JMP059: Functional Data Analysis for HPLC Optimization

Functional Data Analysis, Functional DOE

Produced by

Benjamin Ingham, The University of Manchester
benjamin.ingham@manchester.ac.uk

Functional Data Analysis for HPLC Optimization

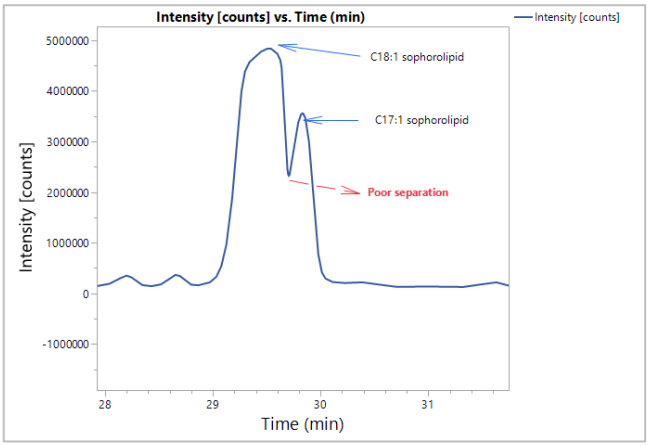
Functional Data Analysis, Functional DOE

Key ideas

This case study deals with functional data analysis and functional design of experiments (FDOE) for the optimization of an analytical method to allow for the accurate quantification of two biological components. Functional data is any form of data that occurs over a continuum, which can include time series, curves, surfaces, and spectra. Functional data analysis considers the whole data set over the continuum, rather than looking at one specific point (e.g., a set time, a set wavelength) and allows an understanding of how curves may vary among one another. FDOE can be used alongside it to provide insights into how input factors affect the curve shape, allowing for optimization to a desired shape. This study uses functional data analysis with JMP Pro.

Background

Bob is attempting to produce a high-performance liquid chromatography (HPLC) method that can separate two closely related sophorolipid biosurfactants (C18:1 and C17:1) that co-elute (appear close together) on the chromatogram – making quantification difficult.



To have a suitable HPLC method, Bob must be able to:

- Sufficiently separate the two peaks.
- Ensure that the peaks are sharp (small in width) and tall (high in sensitivity).

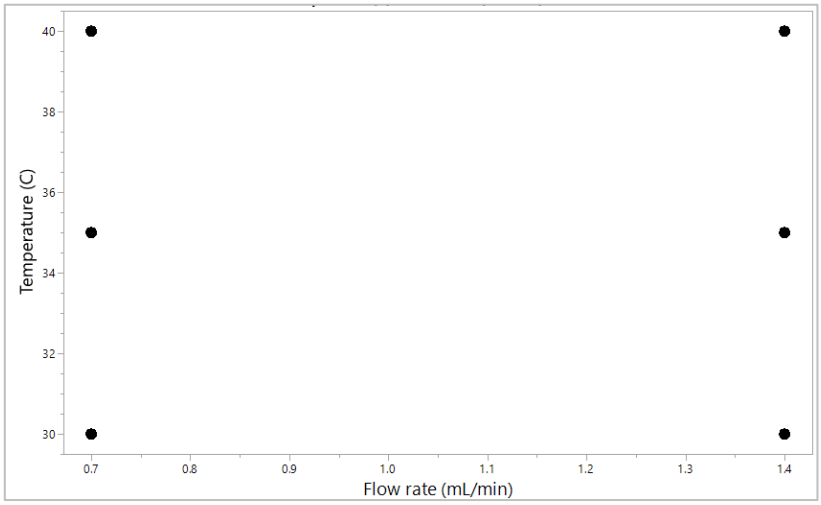
The task

Bob is entrusted with the following tasks:

- Apply statistical design of experiments (DOE) to find HPLC settings that can improve the method.
- Use functional data analysis to understand how the curve shape changes as factors change.
- Identify optimized conditions for the separation of the two peaks.

To improve the separation of the two peaks with the HPLC, two control parameters were chosen to be changed: the mobile phase flow rate and the column temperature. The temperature was altered to 30, 35, or 40°C, and the flow rate to 0.7 or 1.4 mL/min. The final arrangement of the experimental runs is shown in Exhibit 1.

Exhibit 1 Experimental design



The data

Bob decided to optimize his HPLC settings using DOE to change the temperature and flow rate to improve separation. The data set (FDE_HPLC.jmp) contains:

