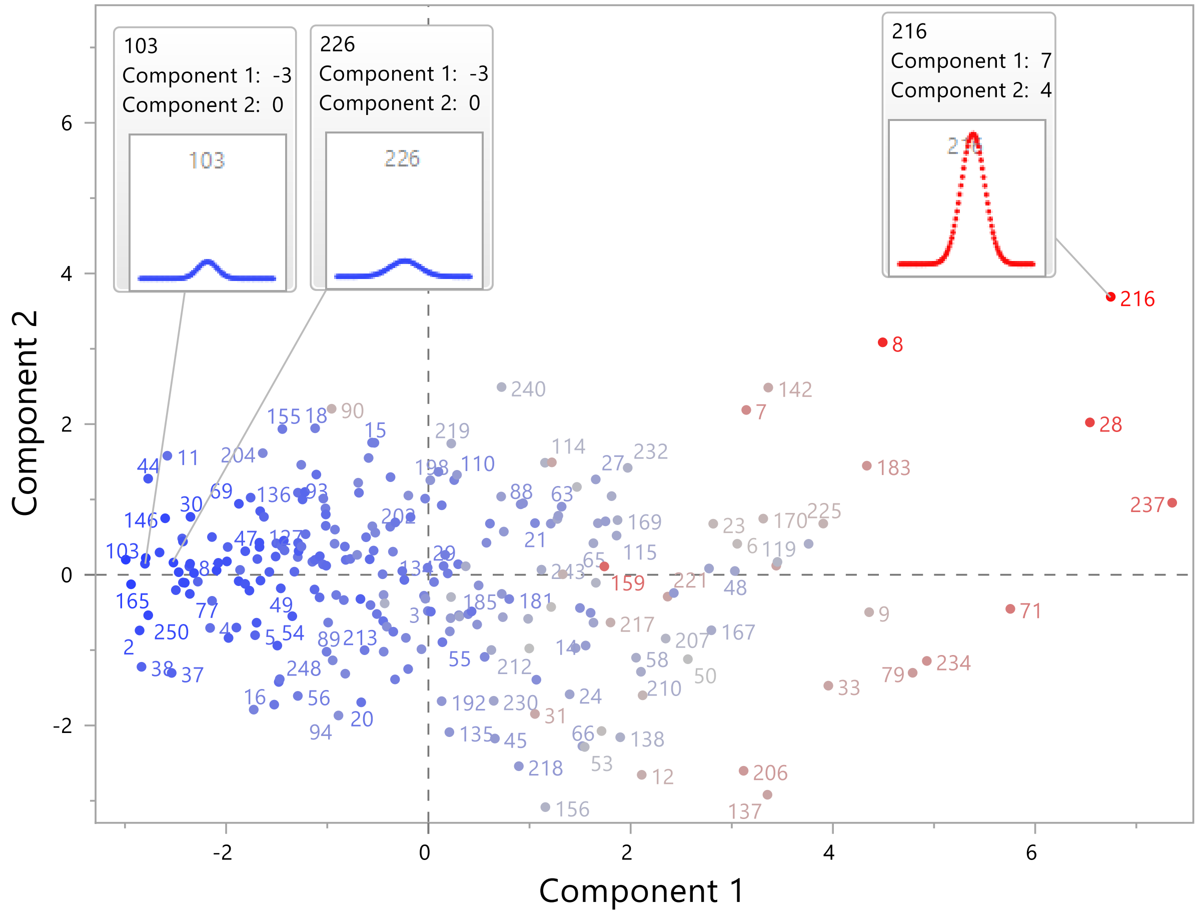
Eigenvalues

FPC	Eigenvalue	20 40 60 80	Percent	Cumulative
1	3.6764		67.2%	67.2%
2	1.1335		20.7%	87.9%
3	0.5203		9.51%	97.4%

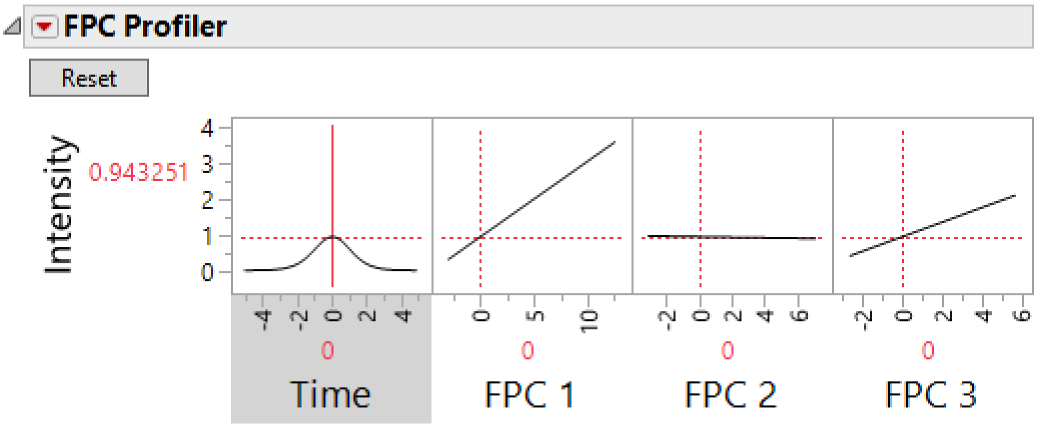
The selected shape components represent the dominant characteristics that account for 97.4% of the variation in the Intensity trends between batches. In this example the shape components represent the change in peak height (component 1), peak shift (component 2) and peak width (component 3). It is important to remember that these 3 "shape components" were determined from the data and this example was simulated so that the 3 shape components are specific meaningful characteristics. Every dataset has its own shape components that reflect the characteristics of that data and they may not be as easy to understand.

