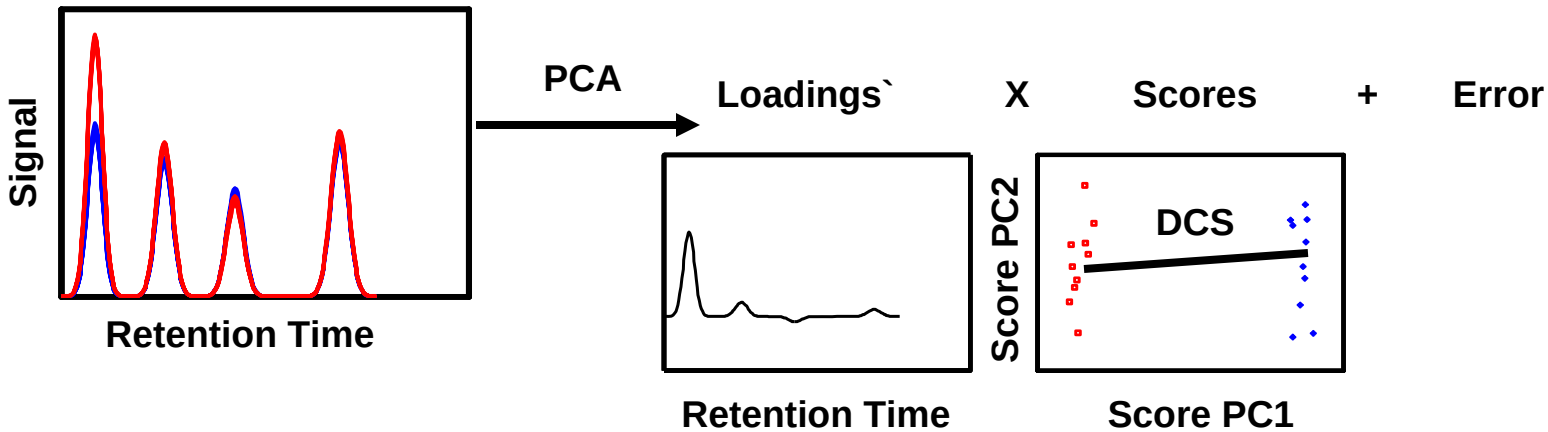
Alignment and Preprocessing for Data Analysis

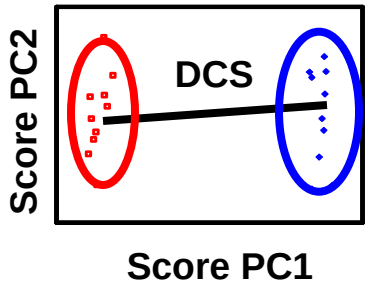
- Preprocessing tools for chromatography
- Basics of alignment
- GC-FID (1D) data and issues
  - PCA
  - F-Ratios
- GC-MS (2D) data and issues
  - PCA
  - F-Ratios
  - PARAFAC
- Piecewise Alignment GUI (available online)
  - synoveclab.chem.washington.edu/Downloads.htm
  - Email for username/password

Tools for Analysis: Classification

- Principal Component Analysis (PCA)



- Degree of Class Separation (DCS)



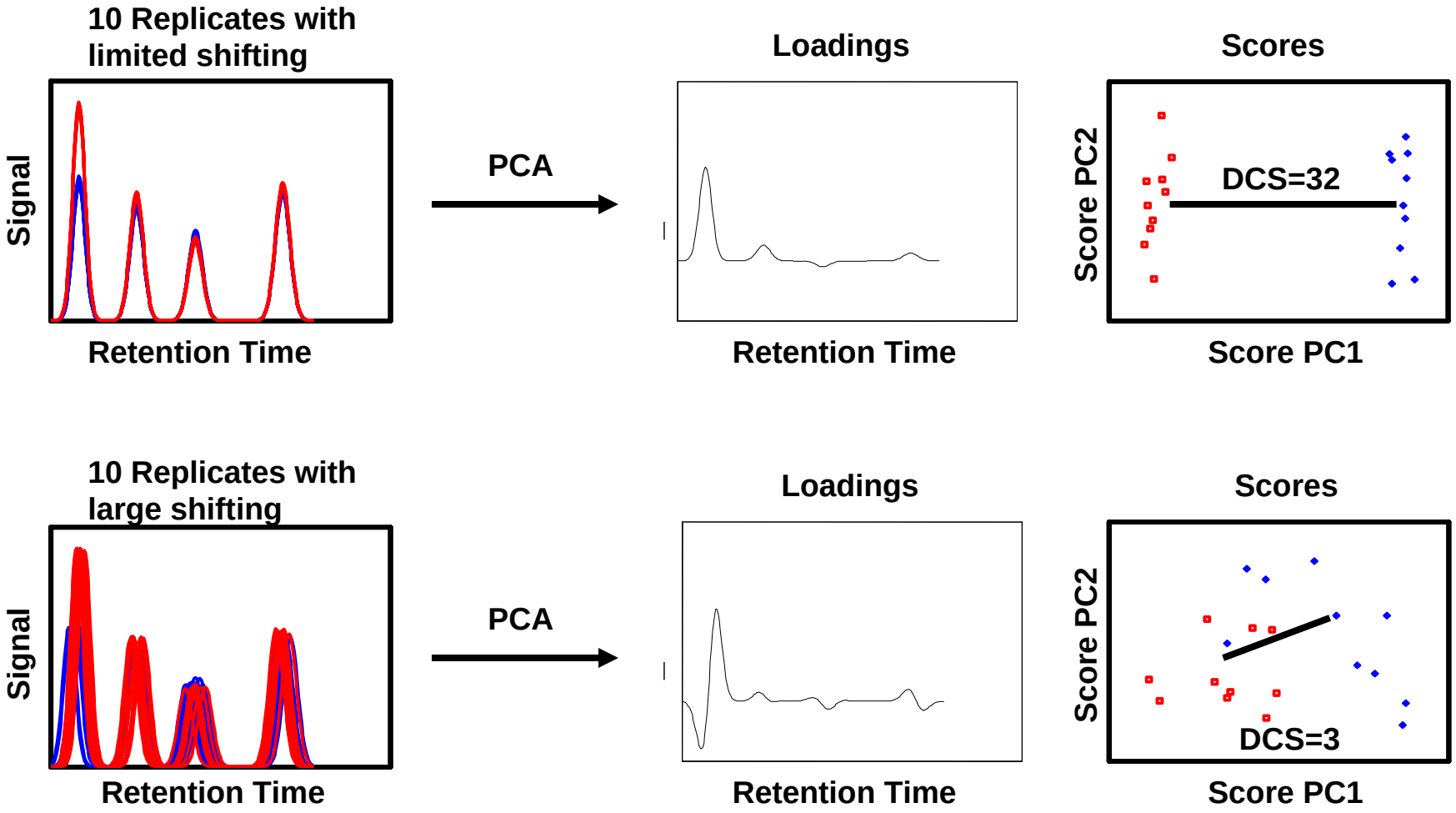
$$DCS = \frac{D_{A,B}}{\sqrt{s_A^2 + s_B^2}}$$

$$D_{A,B} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}$$

Why Align?

- Reduction in classification
  - PCA
- Increase in uncertainty for quantification
  - PARAFAC
- Misalignment occurs frequently
  - Daily instrument variation causes misalignment
  - Correction is necessary to apply these methods

Retention Time Precision & PCA



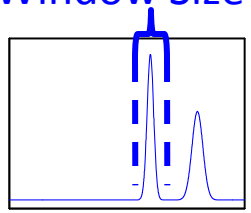
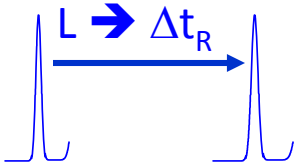
Basics of Alignment

- Types of alignment algorithms
  - Cross correlation coefficient
  - Correlation Optimized Warping (COW)
  - *Piecewise alignment
- Alignment Parameters
  - Window Size
  - Shift
- Target Selection
  - PCA
  - Correlation Coefficient
  - Windowed Target

Alignment Algorithms

- Cross correlation coefficient
  - Move the entire chromatogram to maximize correlation
- Correlation Optimized Warping (COW)
  - Separate the chromatogram into windows
  - Warp and move the windows to optimize the correlation
  - Find the best alignment path to correct the data
- *Piecewise alignment
  - Separate the chromatogram into windows
  - Shift the windows to optimize the correlation
  - Find the best alignment path to correct the data

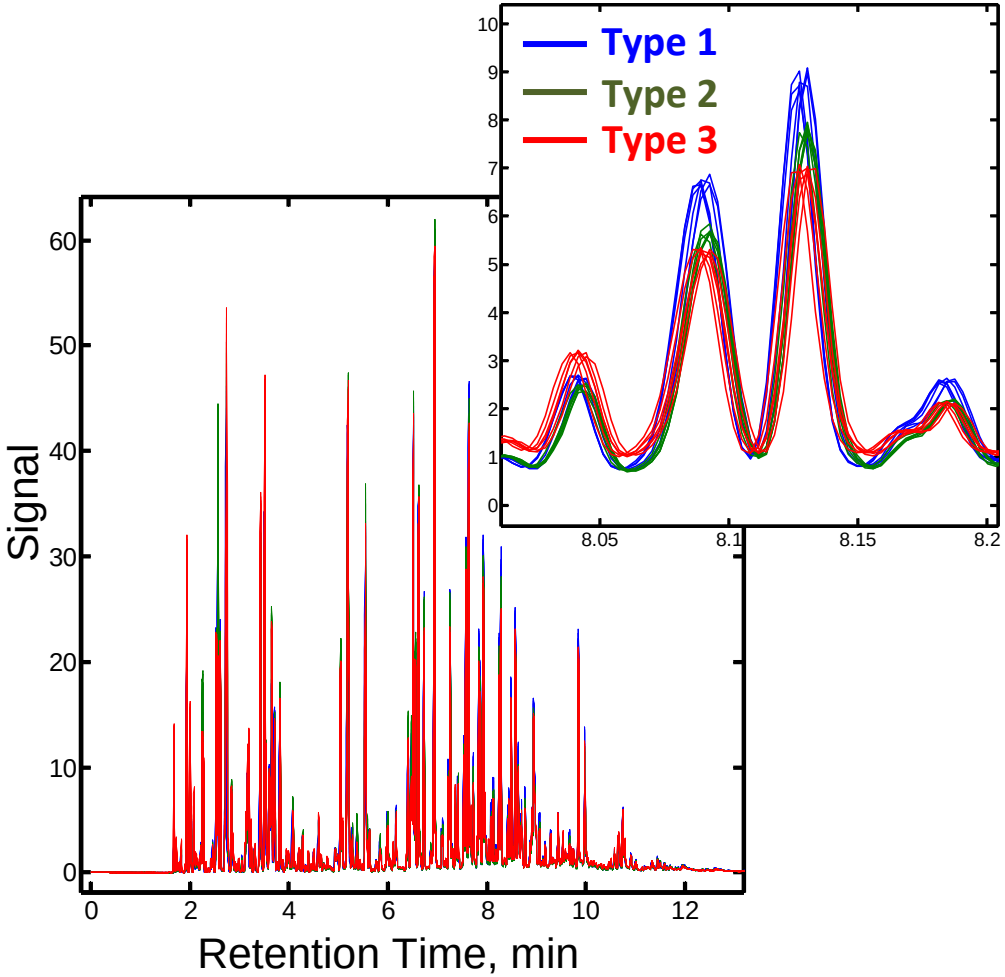
Alignment Parameters

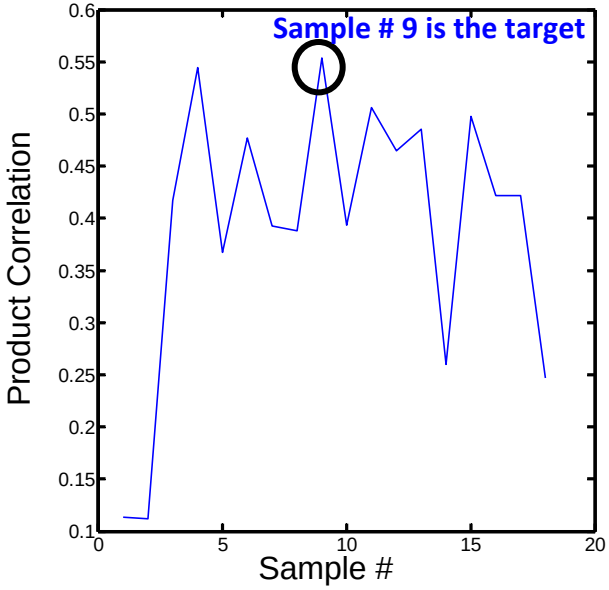
Parameter	Description	Effect Too Small	Effect Too Large	Determining correct values
Window Size, W	Window Size ~1 pk width 	Relative movement high, difficult to determine quality of alignment	Insufficient flexibility to correct peak to peak shifting	Alignment Metric
Shift, L	0<shift<=maximum shift 	Insufficient movement of segments	Increases time	Alignment Metric

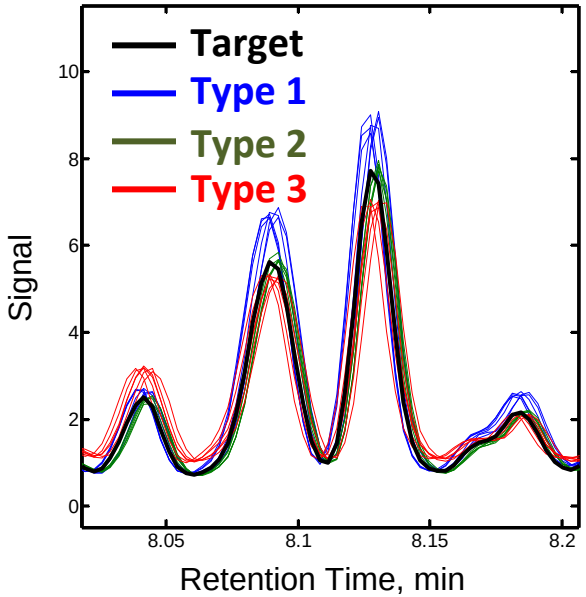
Target Selection

- *Global Approach
  - All chromatograms are initially collected
  - *User chosen target
  - PCA optimized target
    - Scores are produced for every sample
    - The sample with the minimum distance from the center is the target
  - *Maximum correlation target
    - Calculate the product of each chromatograms correlation to the others
    - The maximum correlation is the target
- Online Approach
  - An initial target is set
  - Chromatograms are aligned as they are collected
  - The target changes as new chromatograms are collected

Target Selection (Maximum Correlation)







Alignment of GC-FID (1D) Data and Issues

- Removal of artifacts and solvent peaks
- Baseline correction and normalization
- Alignment
- Improving PCA
  - F-Ratio