Experimental run: A unique identifier for each HPLC run with specific settings

Flow rate (mL/min): The specific flow rate for the experimental run

Temperature (°C): The specific temperature for the experimental run

Time (min): The time reading from the HPLC

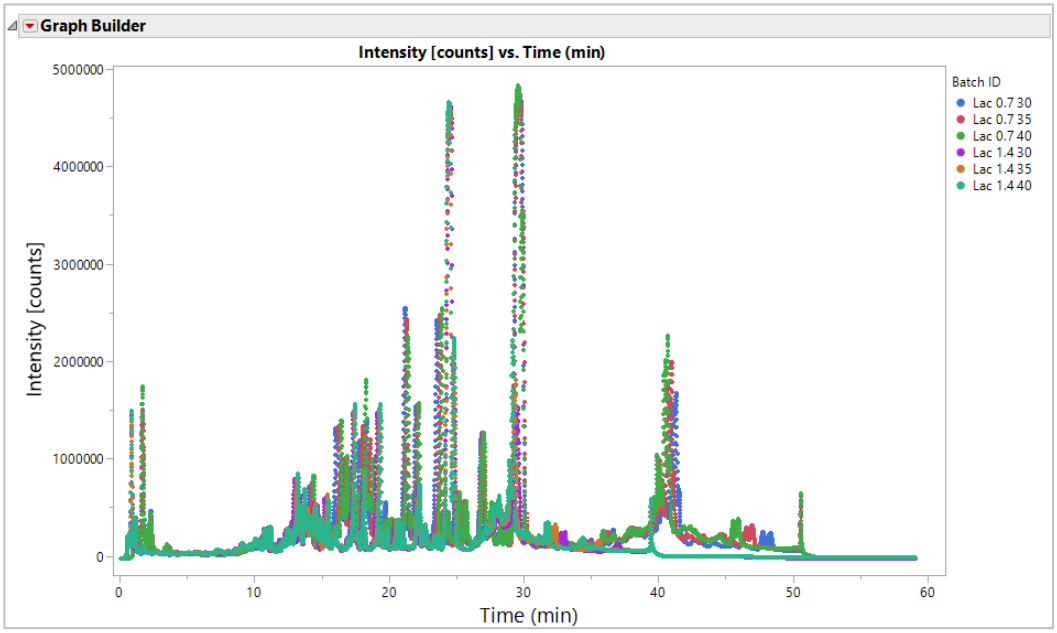
Intensity (counts): The intensity reading from the HPLC

Analysis

Exploring the raw data

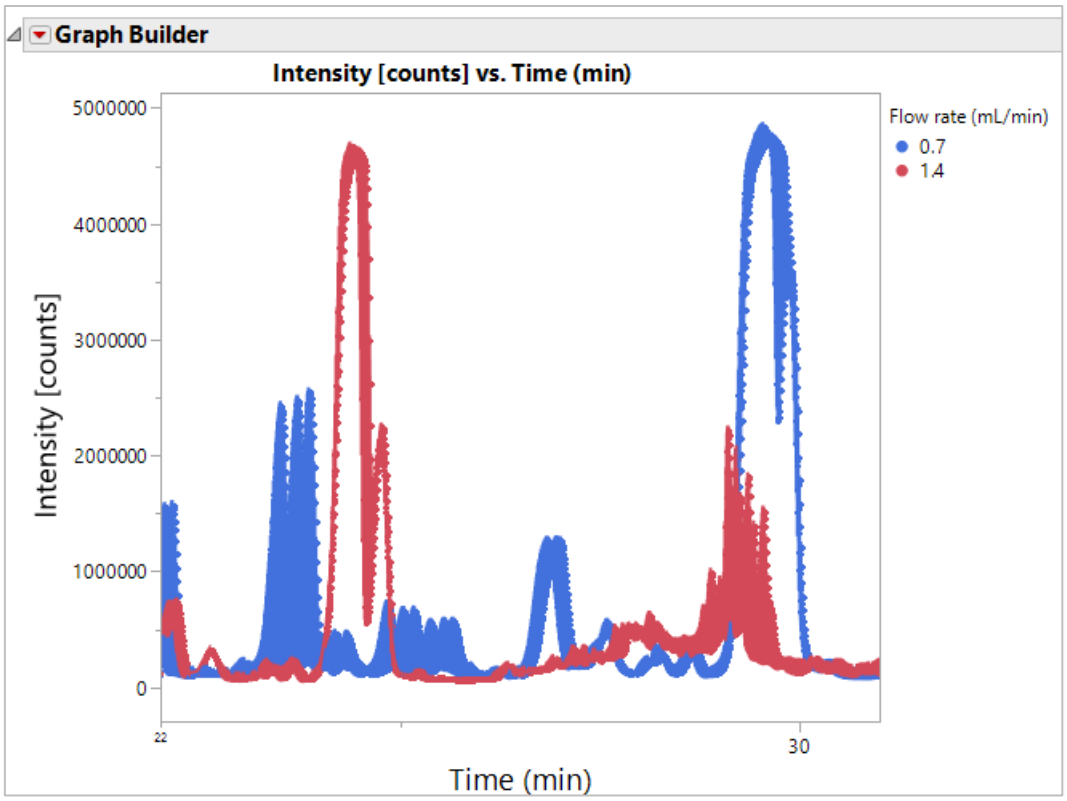
Let's explore the data using the Graph Builder in JMP. The data set shown in Exhibit 2 has all of the chromatographic data for the different experimental runs tested, making it extremely difficult to determine how changes in temperature and flow rate have changed the chromatogram (the output of the HPLC). In HPLC, lots of other compounds unrelated to the analyte of interest (the sophorolipid compound) can turn up, meaning there are a lot of peaks that are not of interest. We need to be able to narrow it down to the peaks relevant to the C18:1 and C17:1 sophorolipids.

Exhibit 2 Raw data output from the HPLC



To create this graph, Graph > Graph Builder. From the Control Panel, drag the Time (min) into the X axis of the graph and Intensity (counts) into the Y axis. To separate each data set into its respective experimental run, drag Experimental run from the Control Panel to Overlay.

Exhibit 3 Narrowed time series data



To create the graph as described in Exhibit 1, replace the Overlay with Flow rate (mL/min) from the Control Panel. Reduce the X axis scale settings to 22 (minimum) and 31 (maximum). Select the Line graph element to join each of the data points.