You can re-create any of the original intensity trends of a given batch by taking the mean trend and adding different amounts of each of these shape components. The 'Score Plot' allows you to understand the differences between the batches.

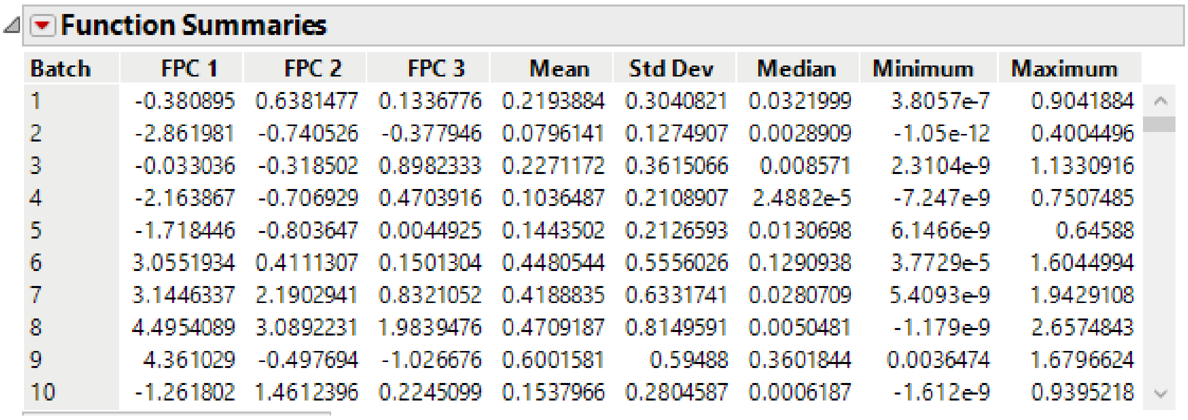
11. In the Score Plot hover over the data points and you will see graphlets of the peak for that batch.



The score plot shows where batches have similarities in shape (similar shape component scores) and where they differ. In the example above, Batch 103 and Batch 226 (left) have similar values of Shape components 1 and 2. We can see that they have similar looking peaks. Batch 216 (right) has a much higher score for component 1, corresponding to a taller peak. It also has a higher score for component 2, which corresponds to a peak shift (harder to see). The score plot can help you to quickly find different groupings and to understand the nature of these differences. You can explore how the scores for each shape component relates to the Intensity peak shape using the FPC Profiler. This profiler allows you to change the score or 'weight' of each shape component that is added to the mean function to produce the final curve. If all the weights are zero, the curve is just equal to the overall mean curve. If we vary the weights, we see how that affects the shape of the curve.



12. Adjust each shape component score in the FPC profiler and see how the resultant curve changes. Remember that the components in this example represent the peak height (component 1), peak shift (component 2) and peak width (component 3).
13. Save the Shape component/FPC scores: from the red triangle next to Function Summaries > Save Summaries.


Function Summaries

Batch	FPC 1	FPC 2	FPC 3	Mean	Std Dev	Median	Minimum	Maximum
1	-0.380895	0.6381477	0.1336776	0.2193884	0.3040821	0.0321999	3.8057e-7	0.9041884
2	-2.861981	-0.740526	-0.377946	0.0796141	0.1274907	0.0028909	-1.05e-12	0.4004496
3	-0.033036	-0.318502	0.8982333	0.2271172	0.3615066	0.008571	2.3104e-9	1.1330916
4	-2.163867	-0.706929	0.4703916	0.1036487	0.2108907	2.4882e-5	-7.247e-9	0.7507485
5	-1.718446	-0.803647	0.0044925	0.1443502	0.2126593	0.0130698	6.1466e-9	0.64588
6	3.0551934	0.4111307	0.1501304	0.4480544	0.5556026	0.1290938	3.7729e-5	1.6044994
7	3.1446337	2.1902941	0.8321052	0.4188835	0.6331741	0.0280709	5.4093e-9	1.9429108
8	4.4954089	3.0892231	1.9839476	0.4709187	0.8149591	0.0050481	-1.179e-9	2.6574843
9	4.361029	-0.497694	-1.026676	0.6001581	0.59488	0.3601844	0.0036474	1.6796624
10	-1.261802	1.4612396	0.2245099	0.1537966	0.2804587	0.0006187	-1.612e-9	0.9395218

This gives you a table with one row per batch. The characteristic shape of the intensity trend is described in that one row for each batch.

You have seen how you can comprehensively describe semi-continuous profiles, spectra, curves or time-courses with functional summaries. In later exercises you will use these functional summaries to model and understand processes.

Optional: Looking at different models

Wavelet models demonstrate wide applicability to functional data and is the best choice in most situations (followed by B-spline, P-spline and then Fourier Basis). In some cases other fitting models may better suit your data – you can compare how they fit through diagnostic plots.

1. Click the Functional Data Explorer red triangle and select Models > B-Splines.
2. Under 'Functional PCA', open the 'FPCA Diagnostic Plots' window. Compare the 'Actual by Predicted' and 'Residual by Predicted' plots between the Wavelet and B-spline data.

Whilst there are only 5 shape components, the predicted values produced do not fit as closely to the actual values of the batches. In this case the Wavelet model looks like the most useful. However, you should always use your domain knowledge to ensure that the models are fitting closely to your expectations. FDE makes it simple and quick to explore the different possibilities.