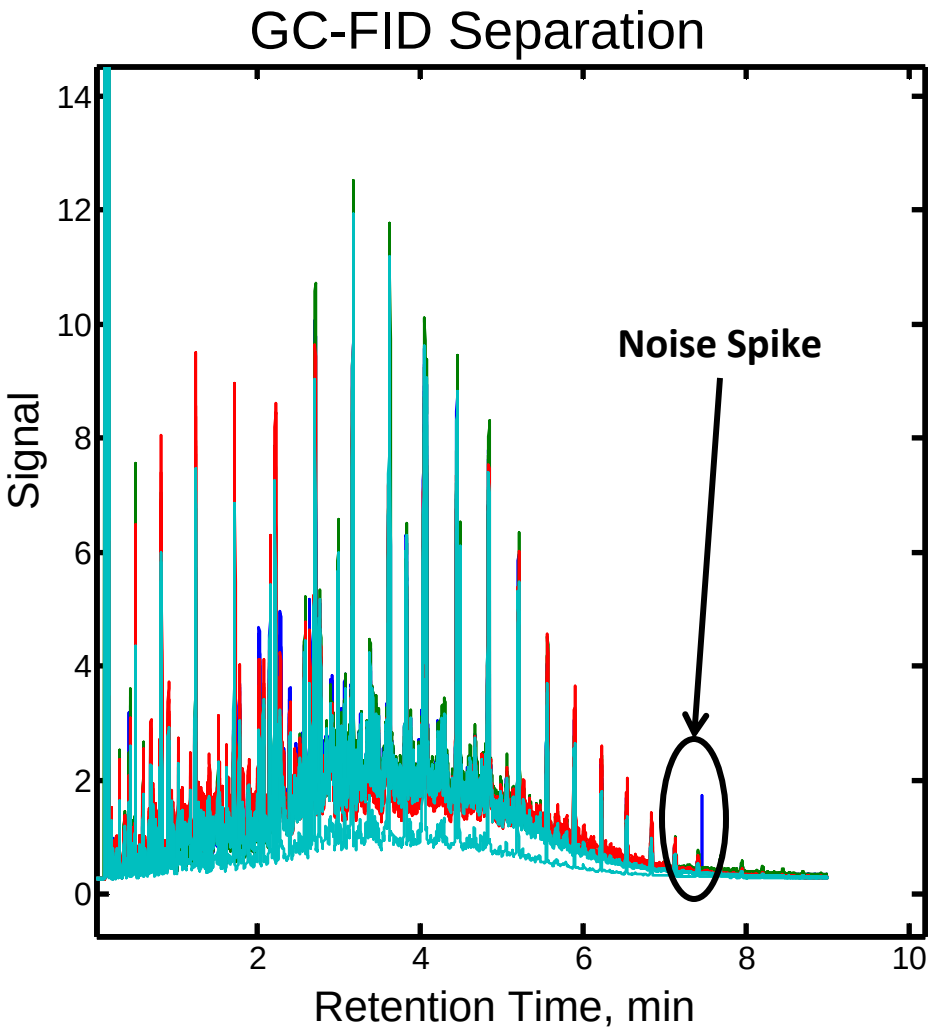
Preprocessing Tools for Chromatography

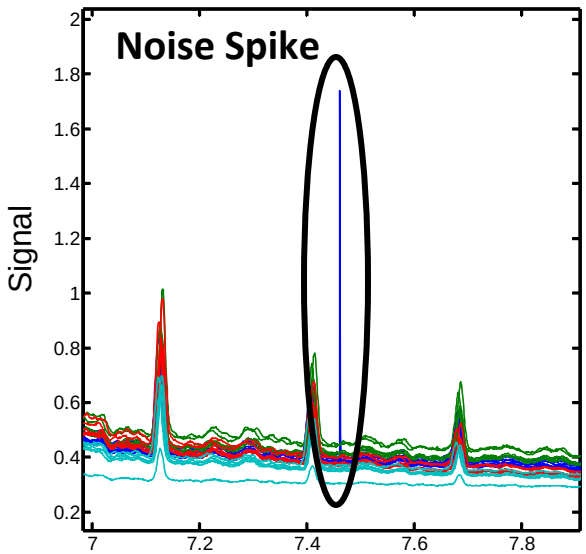
- Noise filtering
  - Median filter
- Baseline correction
- Normalization
- Alignment

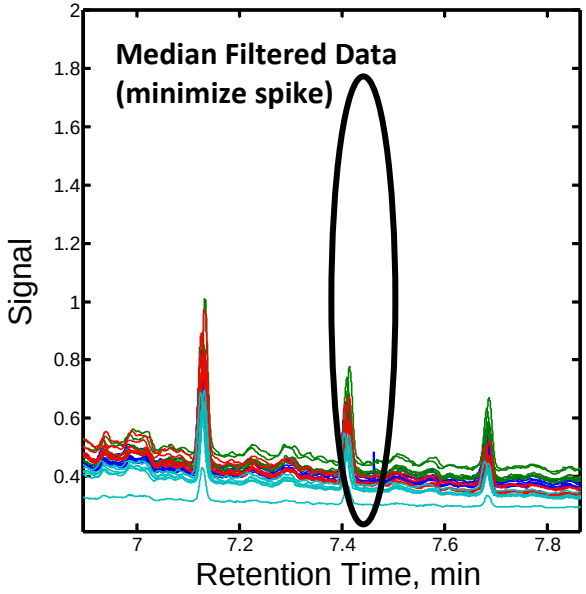
Experimental

- GC-FID Separation
- 4 diesel sample types
- 9 minute separation
- 17 replicate injections over 5 days

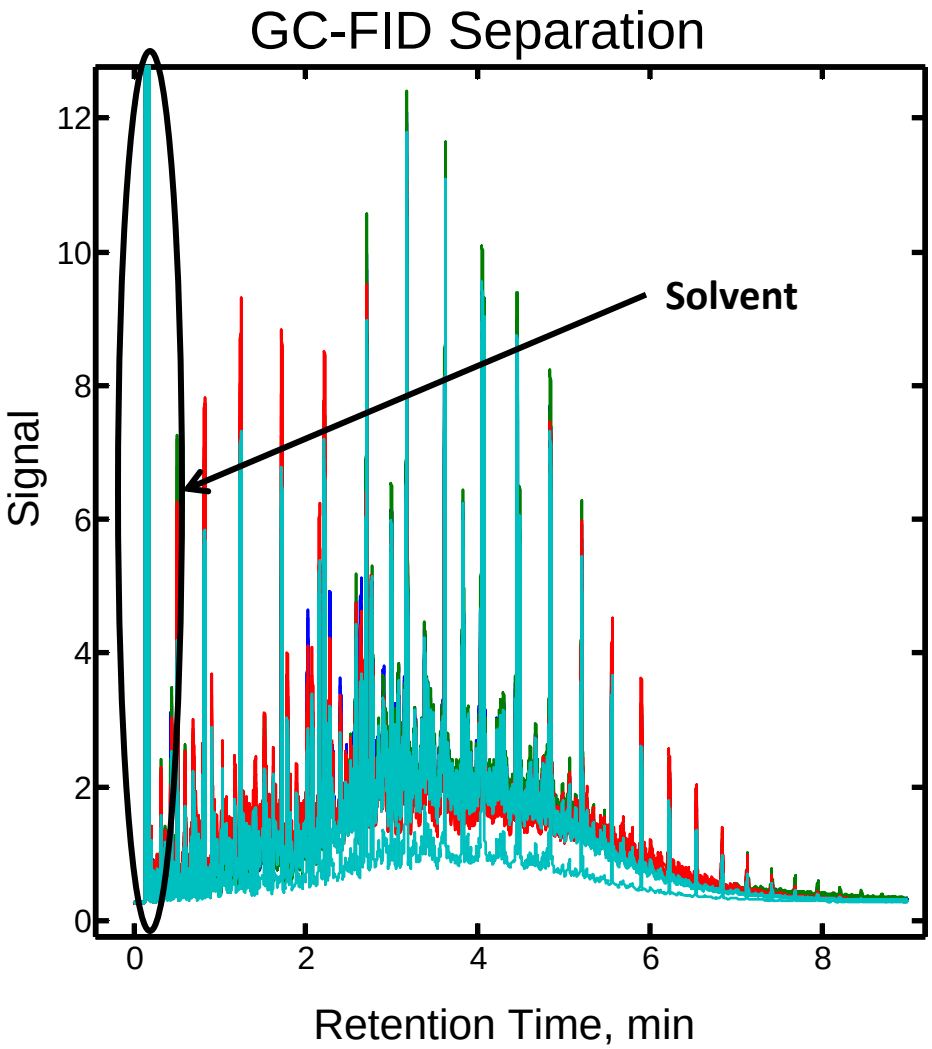
Noise Filtering (Solvent, Spikes, and Outliers)

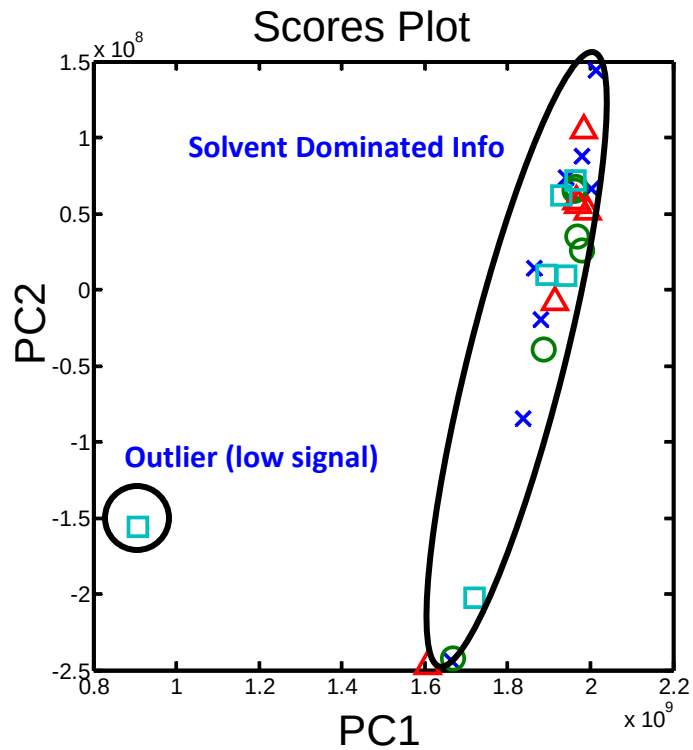




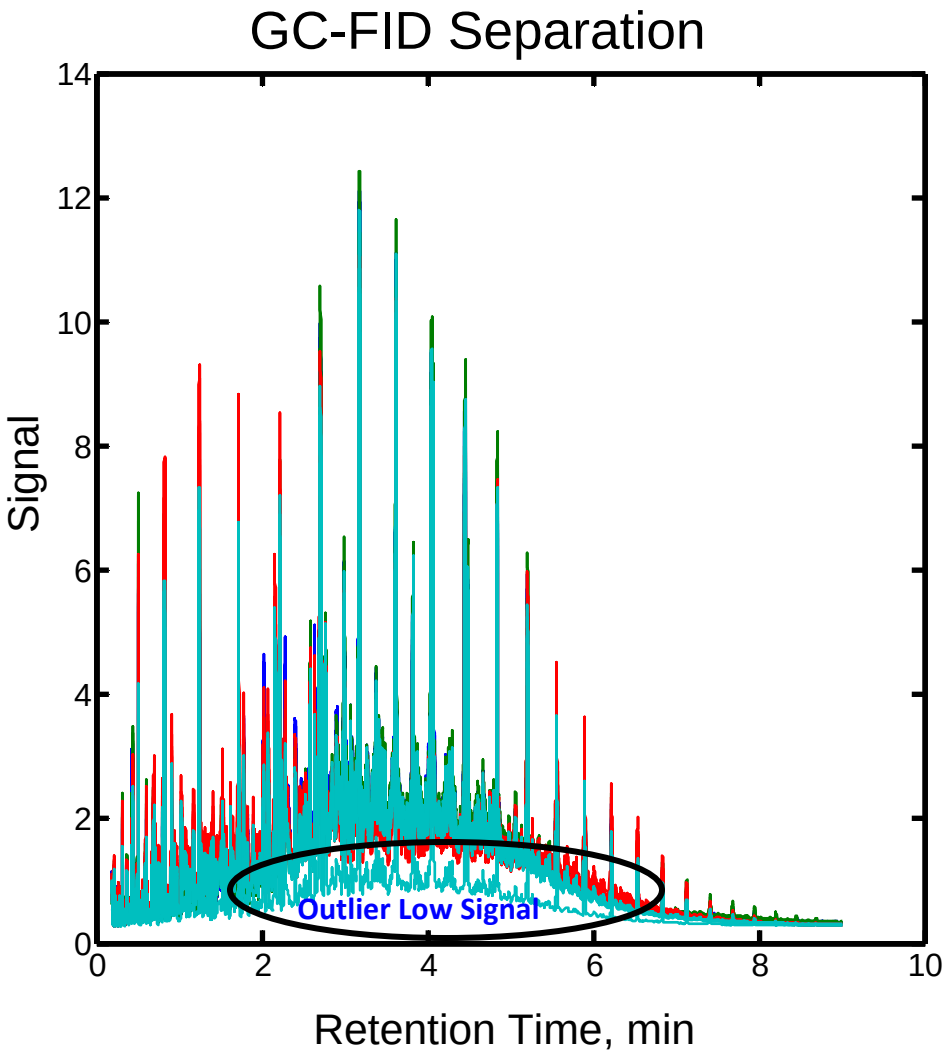


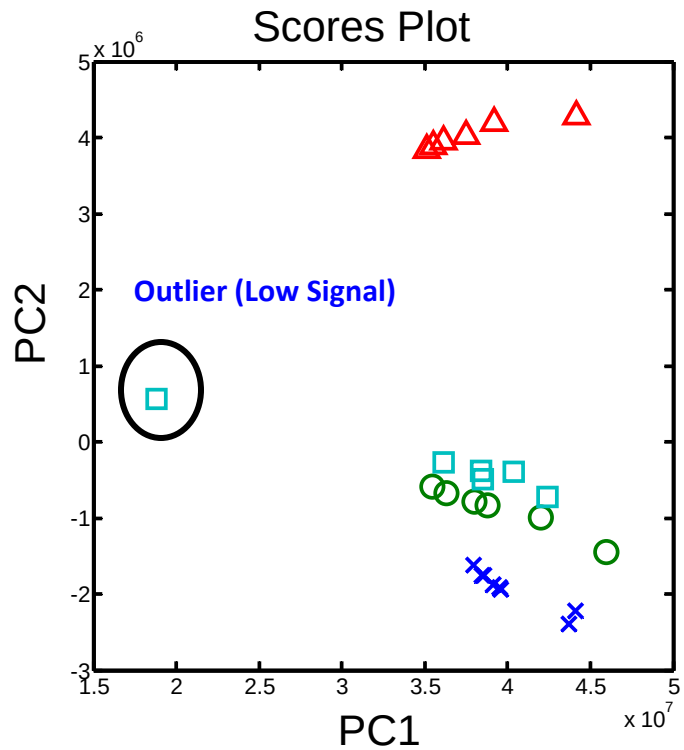
Noise Filtering (Solvent, Spikes, and Outliers)