Previous experience with the HPLC assay helps us know that the peaks for C18:1 and C17:1 typically appear between 22-31 minutes, so the X axis range can be narrowed to look at how just these peaks have changed. Grouping by the flow rate in each experimental run produces two distinct groups; the large peaks in each of these groups represent the C18:1/C17:1 peak cluster. Within these groups, there is variation in the shape, height, and separation of the two peaks caused by the temperature. At this stage, the picture is still too complicated: how can we tell which particular settings caused the best separation? We could attempt to compare the peaks one-by-one (looking at each individual experimental run), but this would be time-consuming and not provide a simple metric for how the shape changes.

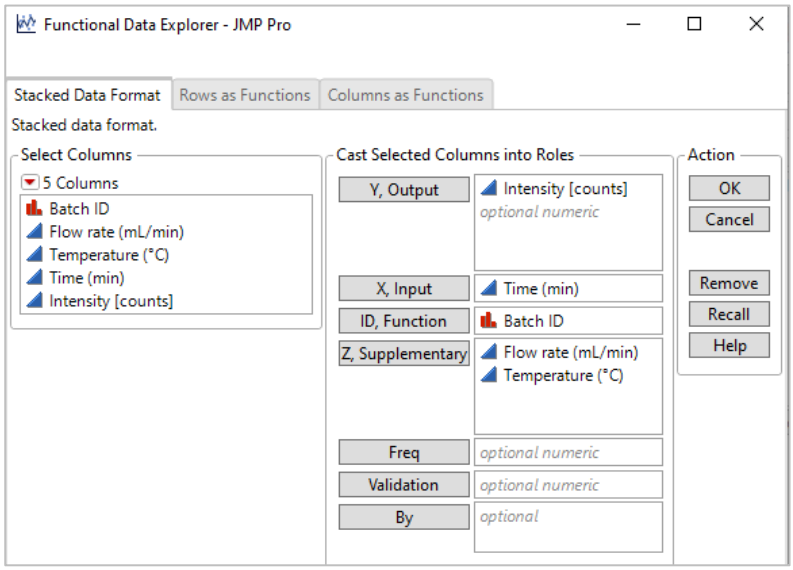
We could also try to approach this with a typical DOE analysis approach – distilling the peaks down to their desired characteristics (Peak 1 height, Peak 2 height, time difference between peaks) and fitting them to a model, such as standard least squares, to find the optimum settings. There are multiple challenges to this approach, including:

- Selecting the right criteria. Does the complexity of your waveforms boil down simply to height and time difference? Are there more complex parameters that need to be accounted for (width, retention time etc.)?
- Understanding the model. How can you visualize the findings from each of the individual models together in a coherent form? The more characteristics you have, the more difficult it becomes to describe the complex relationship between each point.
- Communicating your findings. Once you've collected the findings from each model for each characteristic, how can you easily communicate them? The key to data visualization is ensuring it can be understood by anyone, but the methods described above are prohibitive to communication.

Ultimately, we want to understand the variation of the curves across the changes of temperature and flow rate, finding the settings that best suit our needs (good separation). To analyze this, we can use functional data analysis.

Applying functional data analysis

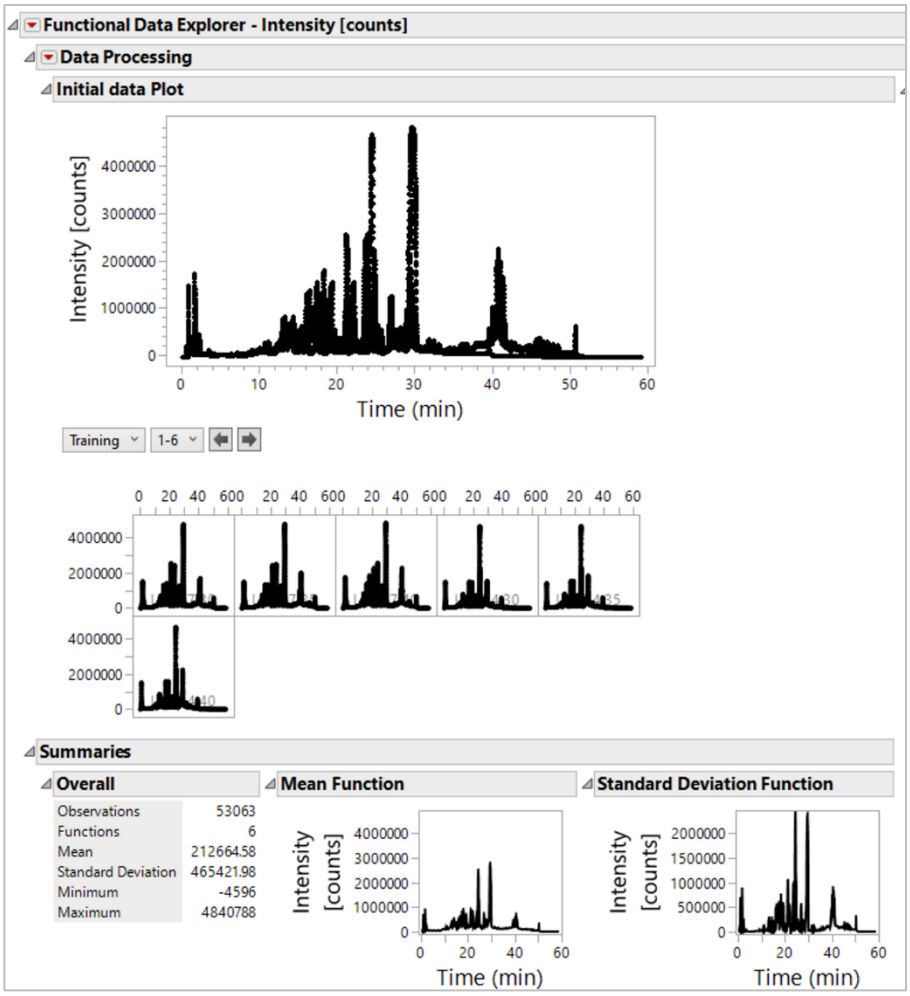
Exhibit 4 Functional Data Explorer selection window



To launch Functional Data Explorer, select Analyze > Specialized Modelling > Functional Data Explorer and input your data as shown above.