Case Study Exercise #2

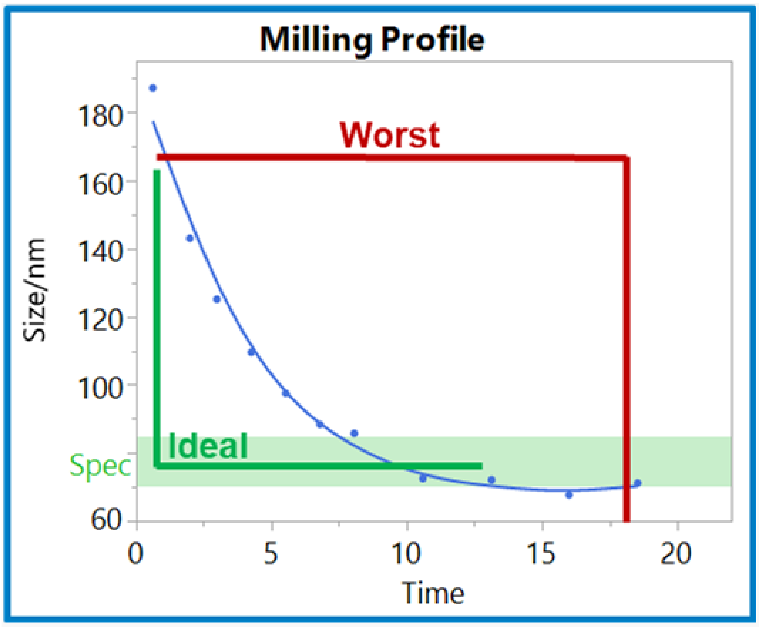
Building a model of a functional response with the factors in a designed experiment

Background

You are developing a process for milling an active ingredient for formulation (pharma, agrichem, printing …). The particle size of the active needs to be within a tight spec (70-85 nanometres) for application.

The process uses bead milling. Dispersion of the active is recirculated around a bead mill. The dispersion is pumped through the mill chamber containing ceramic beads agitated at very high speeds. The particles of the active are ground down by collisions with beads and with each other.

Milling time costs money ($100s per hour) and is potentially a bottleneck in the process, affecting throughput. Sampling to measure particle size through the process also costs money ($100s per sample). Out of spec bxs would be very costly ($10,000s per bx).



The ideal "milling profile" would drop quickly to the middle of spec and remain there. This would be most efficient and robust. The worst profile would take a long time to get to spec but then go out of spec immediately. Feasible profiles will be less extreme.

Goals

This is an example to introduce the functional DOE tool. You will:

- Find a suitable curve model for the milling profiles.
- Perform Functional PCA to capture the characteristic shapes of the milling profiles.
- Identify the effects of the process inputs on the functional profile with Functional DOE.
- Find the optimum settings for the milling process.

Data

You can find the data in Help > Sample Data Folder > Functional Data. Open "Mill DOE.jmp".

This is a 17-run Definitive Screening Design (with data for some additional confirmation runs excluded). The data table contains:

- The factors for each run
- The measured responses at several times points per batch: #Oversize and Size/nm.

We will focus on Size/nm as the functional response of interest.

(You can run the "Original Data Table" script to explore the design.)

Analysis

Analysis to create a model of the functional response versus the DOE factors

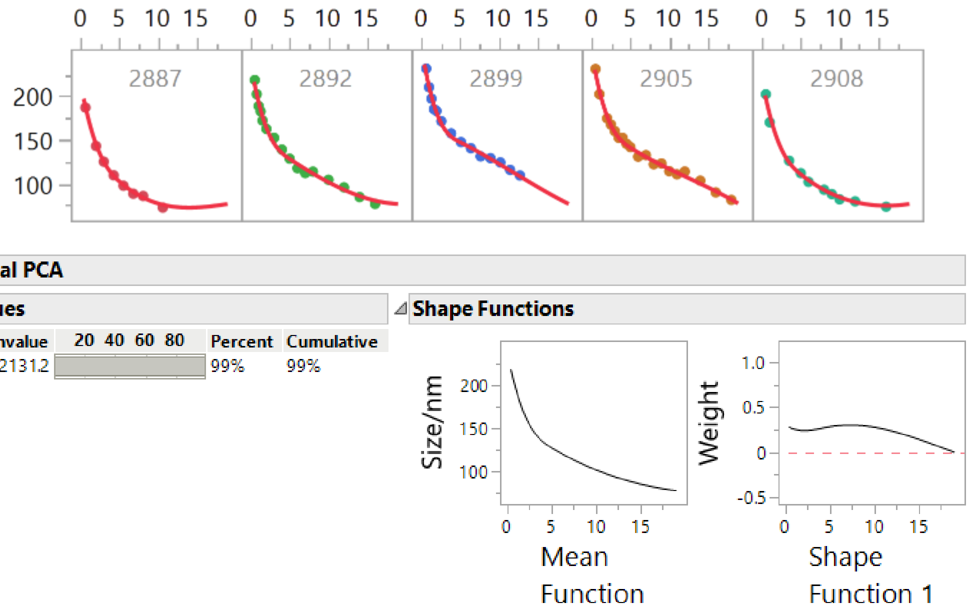
The first job is to turn the functional response - the milling profile - into a single row summary for each run. This is what FDE does.