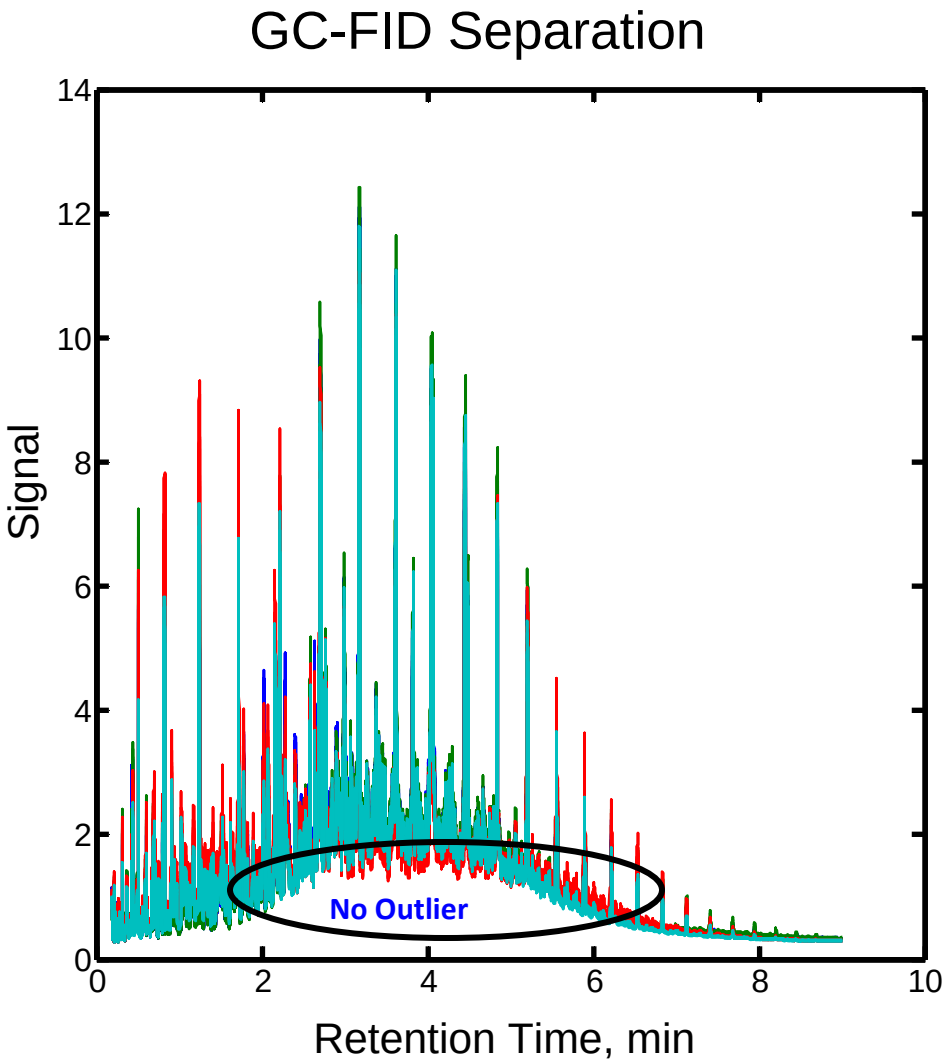


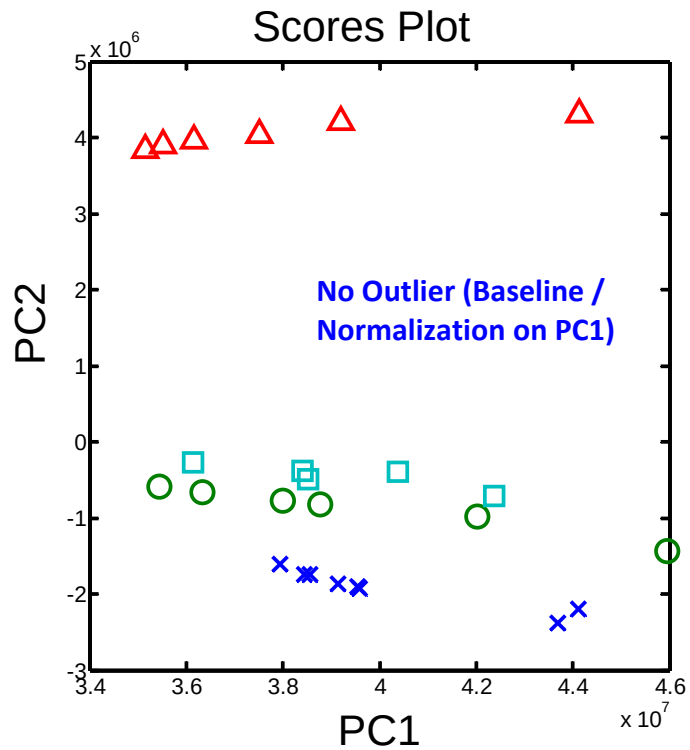
Noise Filtering (Solvent, Spikes, and Outliers)



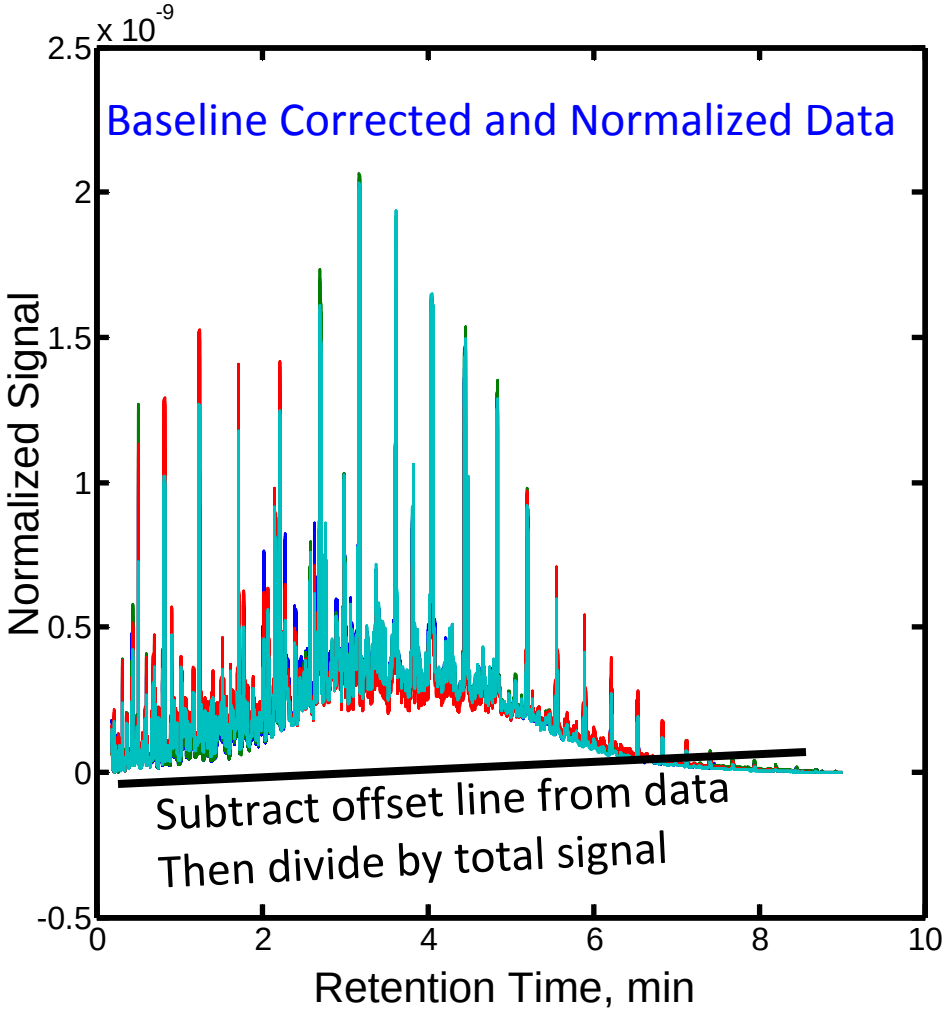


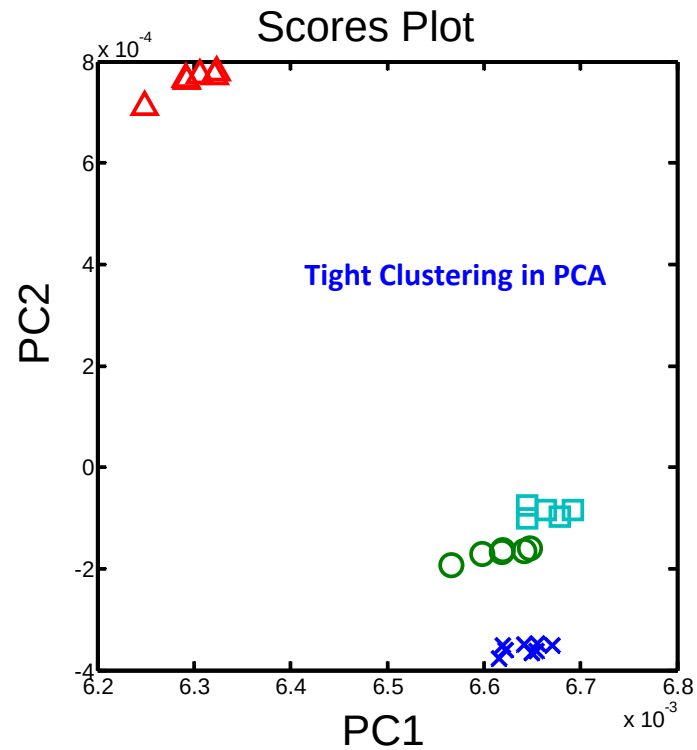
Noise Filtering (Solvent, Spikes, and Outliers)





Noise Filtering (Solvent, Spikes, and Outliers)

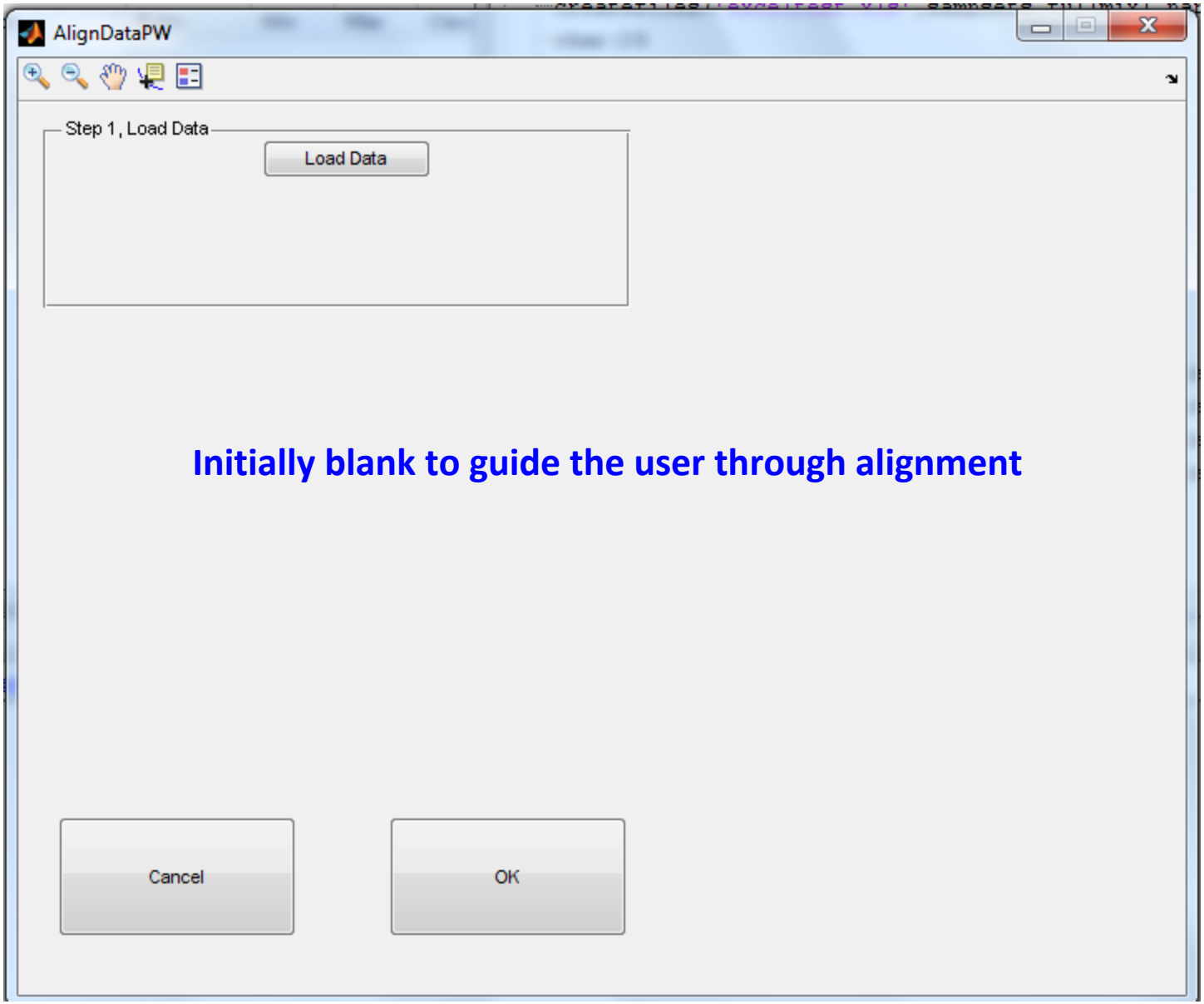




Experimental

- GC Separation
- 3 gasoline sample types
- 15 minute separation
- 6 replicate injections over 2 days
- Misalignment is due to day to day instrument variation