To explore how the characteristic of the peaks (changes in height, width, and position over time) change with the flow rate and temperature, we can apply functional data analysis. With this type of analysis, the discrete observations taken over time (e.g., the intensity measurement of the HPLC sensor, measured every 0.36s) are expressed as a smoothed continuous function.

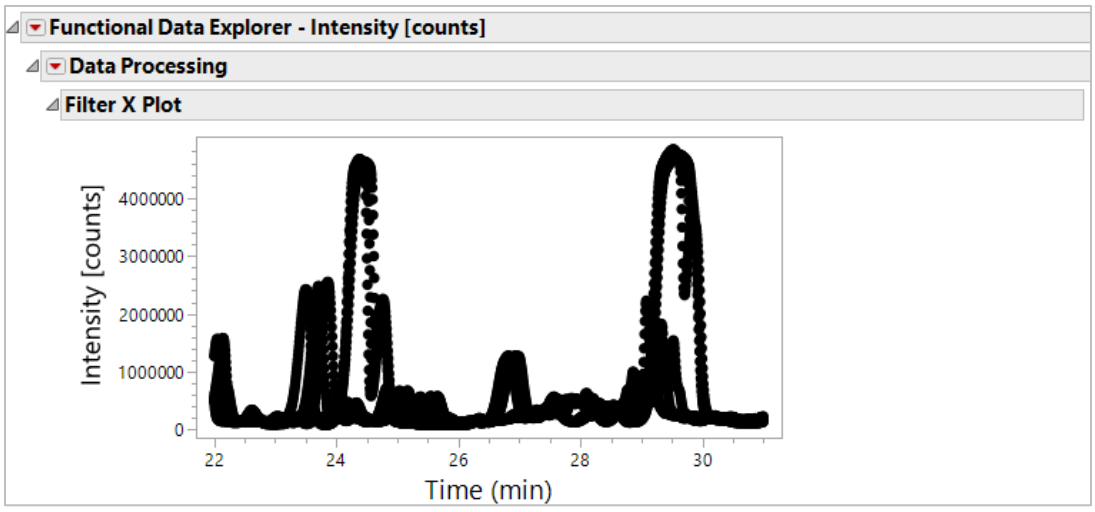
Exhibit 5 Initial Functional Data Explorer output



The initial Functional Data Explorer output is shown in Exhibit 5. The platform displays the data from each HPLC experimental run that are to be used to train the functional data model (Initial Data Plot) alongside summaries of the data. The summaries show the mean function (the average values at each time point across the experimental data) and the standard deviation function (the average deviation from the mean value across runs at each time point). The variation from the mean is mostly caused by the changes in temperature and flow rate; this variation is characterized using the functional data analysis in later steps. As with the Graph Builder plot shown before, the chromatogram is very busy with peaks that we are not interested in. Since we know that the peaks of interest appear between 22-30 minutes, we can reduce the range by filtering the X axis.

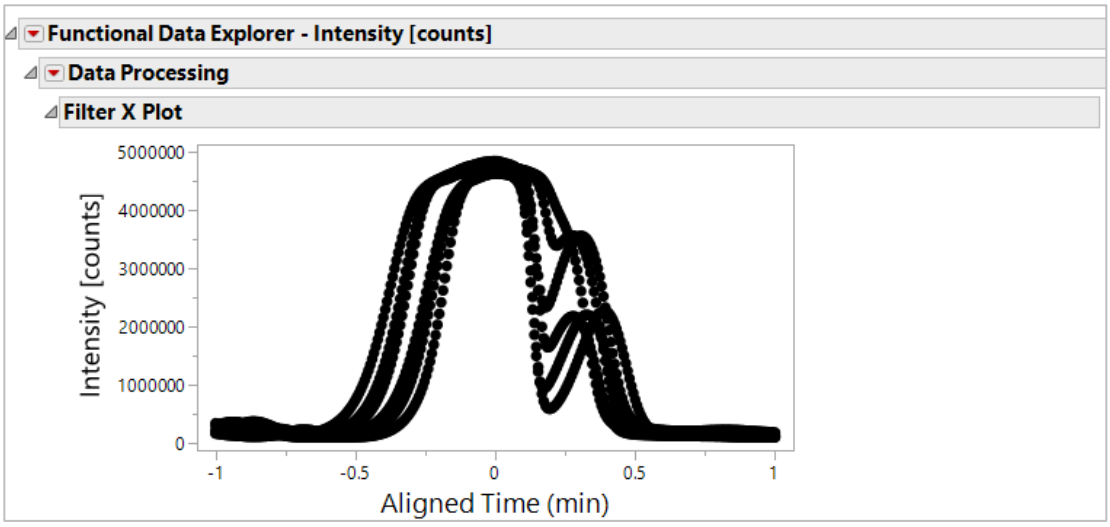
With this reduction, it becomes clearer that we are looking at the larger peaks for each run; however, the variations in flow rate and temperature have also caused a difference in the peaks' retention times (the time they appear on the chromatogram). Our interest is in how the specific peak shape changes and how the two peaks separate, not in when they appear on the chromatogram. To be able to look at these peaks together, they must be aligned. This alignment can be done easily by aligning the maximum peak values: taking the largest intensity count of each peak and placing it as the center at 0 minutes. From here, cleaned discrete time values are created that are ready to be analyzed as functions.