1. Anaylze > Specialized Modelling > Functional Data Explorer.
2. Y = Size/nm, X = Time, ID = Batch
3. Put the DOE factors (%Beads, %Strength, Flow, T) into the Z, Supplementary role.
4. Click OK.

Now we fit the smoothing model that creates a continuous curve from our discrete time points:

5. From the red-triangle menu or "hotspot" at the top of the report > Models > B-splines.

We could explore different numbers of knots and degrees for the spline, and we might consider wavelet or other model types, but for this we will go with the suggested fit of 1 knots, cubic (degree = 3). The smoothing fits look good and we find that 1 Functional Principal Component captures 99% of the variation.



Functional PCA

Eigenvalues

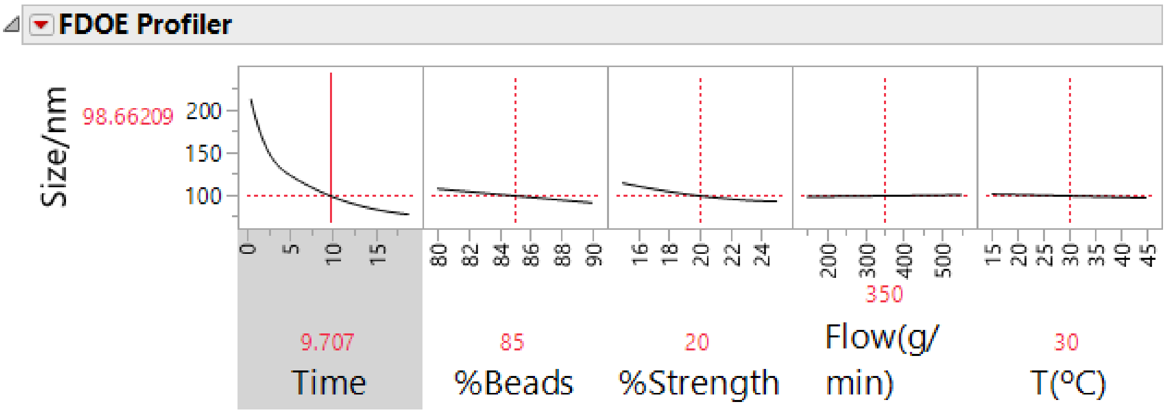
FPC	Eigenvalue	20 40 60 80	Percent	Cumulative
1	2131.2		99%	99%

Shape Functions

The Function Summaries section shows us the scores for shape component/FPC 1, as well as other summary statistics. We can use the FPC Profiler to see how the shape component score relates to the shape of the milling profile.

Now we can model these against our DoE factors – this is why we added them in the Z, Supplementary role.

6. From the red triangle menu for 'B-Spline on Initial data', select Functional DOE Analysis.
7. Scroll down to the bottom of the Functional Data Explorer window and look at the FDOE profiler.

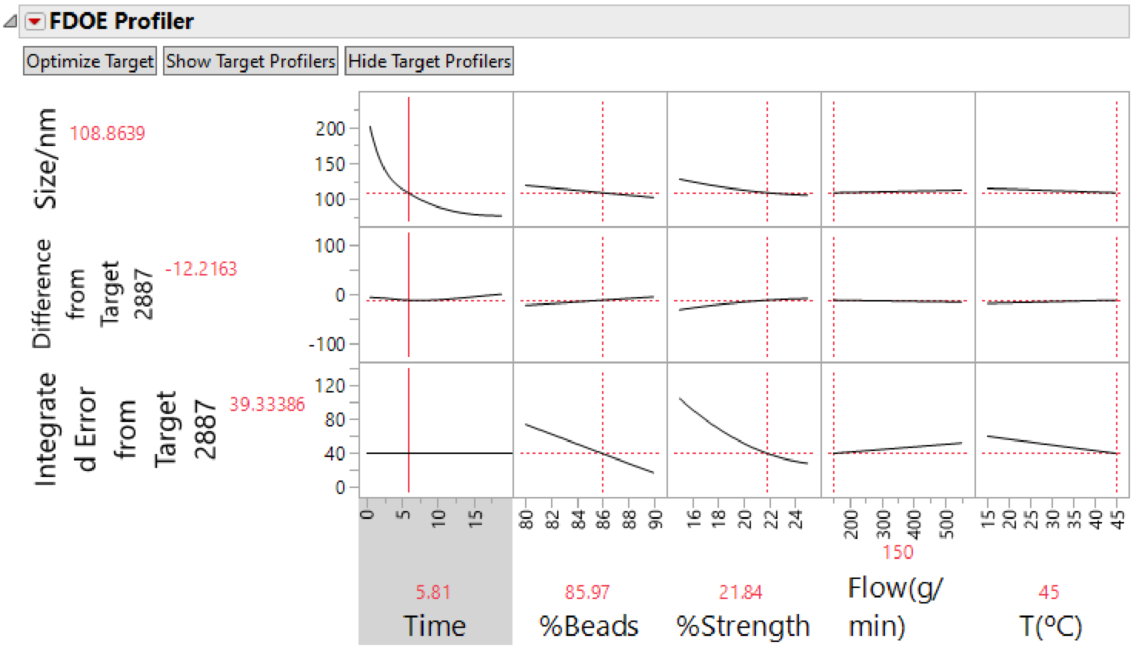


8. Try changing DOE factors and look how the milling profile (Size/nm versus Time) changes. Try to find settings of the factors that give the closest to the "ideal" profile.
9. Optional: If you are familiar with Generalized Regression in JMP Pro, look in the Generalized Regression for Size/nm shape component/FPC 1 outline and see how interacting with the solution path changes the model of the functional response.