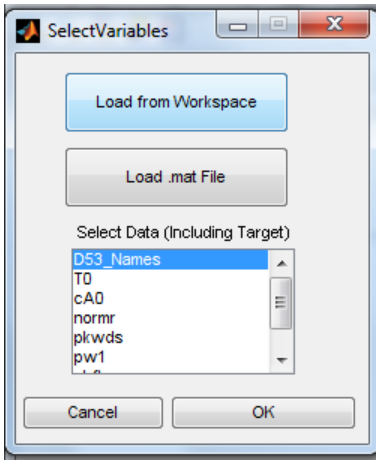
Alignment GUI

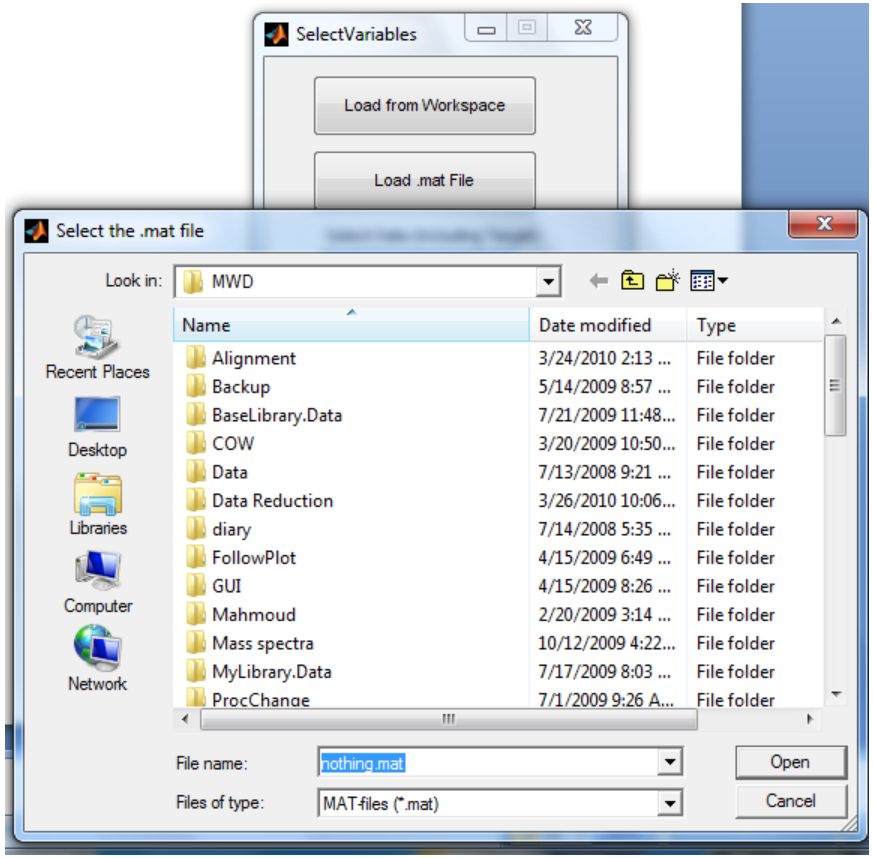


Alignment GUI

Load from file

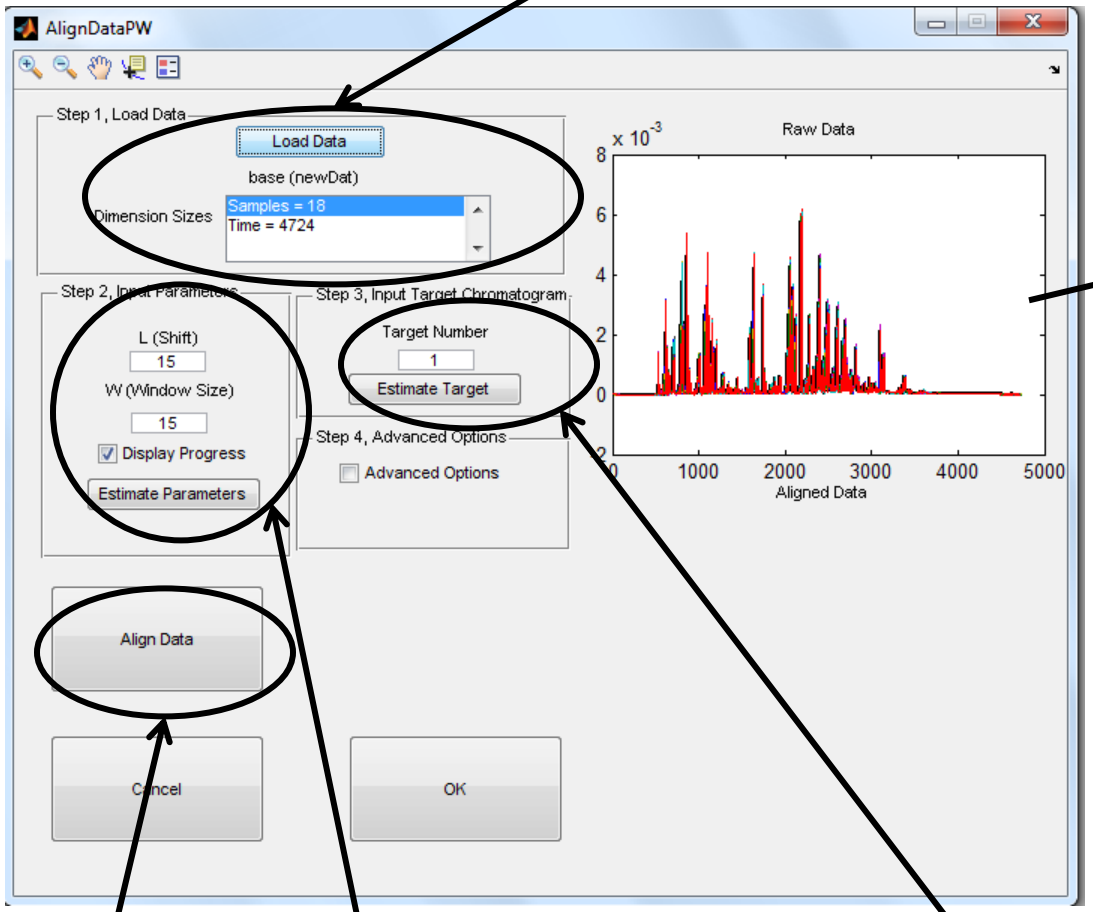
Load from workspace



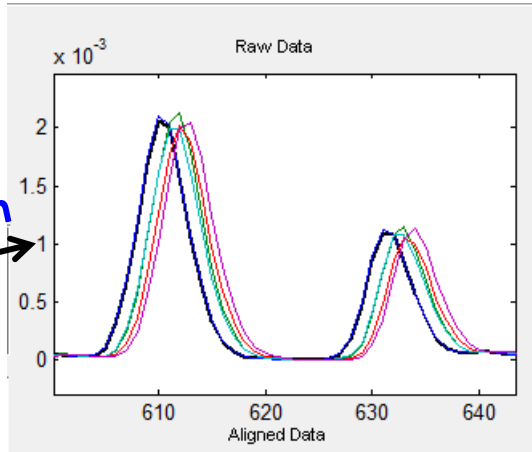


Alignment GUI

Load & View Data Sizes



Zoom & Pan

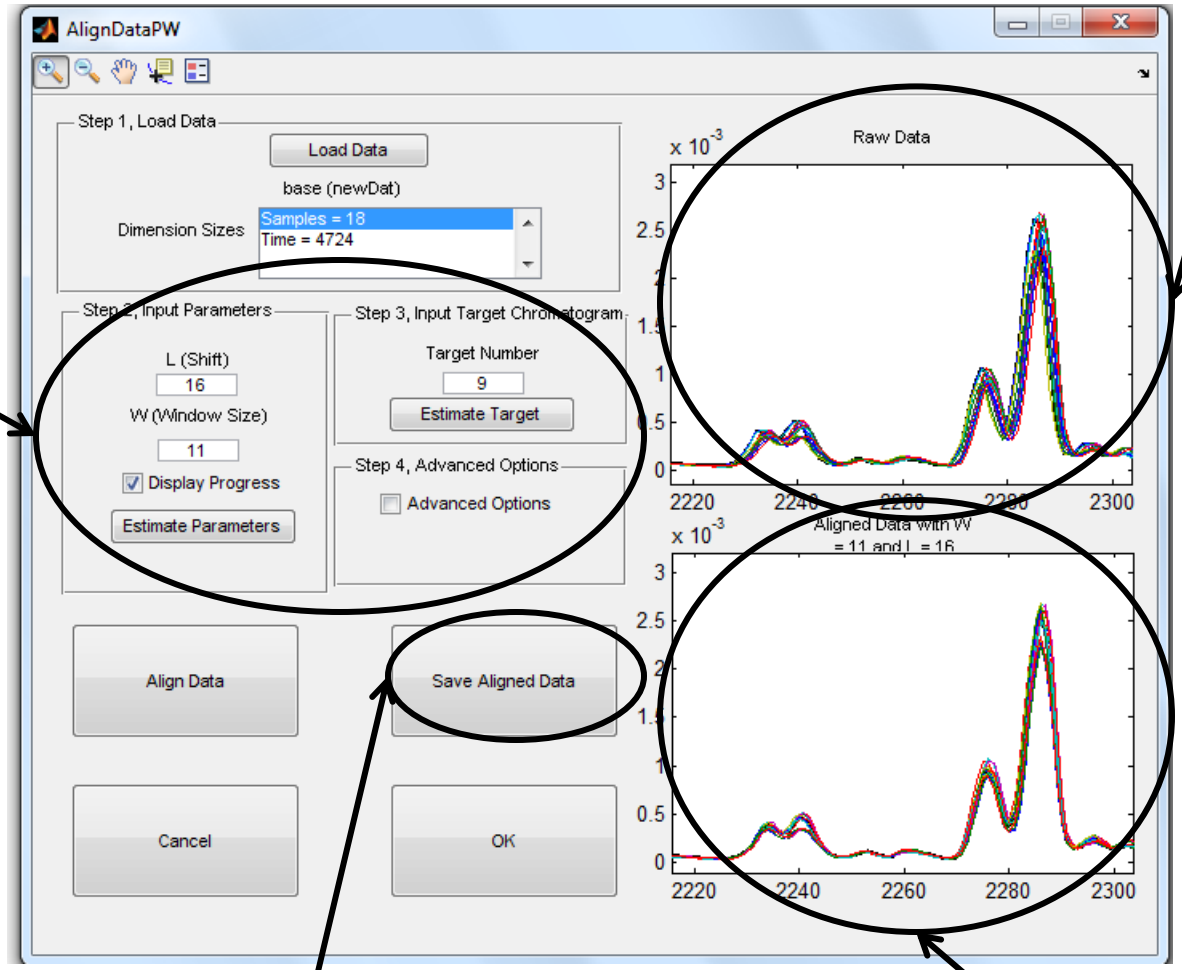


Align Data

Choose or Estimate Parameters

Choose or Estimate Target

Alignment GUI



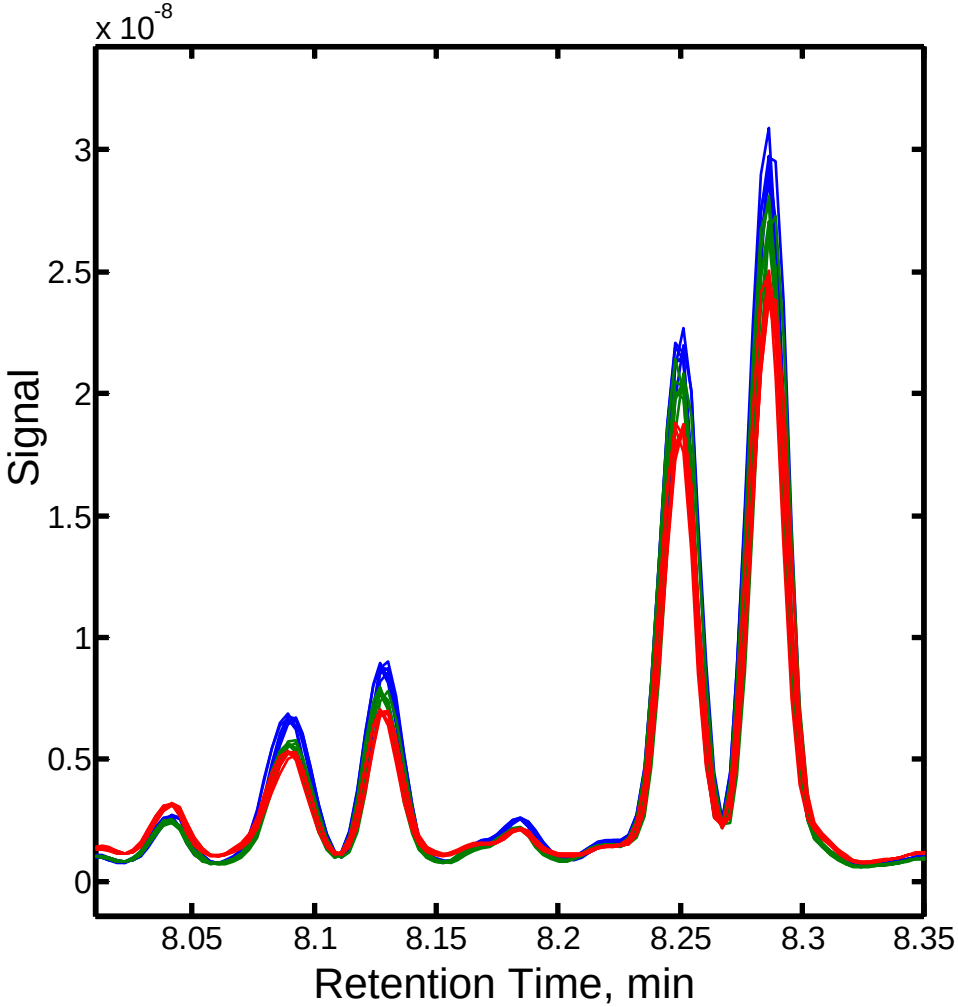
Estimated target and parameters

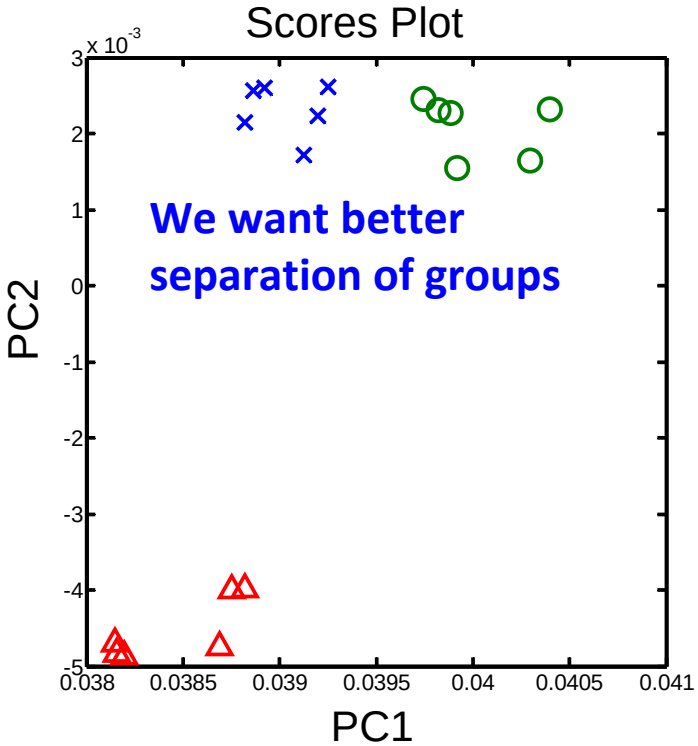
Data before alignment

Now save data to workspace or .mat file

Data after alignment

PCA Classification of Gasoline

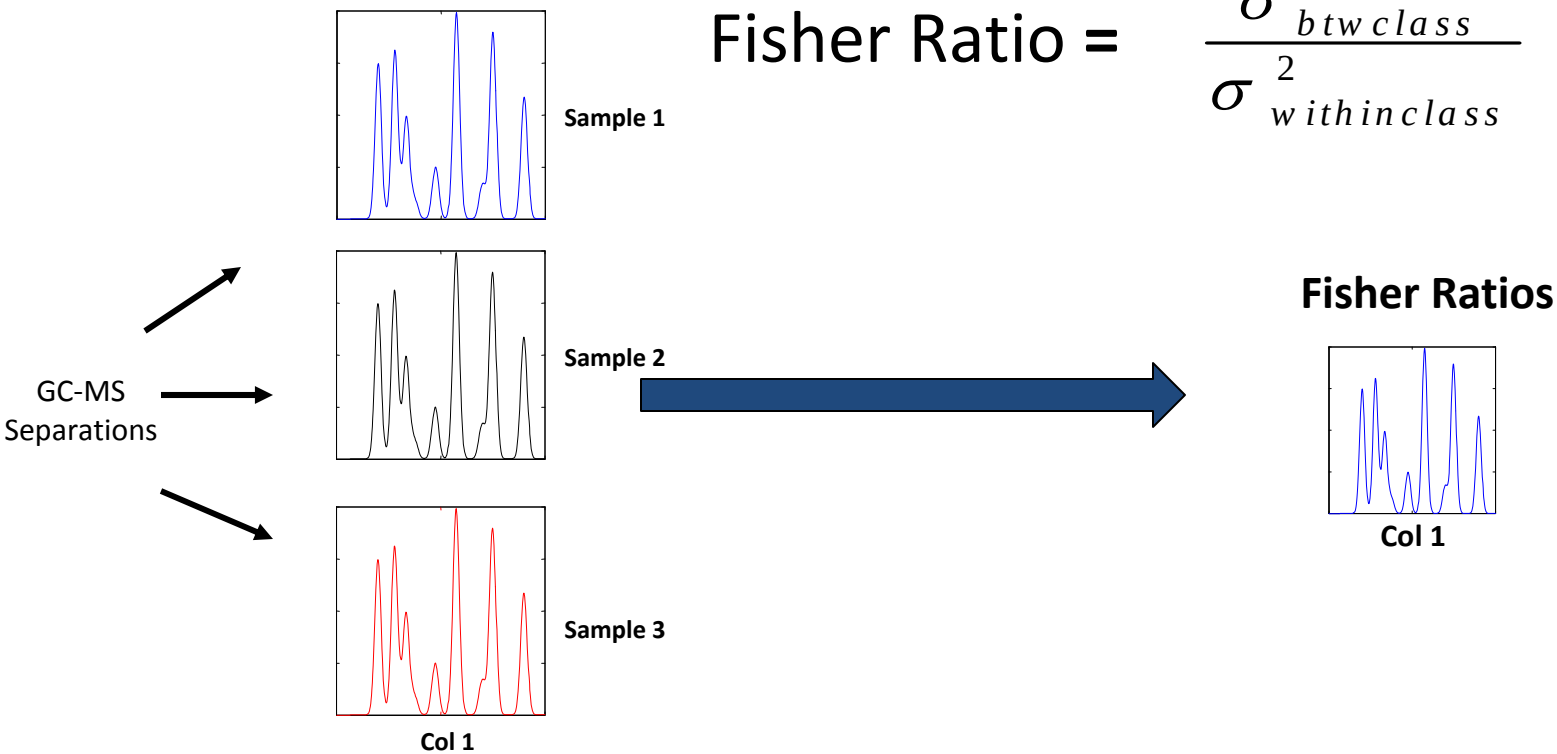




Fisher Ratio Method (F-Ratio)