Exhibit 6 Functional Data Explorer: reducing the time axis



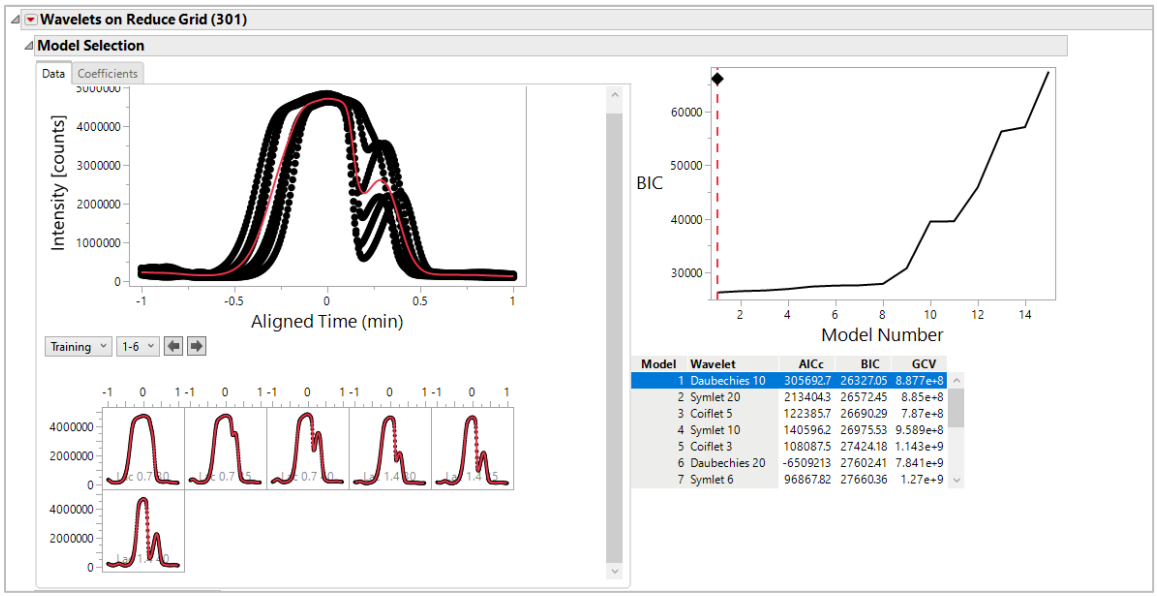
To reduce the time axis, select Cleanup > Filter X > Below 22, Above 30.

Exhibit 7 Functional Data Explorer: aligning waveforms



To align, go the Commands window and select Align > Align Maximum. To remove unwanted peaks, reduce the time range by using Cleanup > Filter X > Below -1, Above 1.

Exhibit 8 Creating a functional wavelet model



To model the data, select Model > Wavelets from the top red triangle. The best model is pre-selected automatically.

Before we can determine the characteristic shape components that describe the variety of chromatograms, we need to first turn the discrete, semi-continuous data into continuous functions using an interpolation or "smoothing" model. Chromatograms, spectra and diffraction patterns are usually modeled effectively with wavelets, which are forms that oscillate from 0 to a maximum before returning to 0. At this stage, it is important to ensure that the smoothing model is sufficiently complex to capture the "real" behaviors in each chromatogram. But we need to avoid overfitting to the "noise." The lowest Bayesian Information Criterion (BIC) is a useful guide to selecting an appropriately complex model, along with your domain experience, which you can apply with visual checks of the fit to each chromatogram.

Now that we have the shape of each curve represented with a continuous function, we need to determine how these curve shapes are common or differ between each experimental run; we can do this by extracting the key characteristic shapes that compose each curve. Functional principal component analysis (FPCA) is performed to decompose each curve into a mean function, a series of characteristic shape functions (eigenfunctions), and functional principal component scores (eigenvalues).

Any original chromatogram can now be described as a combination of the mean function plus or minus some amount of each eigenfunction:

$$\boldsymbol{f_{BatchID}(time) = \mu_{BatchID}(time) + FPC_{1,BatchID} \times \varphi_{1,BatchID}(time) + FPC_{2,BatchID} \times \varphi_{2,BatchID}(time) + FPC_{3,BatchID} \times \varphi_{3,BatchID}(time) + FPC_{4,BatchID} \times \varphi_{4,BatchID}(time) + FPC_{5,BatchID} \times \varphi_{5,BatchID}(time),}$$

where μ = mean function at the given time and φ is the value for each shape function at each given time.