Find settings for the "ideal profile" or "golden curve" using a target function

You can specify a target function so that you can optimise factor settings to give you the "ideal profile" or "golden curve." This target function needs to be specified in the data table. Here we will use one of the runs from the DoE (Batch 2887 is the closest to the ideal) but you can define a target profile in your data set that is not a run of your DOE.

10. Repeat steps 1 to 4, above.
11. Find Target Functions in the Data Processing red triangle > Target Functions > Load Targets or through Commands > Target Functions > Load
12. Select batch 2887 and click OK.
13. Now repeat steps 5 to 7, above.
14. From the FDOE Profiler select 'Optimize Target'. This will adjust the conditions to produce a curve that fits closest to the target function.
15. The differences between the predicted function and the target function can be shown with the Show/Hide Target Profilers buttons. Assess the difference of the current settings from the target function.



You have seen how you can optimise and control a process by modelling the functional response of a designed experiment.

Optional: Analysing Other Functional Response

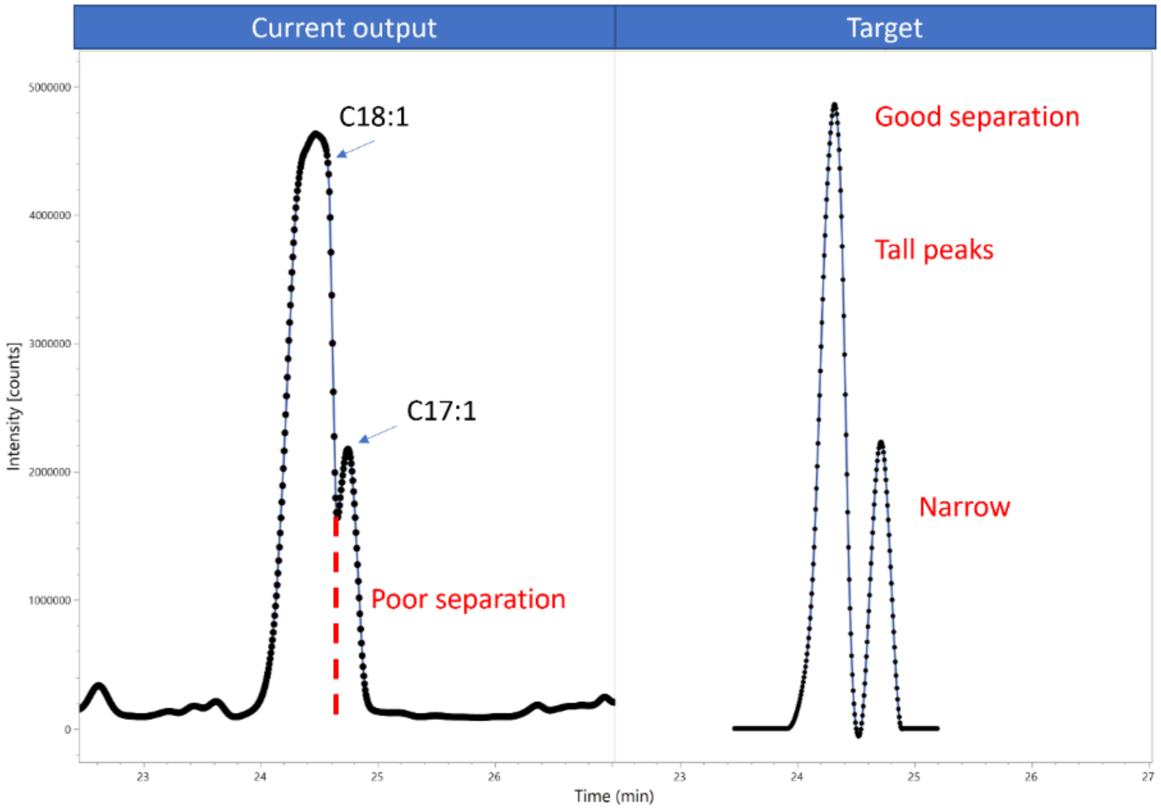
Repeat the analysis for the other functional response: #Oversize. This case may require more functional principal components to capture the variation in the functional response.

Case Study Exercise #3

A designed experiment for chromatography method development

Background

You are attempting to separate two closely related compounds, C18:1 and C17:1 sophorolipids, with high performance liquid chromatography (HPLC) in order to quantify them. The compounds are separated in an HPLC column before arriving at the detector and producing a chromatogram as shown below. To successfully quantify them, the two peaks must be fully separated whilst retaining their sharp, tall shape. Alteration of the HPLC settings can improve the separation.



To achieve this a simple design of experiments (DoE) was performed by changing the flow rate and temperature of the HPLC.

Goals

This is an example to familiarise yourself with the pre-processing of data and its application in optimisation. You will:

- Reduce large chromatographic data sets into the areas of interest.
- Perform data clean-up to reduce, filter and align the peaks
- Apply functional design of experiments (FDoE) to explore the data
- Apply target functions to optimize against.