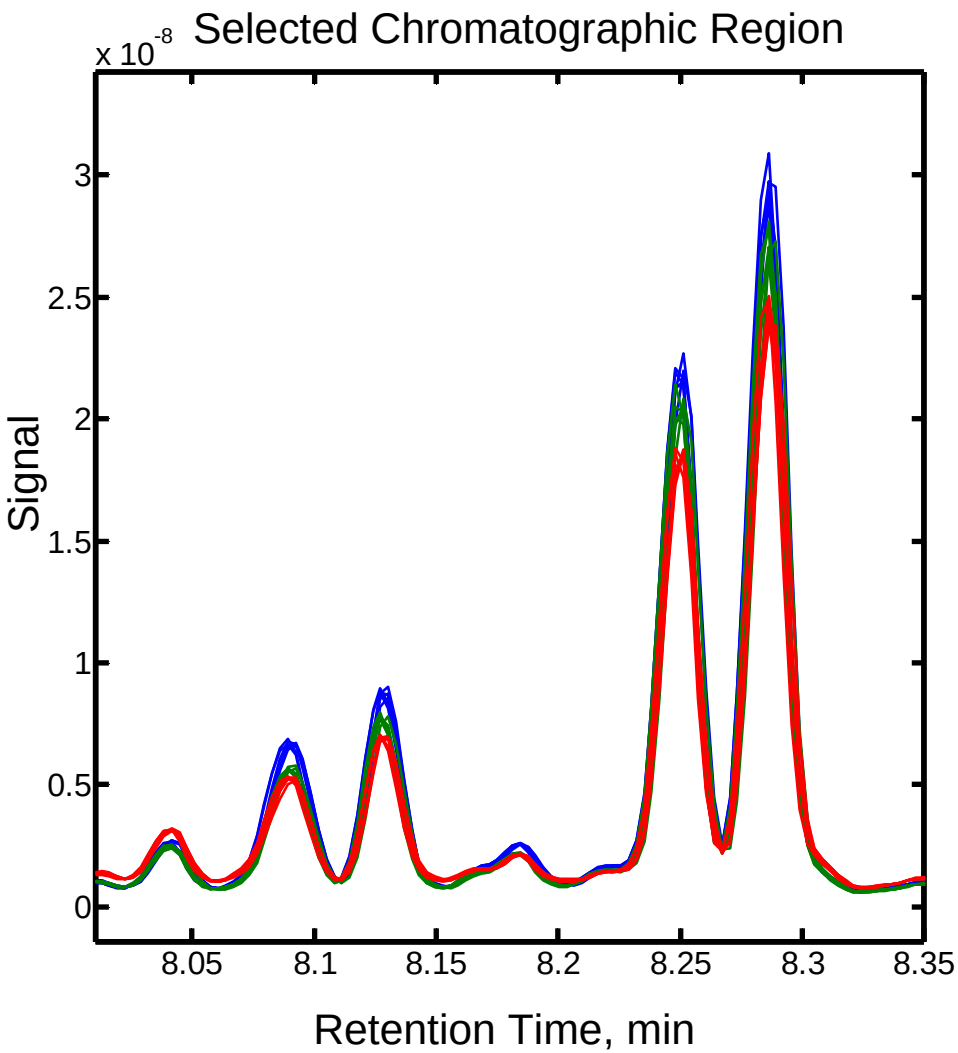
For each mass channel calculate Fisher Ratio at each point in 2D space,

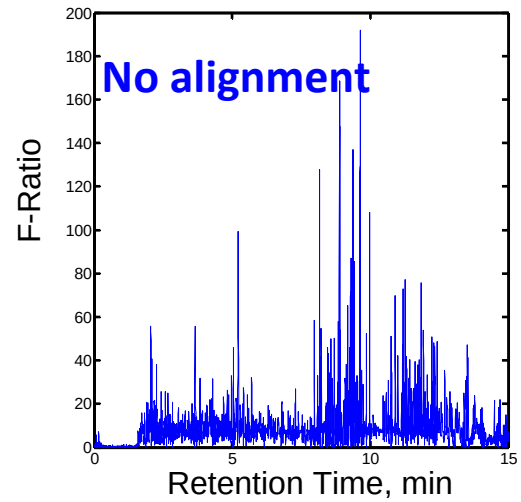
$$\text{Fisher Ratio} = \frac{\sigma^2_{btwclass}}{\sigma^2_{withinclass}}$$

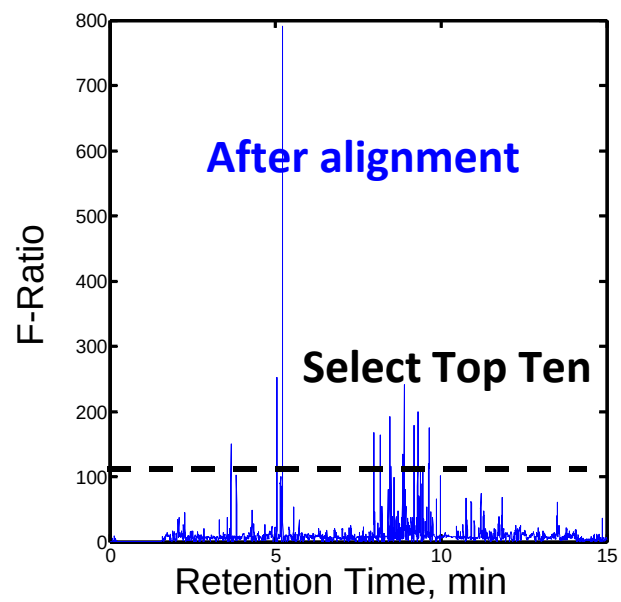


- Works well for samples that have large amounts of within class variance
- Works best when comparing a small number of sample classes

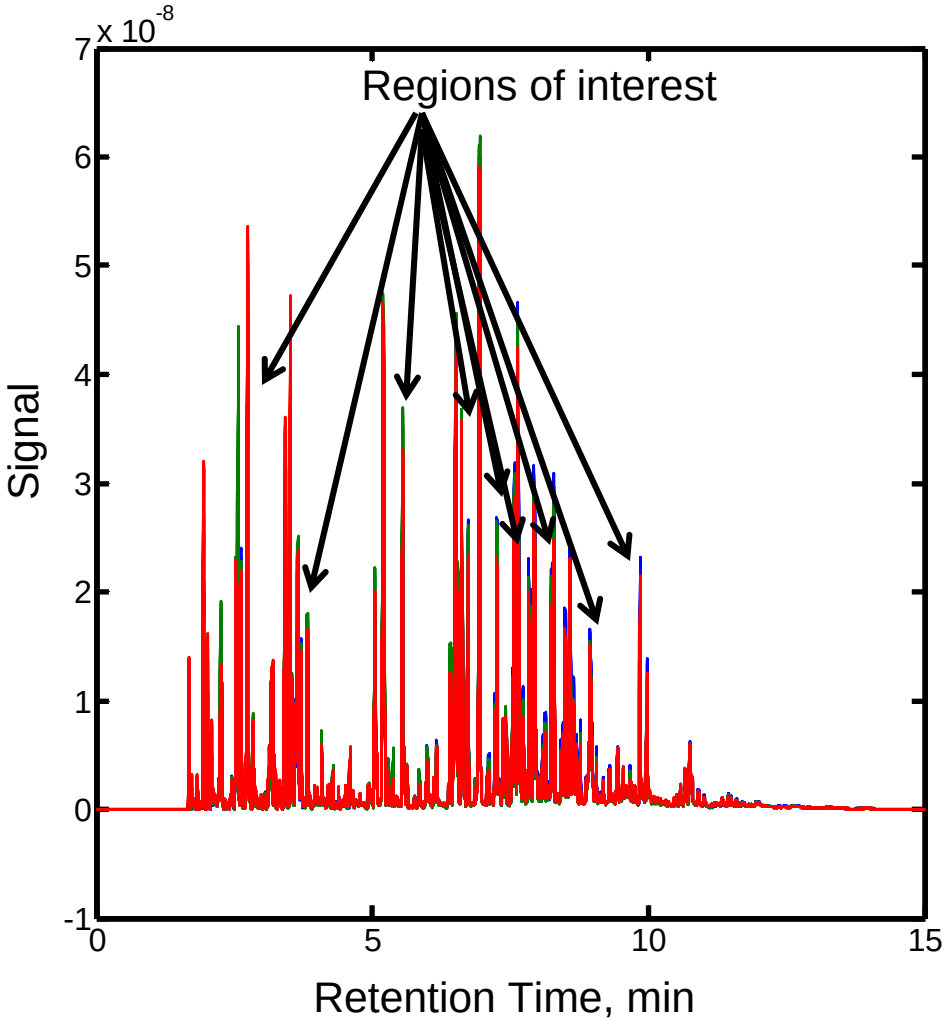
Fisher Ratio Method (F-Ratio)

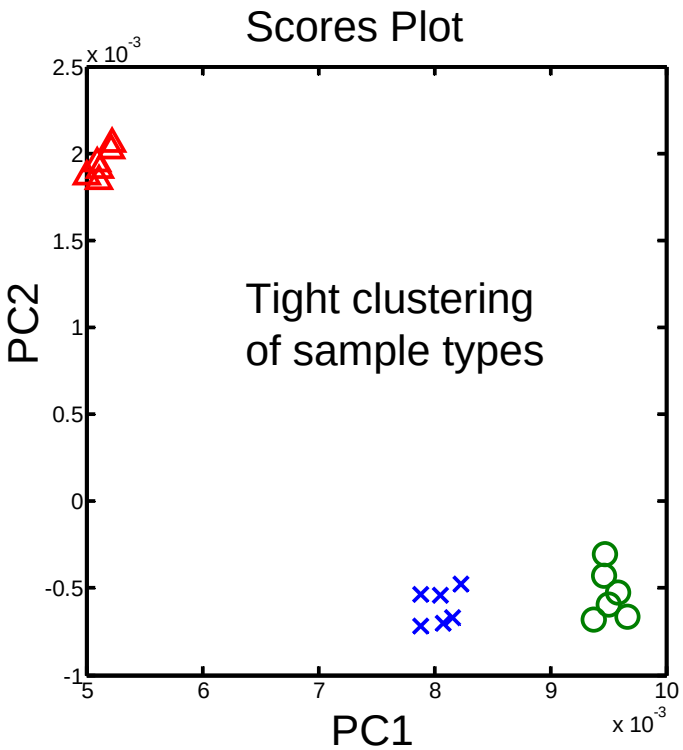






Fisher Ratio Method (F-Ratio)





Summary

- Removal of solvents or artifacts is essential
- Baseline correction is an important step
- Alignment is essential for improving classification
- The F-Ratio algorithm can further improve classification

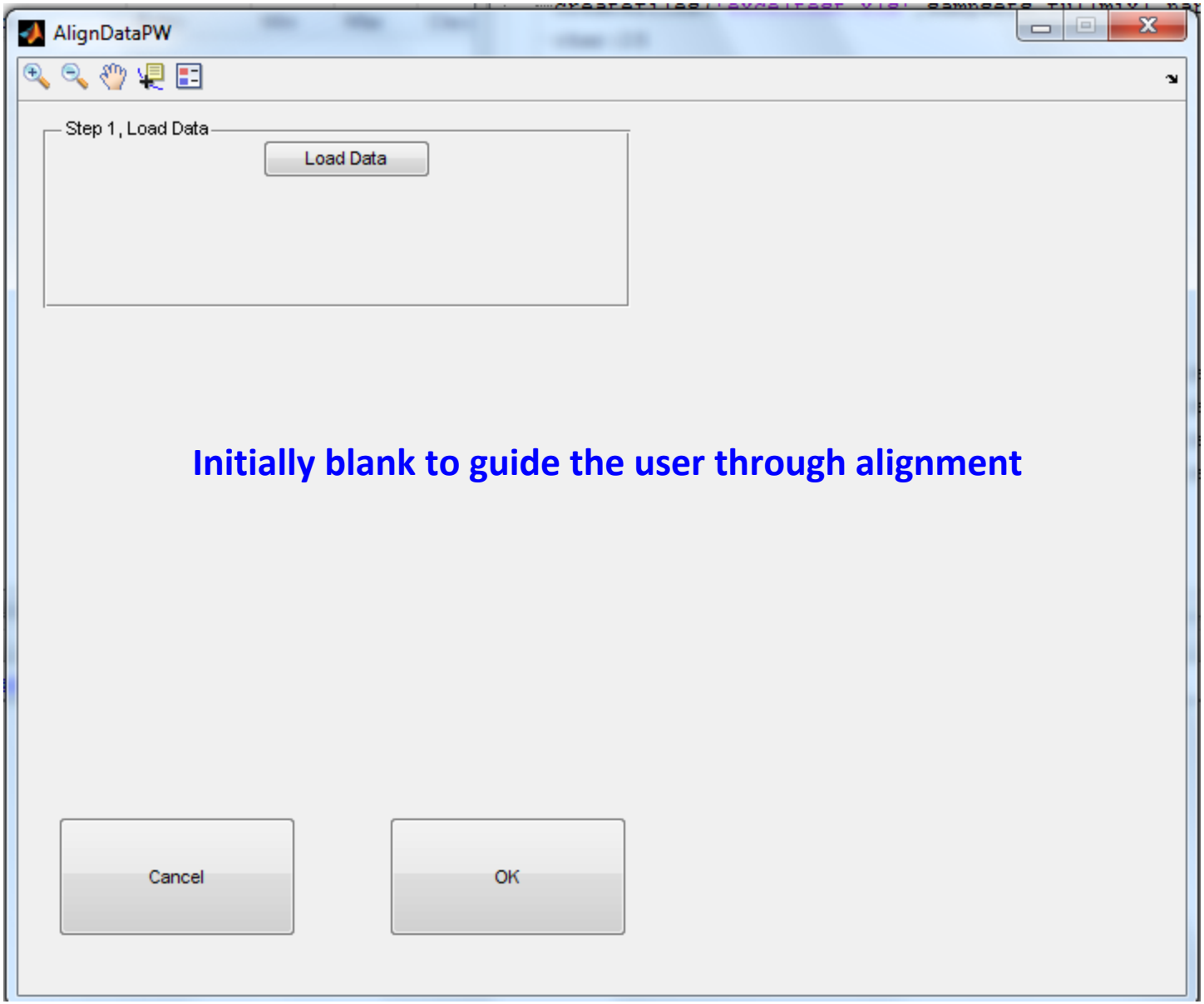
Alignment of GC-MS (2D) Data and Issues

- Removal of artifacts and solvent peaks
- Baseline correction and normalization
- Alignment
- Improving PCA
  - F-Ratio
- PARAFAC

Experimental

- GC-MS Separation
- 3 gasoline sample types
- 15 minute separation
- 6 replicate injections over 2 days
- Misalignment is due to day to day instrument variation