Exhibit 9 Functional PCA output

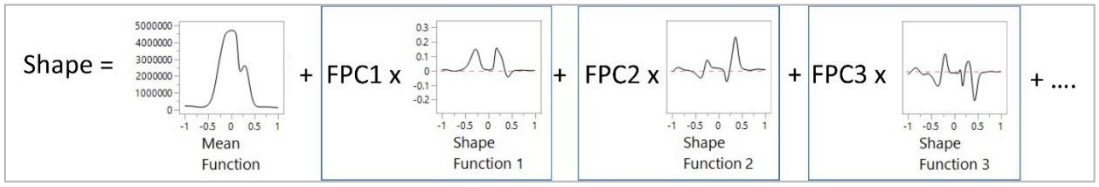


Exhibit 10 Functional PCA output

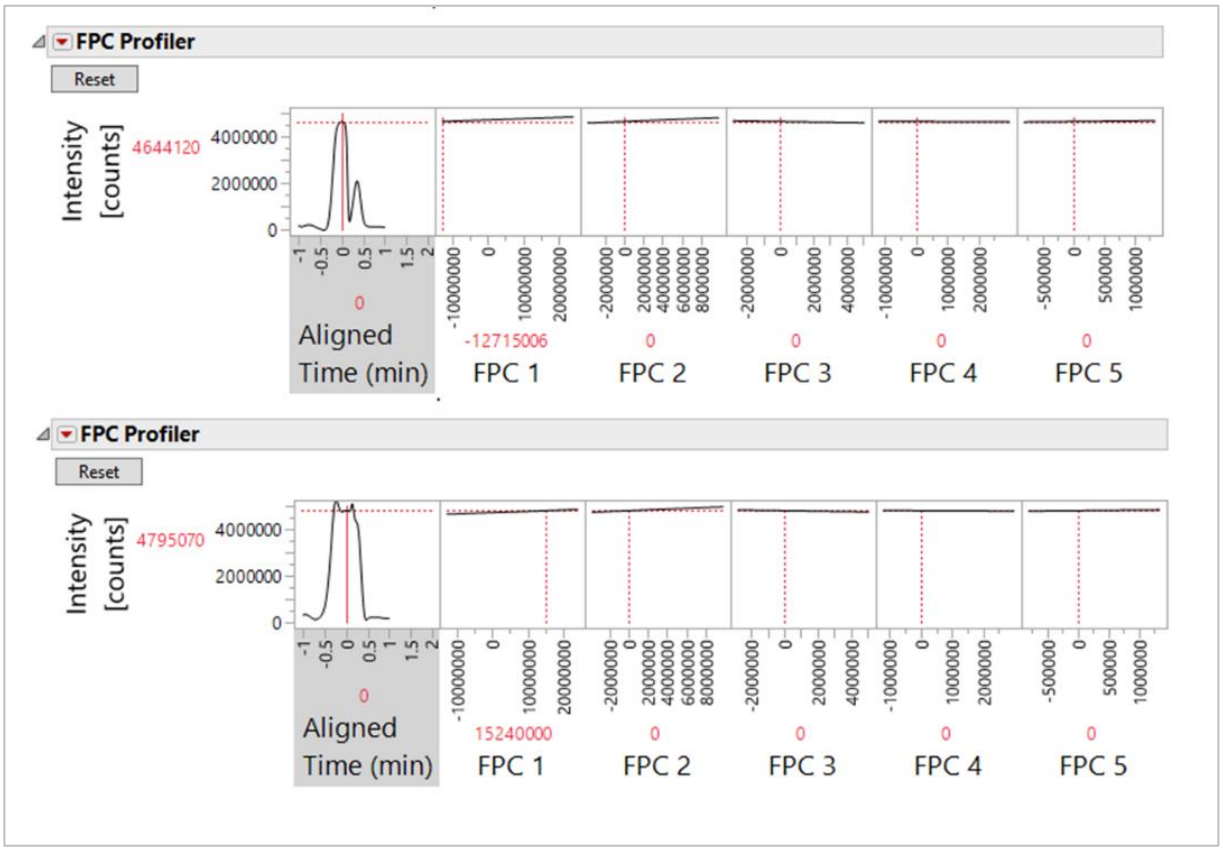
Function Summaries

Batch ID	FPC 1	FPC 2	FPC 3	FPC 4	FPC 5	FPC 6	FPC 7	Mean	Std Dev	Median	Minimum	Maximum
Lac 0.7 30	14762114	-3874738	3222206.7	1531684.7	-168411.5	282062.01	-20901.47	1869357.5	1820919.1	636747.19	127698.85	4735939.2
Lac 0.7 35	13367665	-1243860	-4221778	-2026799	-302554.8	114502.2	29269.556	1998162.8	1677109	1229909	131511.79	4755215.6
Lac 0.7 40	8511829	4720950.6	-1774488	1583735.6	564369.56	-602848.4	-23545.79	1743005.6	1747676.3	514972.83	156084.49	4840670.4
Lac 1.4 30	-638141.5	1353205.6	4442528.5	-1955676	863691.79	-171596	105051.74	1241089.6	1614531.9	195748.1	90659.604	4629133.2
Lac 1.4 35	-4635484	3467716.1	1679632.5	-342140.2	-1306115	295453.21	-215161.3	1175427.1	1547093.6	217972.21	89589.466	4640206.1
Lac 1.4 40	-9273640	911620.02	-1507330	819548.5	256964.56	1034621.1	250326.43	1115324.6	1509569.9	211384.32	93444.247	4675462.8
Lac 1.6 45	-11288635	-2873547	-1500170	43914.604	825903.35	-5224.935	-326474.1	1051311.6	1464251.6	195769.01	93209.392	4633972.3
Lac 1.8 45	-10805707	-2461347	-340600.4	345731.34	-733847.8	-946969.2	201434.98	1027022.7	1430223.7	197933.16	90436.878	4613084

We can capture the equation above with a simple graphic (Exhibit 9), taking the mean and shape functions from the functional PCA output to represent how a given shape is formed from the mean and shape functions. The shape function characterizes the dominant characteristic shapes and patterns that appear within the experimental run (within run variation), while the FPC scores represent the weight of each of those shape functions that are added to the mean function, with each experiment having a unique value (run-to-run variation). In this example, there are seven FPC scores/shape functions that are used to recreate the curve of a run (Exhibit 10). These scores can be negative or positive, leading to the subtraction or addition of each shape function from or to the mean function. The shape functions are chosen so that they are orthogonal to each other, meaning they are uncorrelated, and capture the variation in the data in an ordered manner, with the first shape function capturing the most important mode of variation, the second capturing the second most important mode, and so on. By using shape functions to represent the data, it becomes possible to identify and analyze the most important patterns in the data in an efficient way.