Data

Open ‘HPLC FDE.jmp’

This is a 6 run screening design. The data table contains:

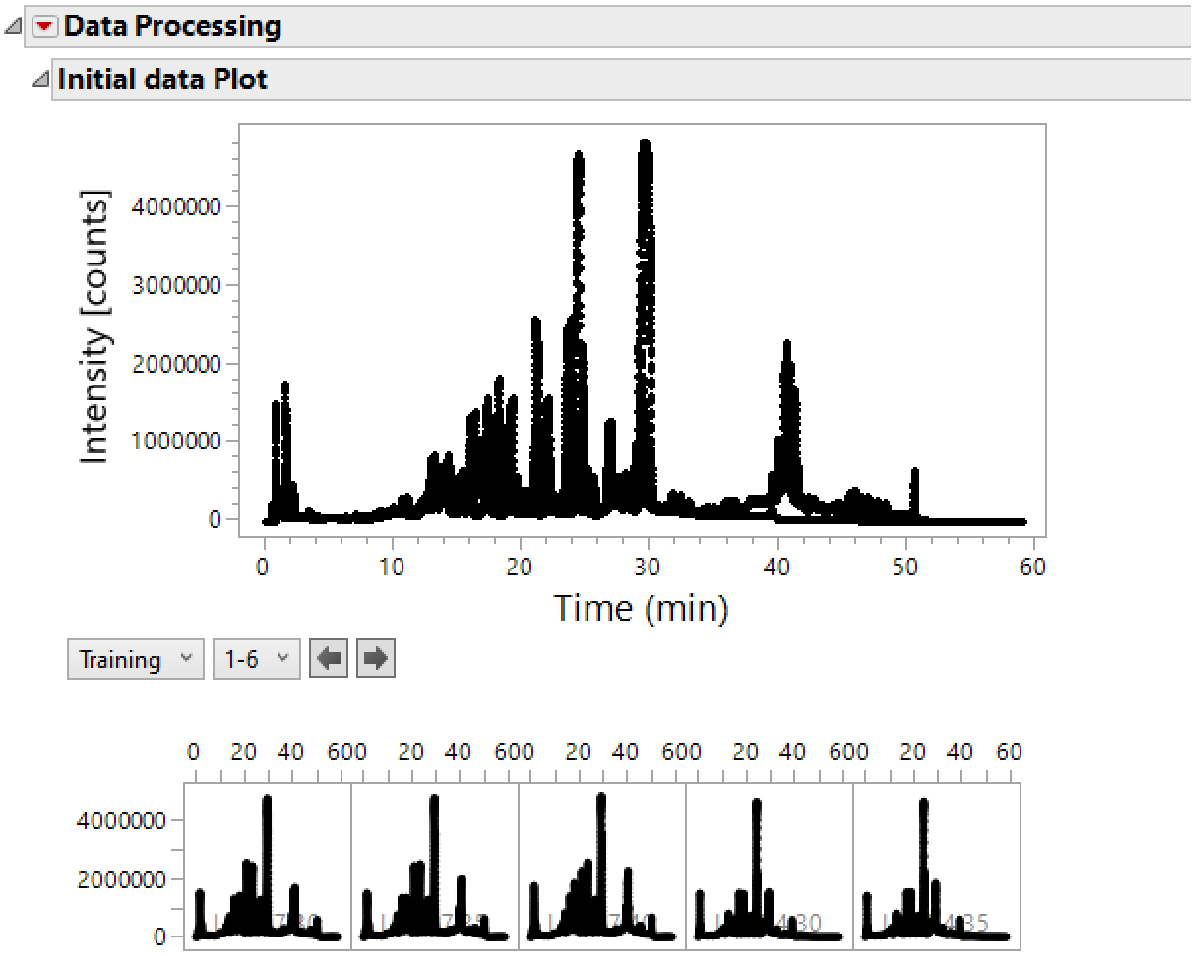
- The measured response (Intensity (counts)) over Time (mins)
- The factors - Flow rate [mL/min] and Temperature (°C)
- The ID of each run (Batch ID)

Process

Preparing the data set for modelling of the functional response

Firstly, define the variables roles in the launch dialog for the Functional Data Explorer platform.

1. Analyze>Specialized Modelling> Functional Data Explorer
2. Y = Intensity [counts], X = Time, ID = Batch ID
3. Put the DOE factors (flow rate [mL/min] and Temperature [°C]) into the Z, Supplementary role.
4. Click OK