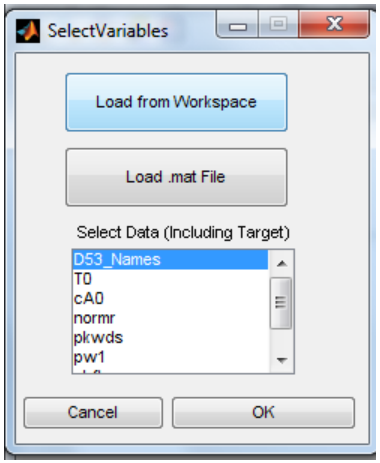
Alignment GUI

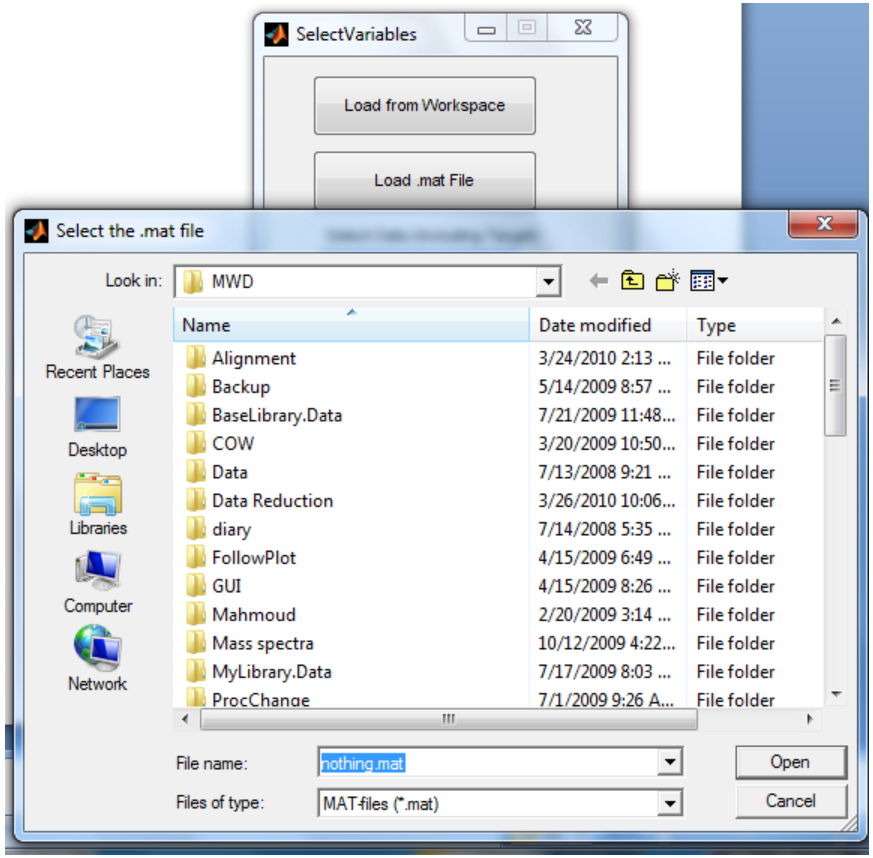


Alignment GUI

Load from file

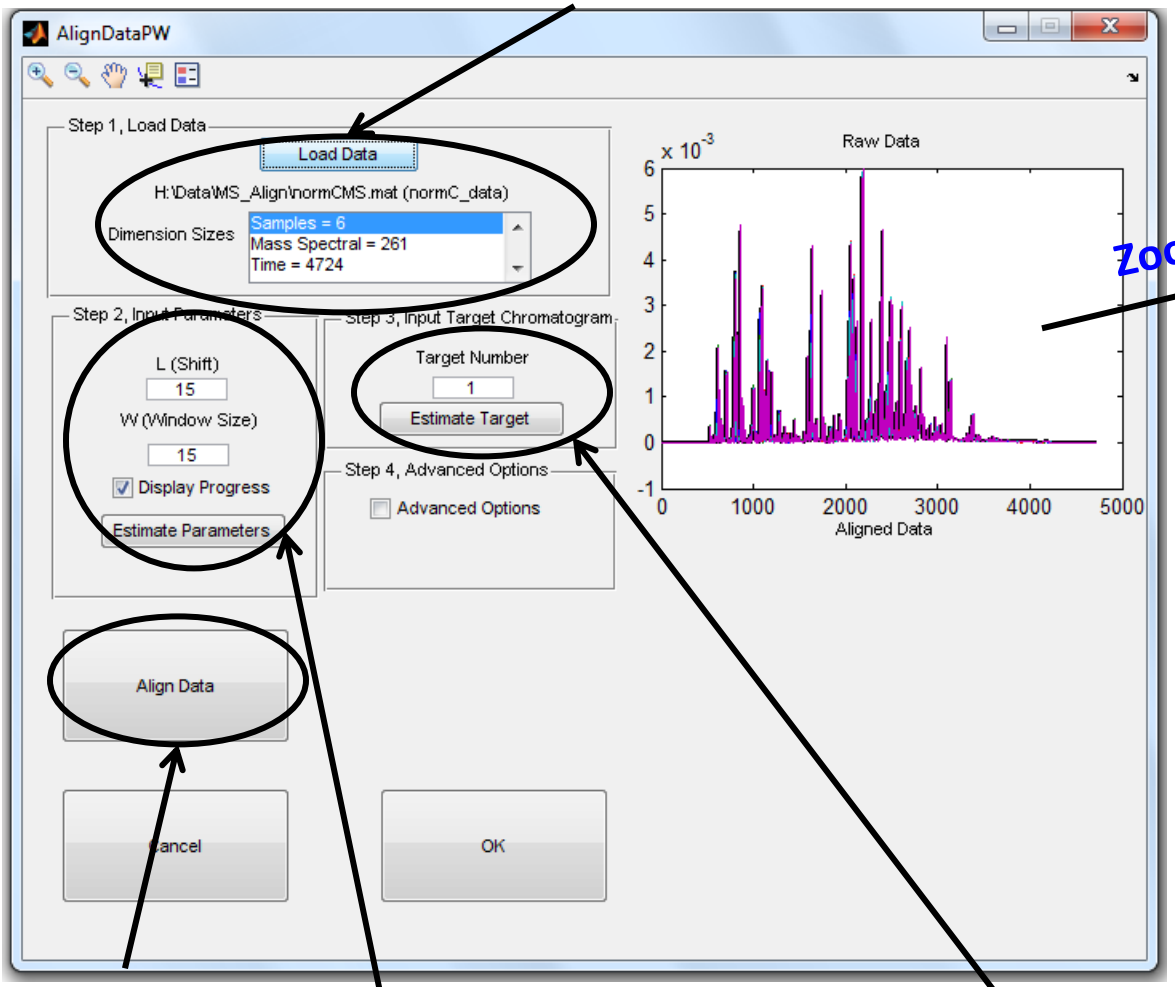
Load from workspace



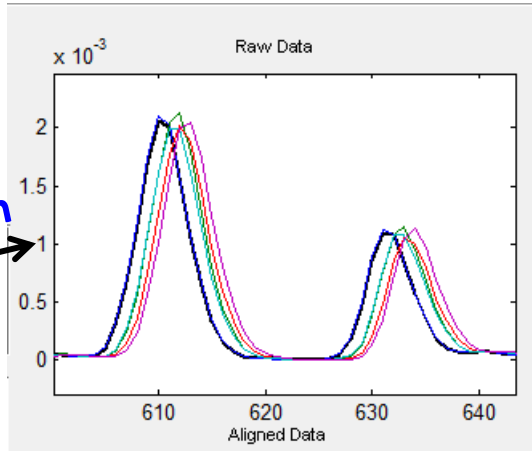


Alignment GUI

Load & View Data Sizes



Zoom & Pan

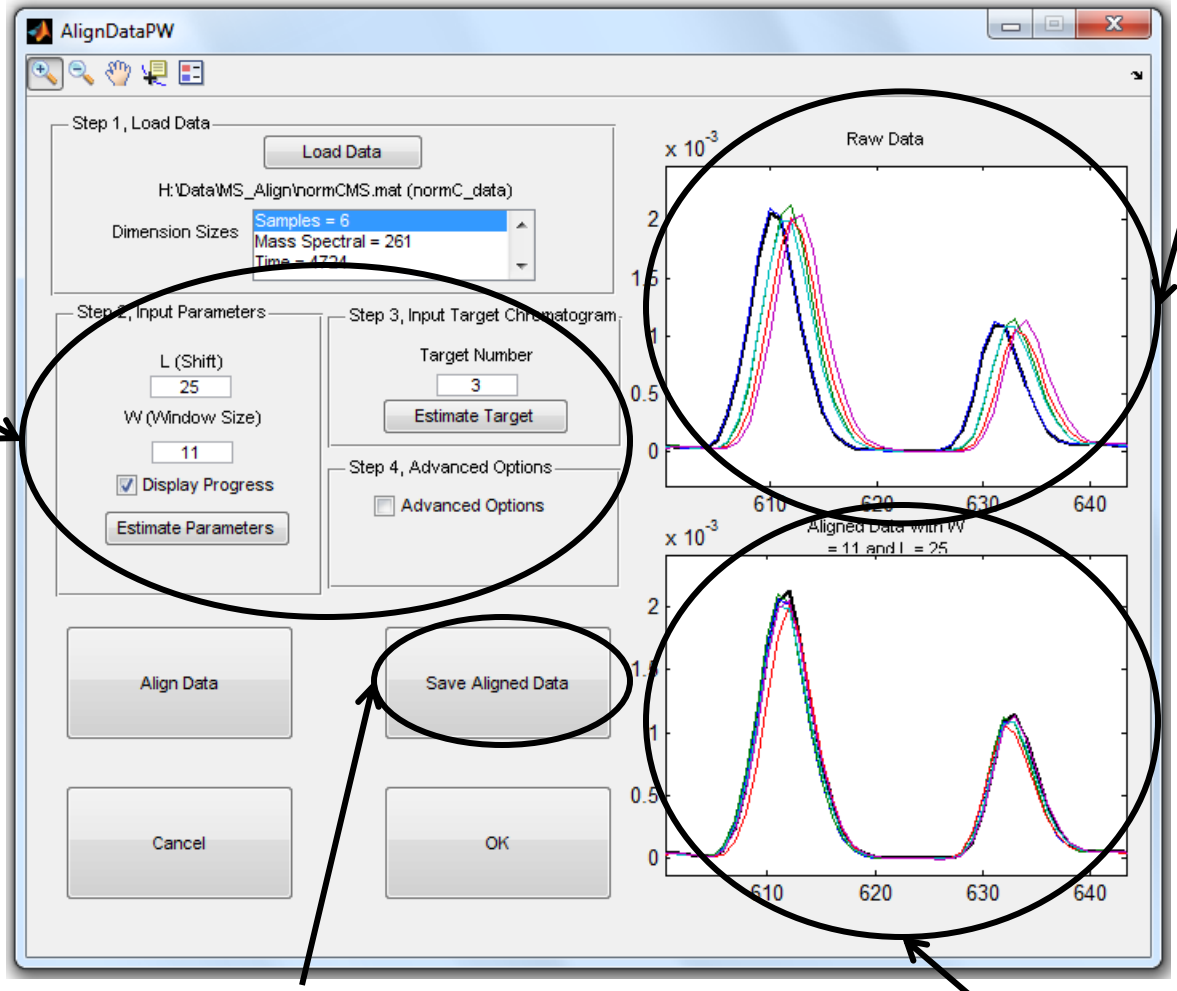


Align Data

Choose or Estimate Parameters

Choose or Estimate Target

Alignment GUI



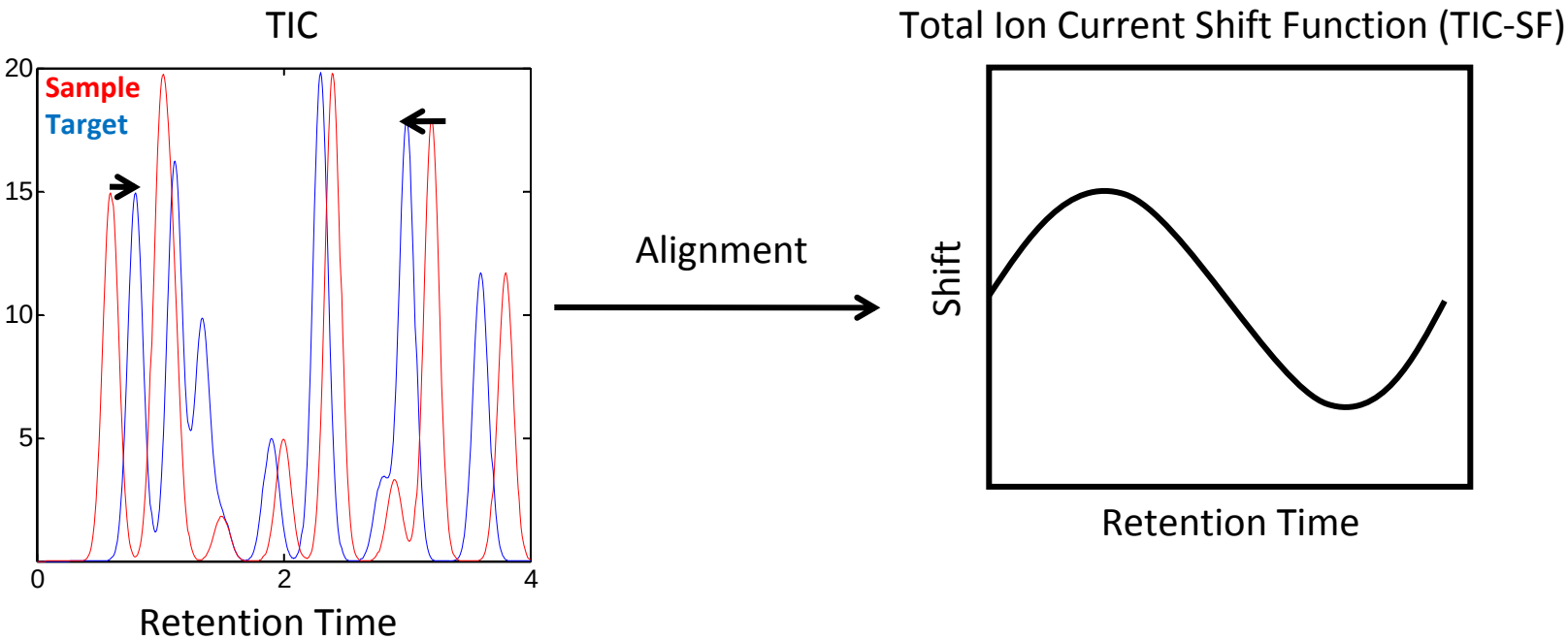
Estimated target and parameters

Data before alignment

Now save data to workspace or .mat file

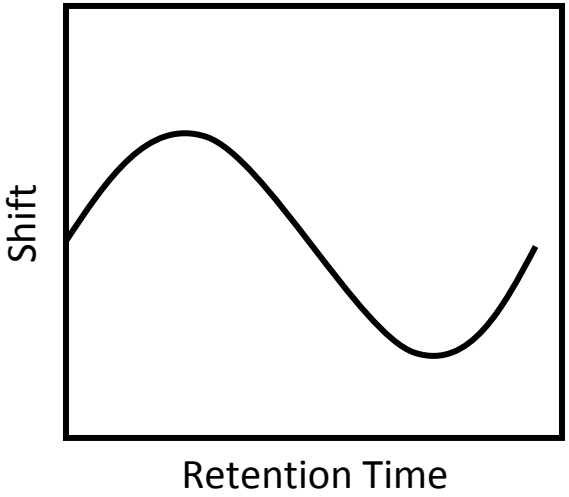
Data after alignment

Alignment

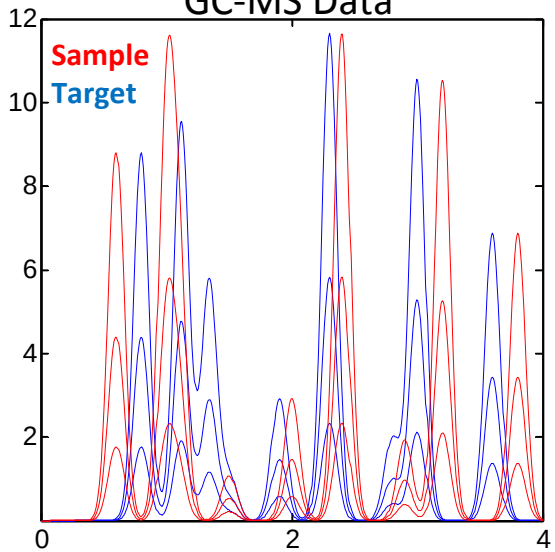

TIC
Sample
Target
20
15
10
5
0
2
4
Retention Time
Alignment
Total Ion Current Shift Function (TIC-SF)
Shift
Retention Time