Understanding what these shape components represent can be difficult, but we can surmise their effects on the curve by looking at what the addition/subtraction of the shape function does. With the Functional Principal Component Profiler (Exhibit 11), set each FPC score to 0. The resulting peak shape will match the mean function of all of the curves. We can then increase and decrease the value of FPC1. Increasing values over 0 causes the two peaks to join and become wider. Decreasing this value causes the two peaks to split apart and become more narrow. We can then infer that FPC1 represents the separation/joining of the two peaks and that our desired experimental run should have a low FPC value.

Exhibit 11 Functional Principal Component Profiler



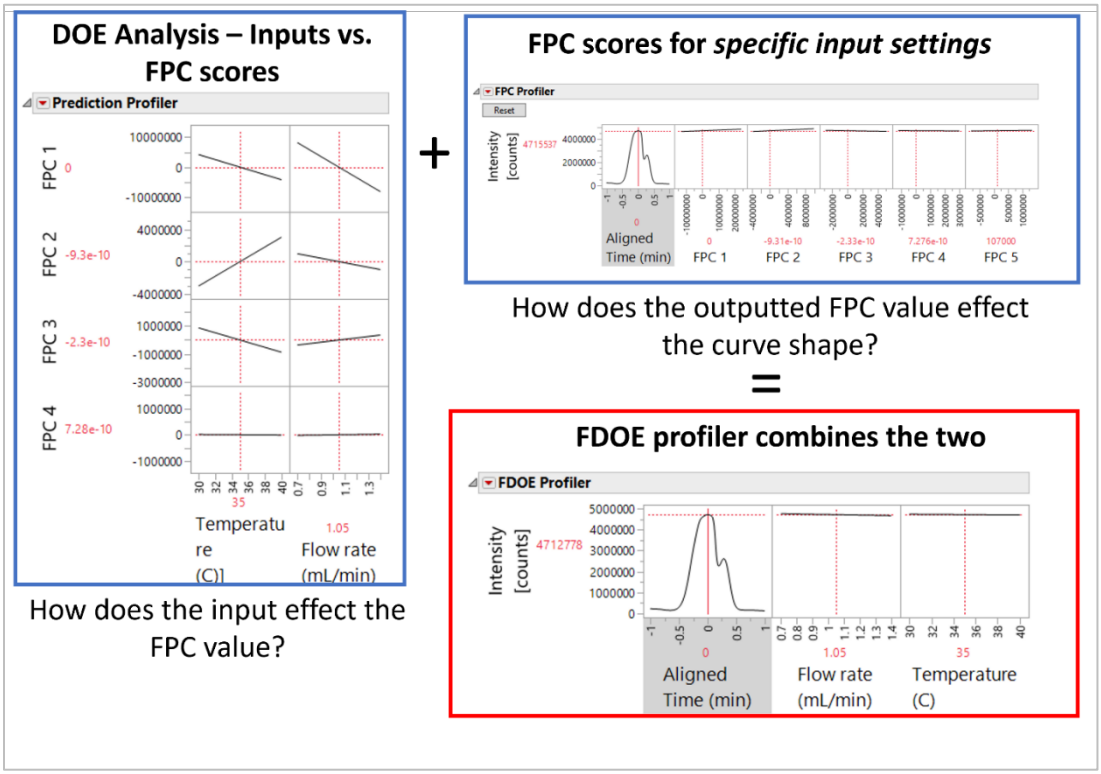
The Functional PCA section is displayed once any given model is selected. The appropriate number of components is selected by default but can be altered in the FPCA Model Selection. The setting here shows the effect of reducing the FPC1 value

We have obtained functional components that represent the variation in chromatograms across the runs. We can now usefully describe each chromatogram using only the FPC scores and can successfully recreate them with these values. However, the goal of this study is to find the direct effect of the inputs (flow rate and temperature) on the shape of the chromatogram curves. To achieve this goal, we can use the FPC values as responses that we model against the input factors (temperature and flow rate) in a standard DOE model, allowing us to model the effect of the input factors on the chromatogram curves.