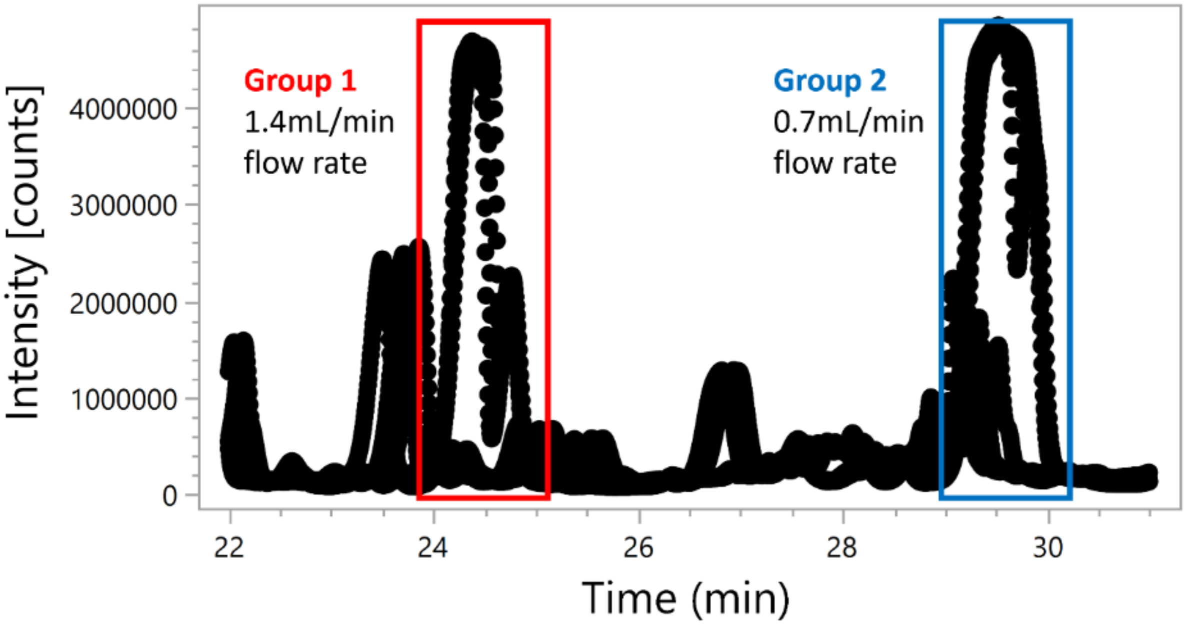


Data Processing

Initial data Plot

The discrete numeric data points are visualised in the Data Processing section, ready to have a continuous curve model applied. However, the whole 60 minutes of recorded data has been saved for each chromatogram. For our purposes we are only interested in analysing the change to the 2 sophorolipid peaks that occurs over a smaller period of time (approximately 22-31 minutes), so the time range needs to be reduced to the region of interest.

5. Under 'Commands', click 'Cleanup', Filter X, Below = 22, Above = 31



Now that we have the x-axis reduced we can see two distinct large peaks at ~24 and 29 minutes – these are the peaks of interest. However, they are separated at different times due to the change in flow rate. This separation makes it difficult to accurately model the changes between the two peaks, particularly as the chromatogram is very busy with other 'non-sophorolipid' peaks that are not of interest. Therefore we will align the main peaks and further reduce the time range to exclude other peaks.

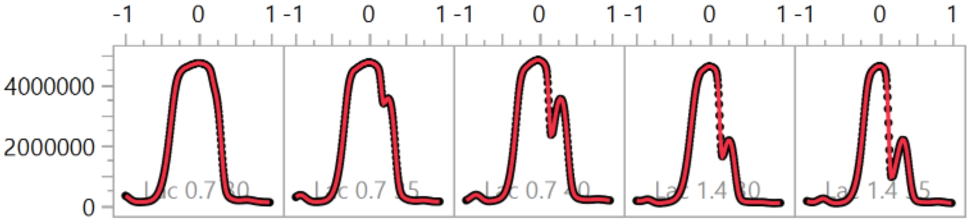
6. Under 'Commands', click 'Align', 'Align Maximum'
7. Reduce the range further to remove extra peaks with 'Commands'>'Cleanup'>'Filter X', Below = -1, Above = 1.

This data has now been processed ready for functional principal component analysis. Note that Functional Data Explorer includes other data processing tools, including methods for Spectral data processing.

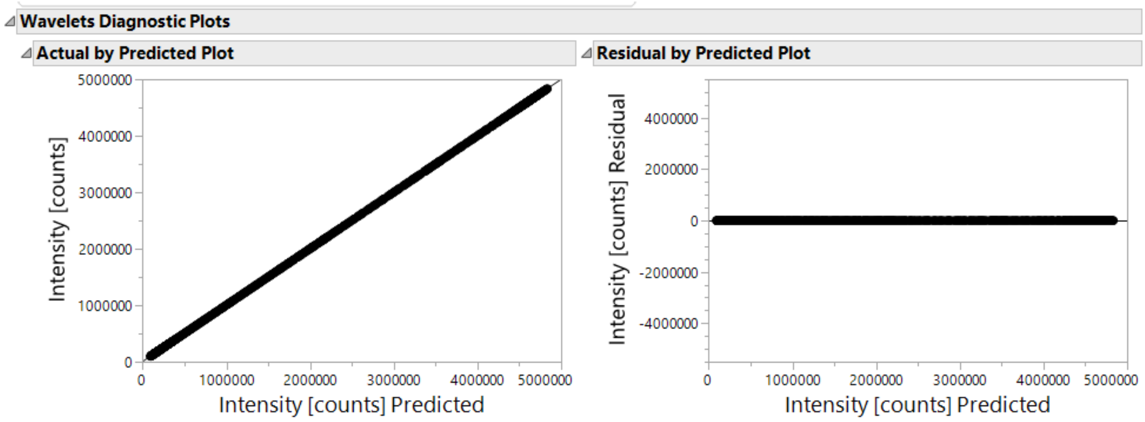
Applying the continuous curve model and checking goodness of fit

Now that the data has been pre-processed we can fit a smoothing model to create a continuous curve from our discrete values.

8. From the red-triangle menu or "hotspot" at the top of the report > Models > Wavelets



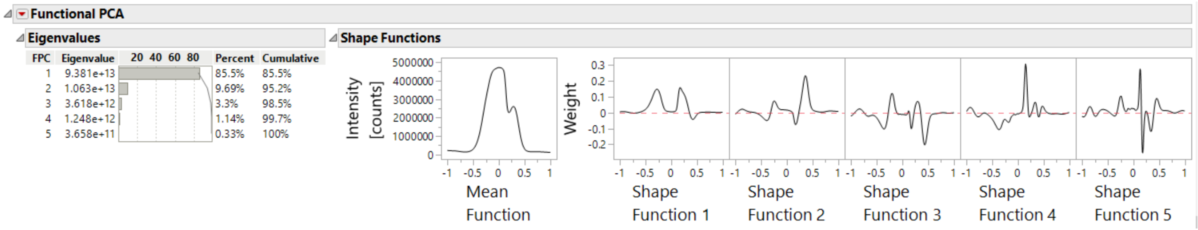
Wavelets are forms that oscillate from zero to a maximum before returning to zero, which is well suited to the chromatographic data collected here. The model appears to fit well to the data for each run and accounts for the specific peak shape of each chromatogram.