Alignment

Total Ion Current Shift Function (TIC-SF)

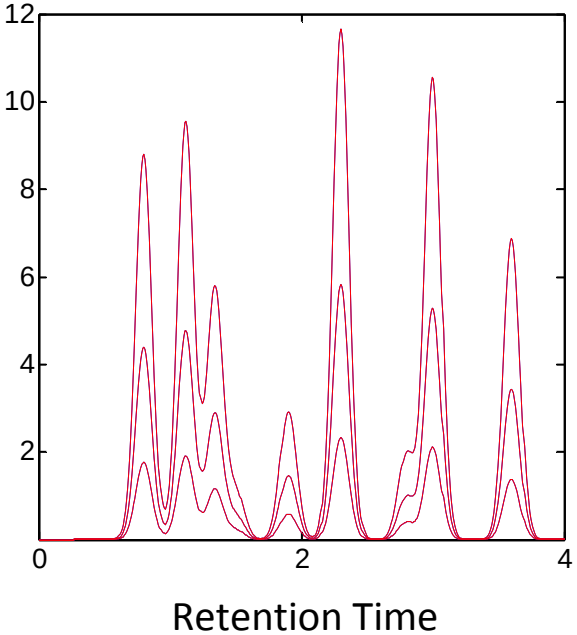


GC-MS Data

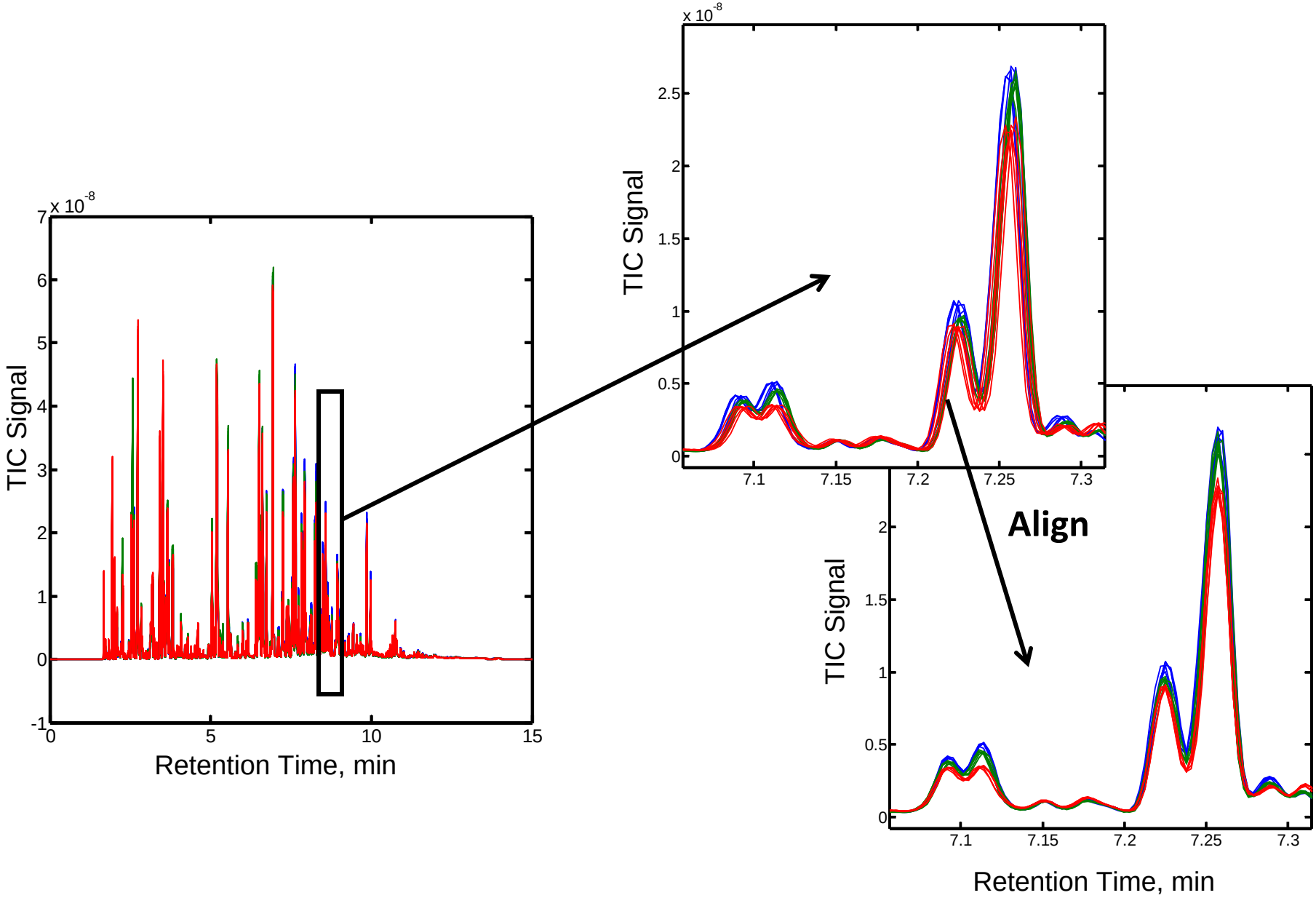


Apply alignment

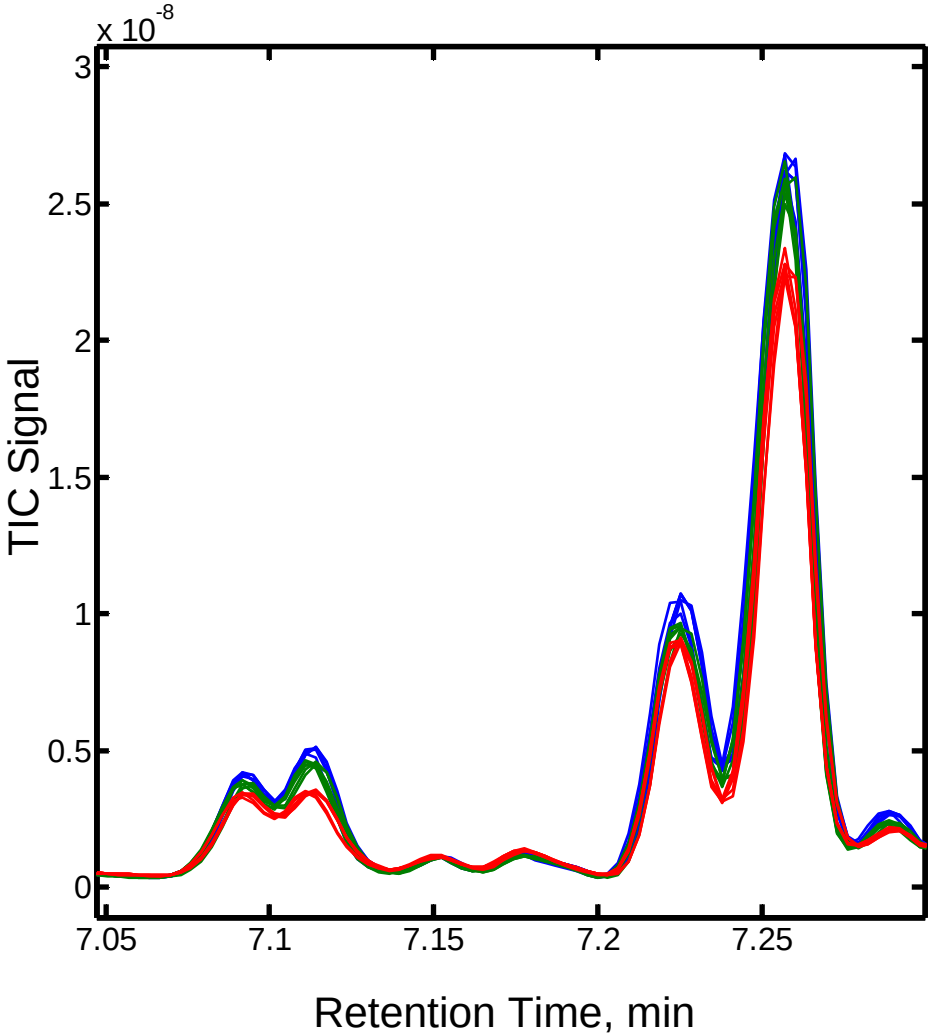
GC-MS Data

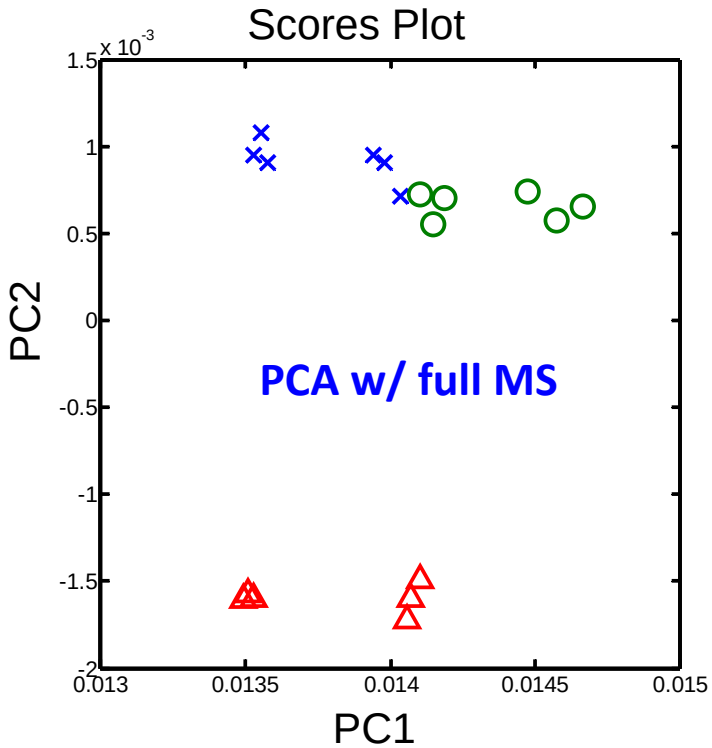


Alignment



Classification of Full MS Data

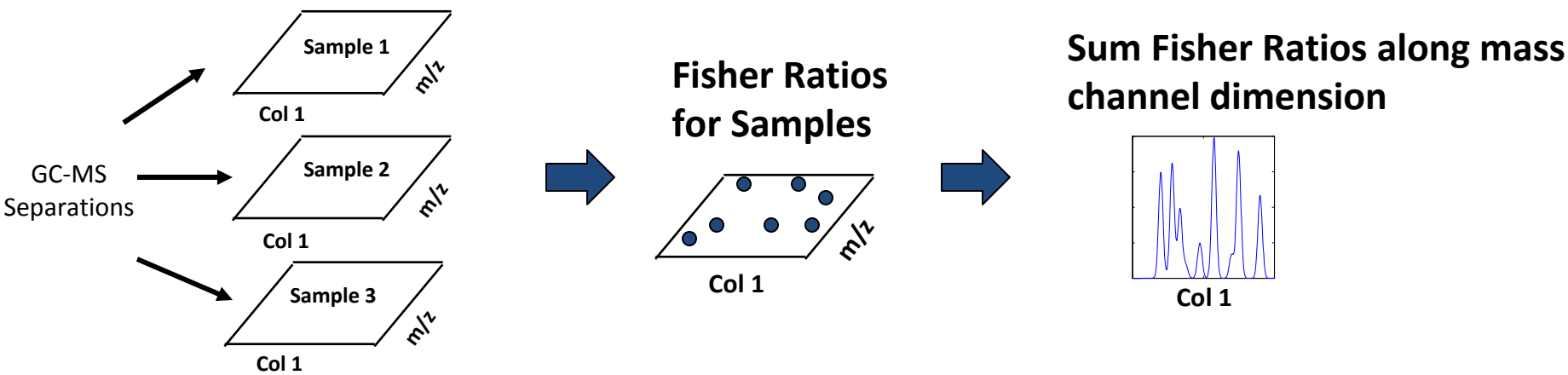




Fisher Ratio Method

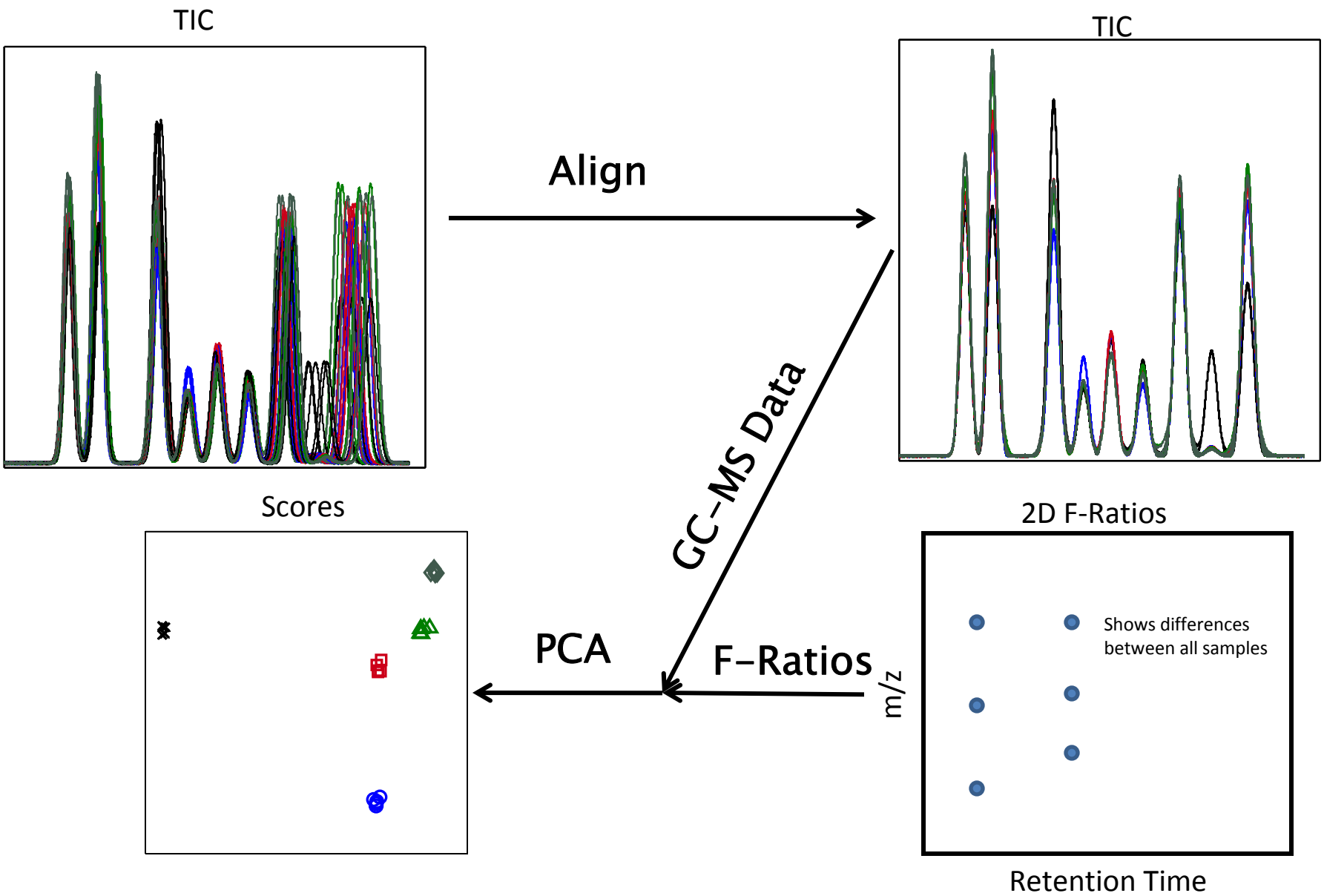
For each mass channel calculate Fisher Ratio at each point in 2D space,