Functional design of experiments

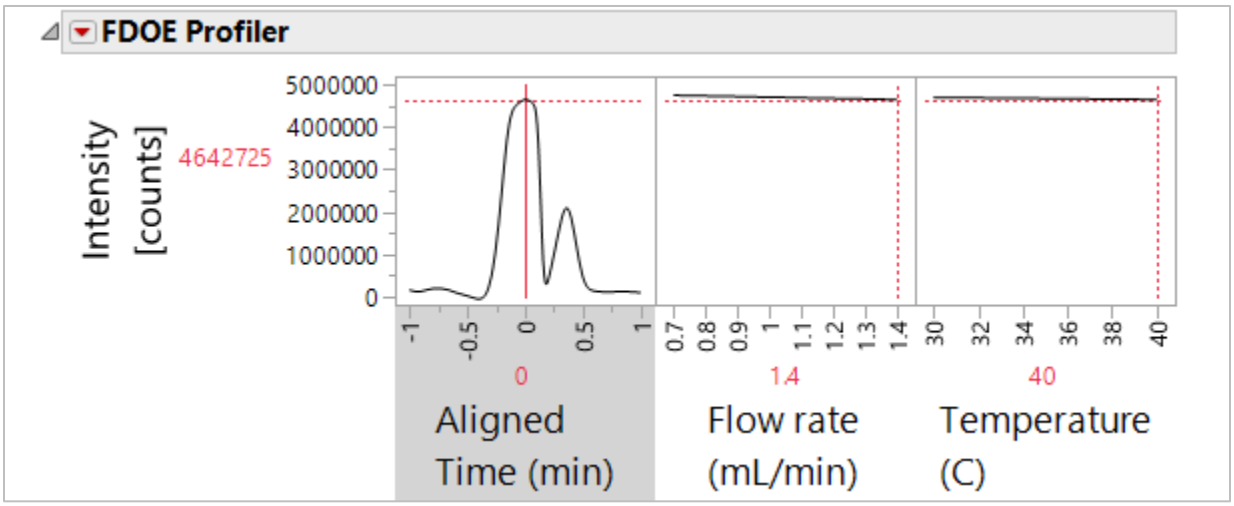
From the functional model we have generated single values (FPC scores) that represent the shape of the chromatogram curves, but how do we go from these values to relating the effect of the input on the overall shape? Functional DOE works by using these FPC scores as regular responses (outputs) in a DOE model against the flow rate and temperature (inputs). This method produces a profiler where FPC scores can be predicted for any given settings of temperature and flow rate, allowing us to link the inputs (flow rate and temperature) to the curve shape via the FPC scores. A simplified flow chart is shown in Exhibit 12, demonstrating when the temperature and flow rate are at 35°C and 1.05mL/min, respectively. The generalized regression model tells us the predicted FPC values (FPC1 = 0, FPC2 = -9.3e-10). We can take these FPC values into the FPCA profiler and see what the resultant curve shape would be. The FDOE Profiler in JMP Pro combines these two aspects into a simple, easy-to-use profiler so that you only see the inputs' effect on the curve shape.

Exhibit 12 Functional DOE process



From the FDOE Profiler, the optimum separation was found with maximum settings (flow rate 1.4mL/min and temperature 40°C). The FDOE platform provides a simplified tool that characterizes the change of the shape of the curve, not just the change of a single parameter. In cases such as chromatography where numerous characteristics are desired (peak height, peak width, sensitivity, separation, etc.) and may conflict with one another, functional analysis provides the best tool to accurately predict the best conditions. The FDOE model shows how flow rate and temperature affects each FPC score. The combination of each eigenfunction multiplied by its FPC score produces the specific curve shape for that combination of scores. And the FPC scores are modeled with the DOE factors using generalized (penalized) regression to produce the predictive profiler shown in Exhibit 13.

Exhibit 13 Functional DOE Profiler



To add this profiler, select the red triangle on the Wavelets model section and click Functional DOE Analysis.

Summary

Managerial/business implications

Gaining accurate analysis is vital when performing analytical measurements in business, as it is used to quantify the level of productivity and confirm quality (purity analysis). Traditional method development is laborious when analyzing chromatographic data, due to the complexity and size, typically requiring analysis “by eye” for each experiment one by one. Functional data analysis allows for rapid assessment of experiments by summarizing all the experimental runs into a single FDOE profiler – simplifying data analysis, reducing turnaround time for results, and improving the quality of analysis.

Statistical insights

Gaining insights from curve data can be difficult, due to the complexity (multiple shapes, misalignment), size (large data sets), and inability to gain meaningful insights with traditional “single output” models (peak height, width, slope). In this example, the data has been simplified with pre-processing (X axis filtering and peak alignment) and converted to continuous curves with curve fitting (wavelets) to allow for the whole curve shape to be modeled. The key variations in shape have been distilled from the curves into shape functions with functional principal component analysis and modeled with functional DOE to link the changes in temperature and flow. From this we can conclude:

- The optimum settings for separating the two peaks are 40°C and 1.4mL/min.
- The separation is affected by both the flow rate and the temperature.
- Peak height is not affected by the flow rate or temperature, meaning sensitivity is not reduced.

JMP features and hints

JMP was used in this case study to:

- Visually represent the chromatographic data using Graph Builder.
- Reduce the data set by narrowing of the X axis with the Data Filter tool.
- Clean up the data set in the Functional Data Explorer tool using X axis filtering and peak alignment (Align Maximum).
- Fit curves to the chromatographic data with wavelet curves.
- Find shape functions and FPC scores with functional principal component analysis.
- Model the effect of experimental conditions (temperature and flow rate) to the shape of the chromatogram curves with functional DOE and the Functional Profiler.

Exercise

1. Apply a B-spline, P-spline, and wavelet model to the data set. How does the model accuracy change? Explore the Actual by Predicted plots.
2. Look at increasing and reducing the number of FPCs in the wavelet model from the default value. How does this affect the accuracy of the functional DOE model? What is the ideal number of FPCs?
3. Simulate the intensity profile at 0.7mL/min flow rate and 40°C and 1.4mL/min and 30°C from the Prediction Profiler. Compare these results to the actual readings using Graph Builder. How do they compare? Where does it not fit well?
4. What would you suggest as a follow-up experiment?