Wavelets Diagnostic Plots
Actual by Predicted Plot
Residual by Predicted Plot
Intensity [counts]
Intensity [counts] Predicted
Intensity [counts] Residual
5000000
4000000
3000000
2000000
1000000
0
-2000000
-4000000

For a further look at the model fit, the 'Wavelets Diagnostic Plots' outline provides the Actual by Predicted and Residual by Predicted plot for analysis.

9. *Optional*: You can have a look at the effect of model selection on the diagnostic plots by changing the model number in the 'Model Selection' section.

Analysis to create a model of the functional response versus the DOE factors



Eigenvalues

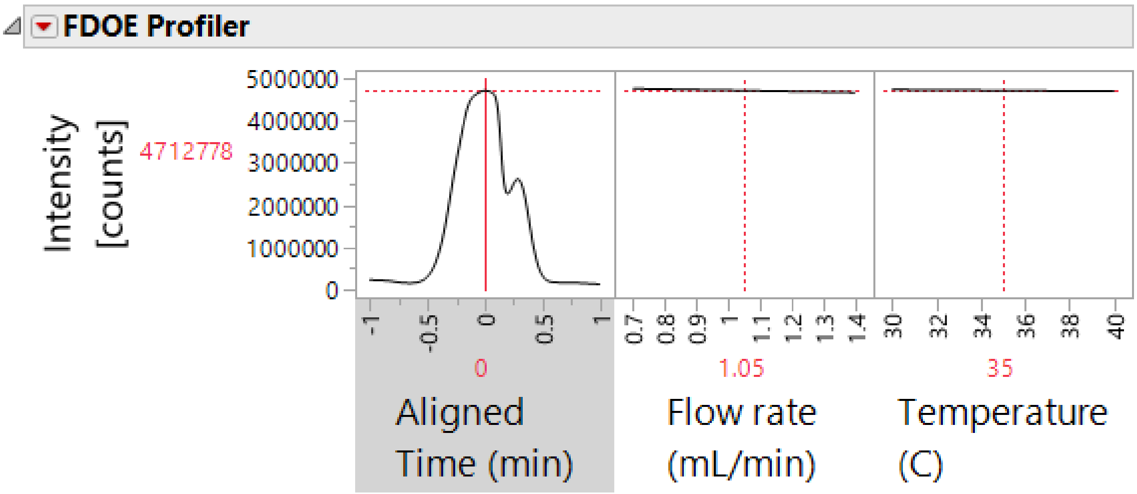
FPC	Eigenvalue	20 40 60 80	Percent	Cumulative
1	9.381e+13		85.5%	85.5%
2	1.063e+13		9.69%	95.2%
3	3.618e+12		3.3%	98.5%
4	1.248e+12		1.14%	99.7%
5	3.658e+11		0.33%	100%

The functional PCA has been applied to the data and 5 functional principal components have been selected to account for the variability in the data. More complex data sets will produce a more complex model with higher shape components, however care should be taken to avoid having a high number of shape components where little variation (i.e. <1%) is accounted for. For example, the 5th shape component (FPC 5) only accounts for 0.33% of the variation in the data. Is it important to retain this?

10. *Optional*: You can see the effect of each shape component on the shape under the 'FPC Profiler' and reduce/increase the number of shape components in the 'FPCA Model selection' tool.

Now we can model the shape components against our DoE factors to find optimum settings for separation.

11. From the red triangle menu for 'Wavelets on Reduced Grid', select Functional DOE Analysis.
12. Scroll down to the bottom of the Functional Data Explorer window and look at the FDOE profiler.



13. Try changing DOE factors and look how the chromatogram changes. Try to find settings of the factors that give the closest to the "ideal" profile shown at the top of the case study.
14. *Optional*: The fit of the DOE model can be explored with 'FDOE Diagnostic Plots' – as with the Wavelet model, this provides Actual vs. Predicted and Residual vs. Predicted plots.

Predicting optimum settings with a functional target

The functional data explorer platform provides a quick visual tool that allows for rapid analysis of the effects of settings on the functional data sets. For the HPLC case, there was a 'target' profile that was needed to properly separate the two compounds. This target function can be specified in the FDE platform to predict optimum settings. Guided by the functional DOE output, additional runs were performed at higher flow rate and temperature settings to see if separation could be further improved. With the new data and target function, we will repeat the Functional DOE analysis and find the optimum settings.

15. Open 'Additional Runs with Target' in the script menu of the main data table.
16. With the new data set, repeat steps 1-4.
17. Select the red triangle on 'Data Processing' and select 'Target Function' – find the run ID called 'Target' and select.
18. Repeat steps 5-12 to create the Functional DOE
19. Select 'Optimize Target' on the FDOE Profiler and see how the flow rate and temperature should be set to achieve the best peak separation. Optional: Select 'Show Target Profilers' to explore the difference from target and integrated error from target values.

You have seen how to optimise an analytical chemical method with a chromatographic response using functional data processing and Functional DOE.