$$\text{Fisher Ratio} = \frac{\sigma^2_{btwclass}}{\sigma^2_{withinclass}}$$

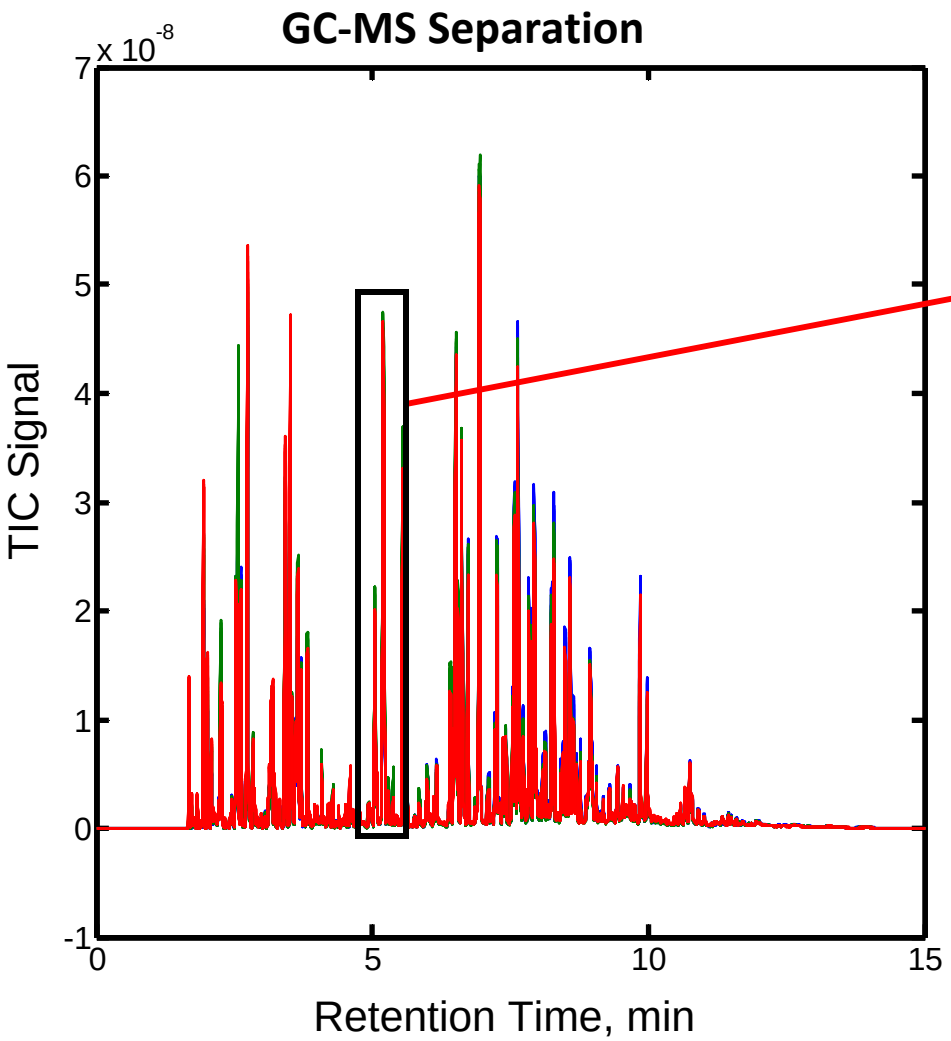


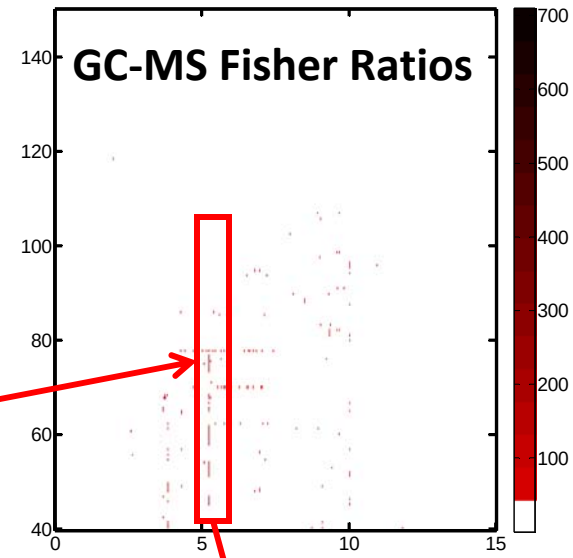
- Works well for samples that have large amounts of within class variance
- Works best when comparing a small number of sample classes

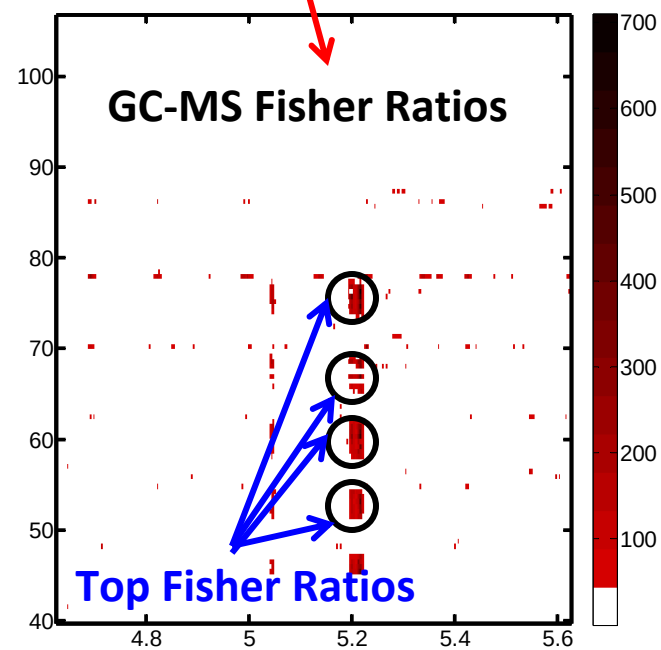
Simulated Example Using F-Ratios for PCA



F-Ratios


GC-MS Separation
x 10^-8
TIC Signal
Retention Time, min

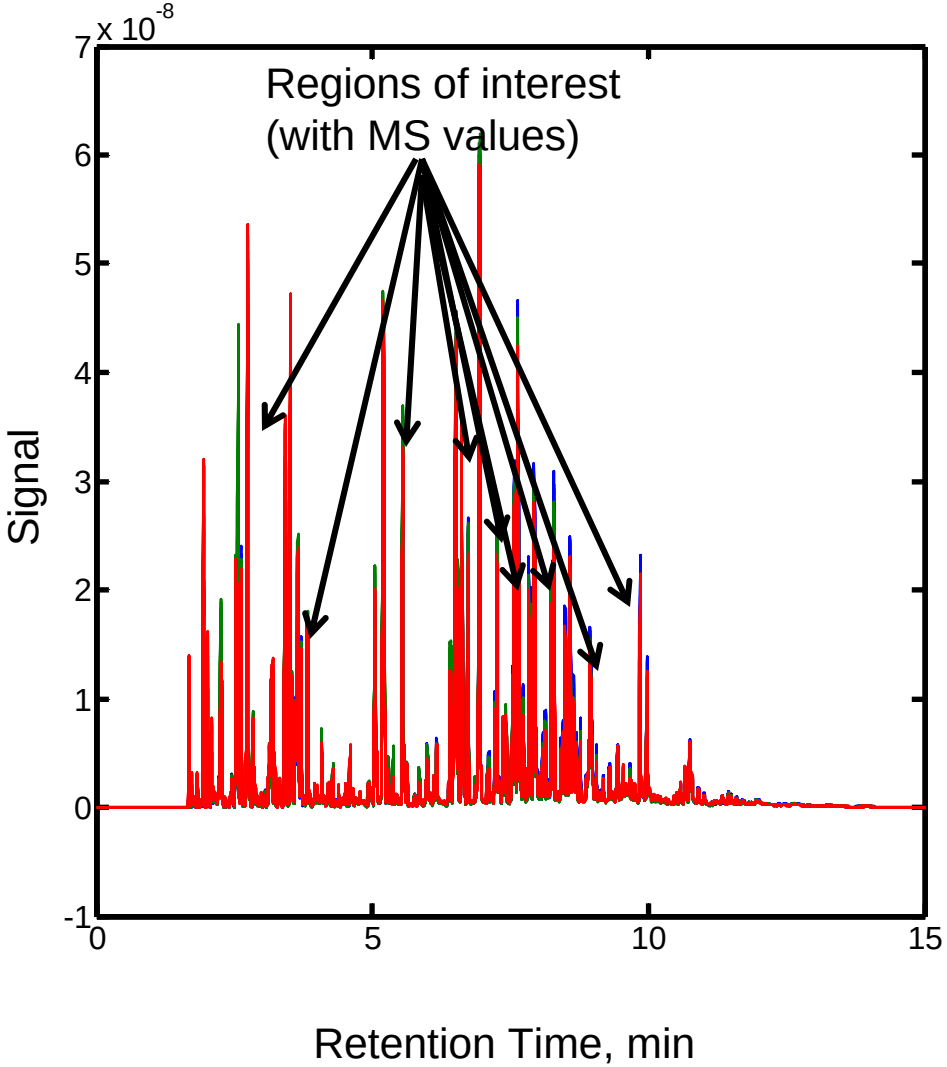

GC-MS Fisher Ratios

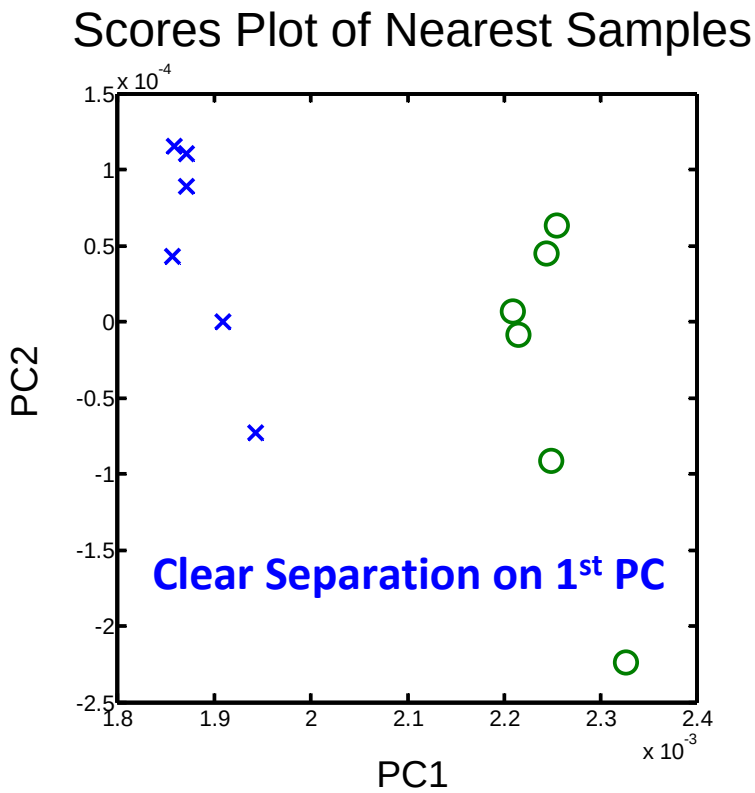

GC-MS Fisher Ratios
Top Fisher Ratios

Procedure

- Select masses using mass spectral information
- Select time regions using F-Ratios
- Combine to reduce the data set
- Improve results of PCA

Classification of Full MS Data





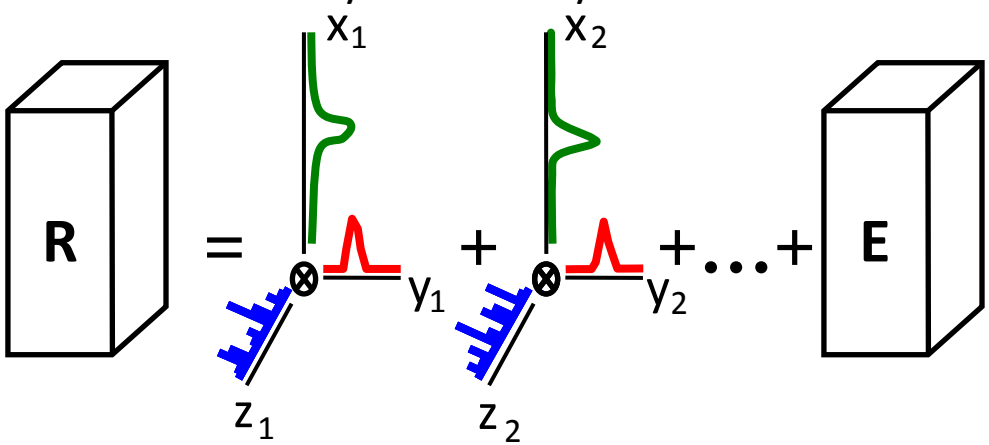
Quantification of GC-MS Data

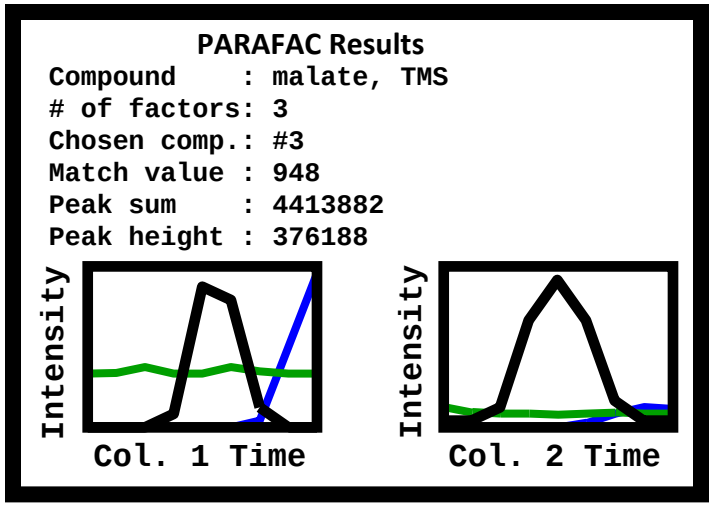
- Use aligned, baseline corrected and normalized data
- Use PARAFAC of small regions for analysis
  - Match values to mass spectra
  - Peak Sums
  - Peak Profiles

Quantification

- Target analyte Parallel Factor Analysis (PARAFAC) isolates the pure component peak and mass spectral information from overlapping peaks and background for both *identification* and *quantification.*

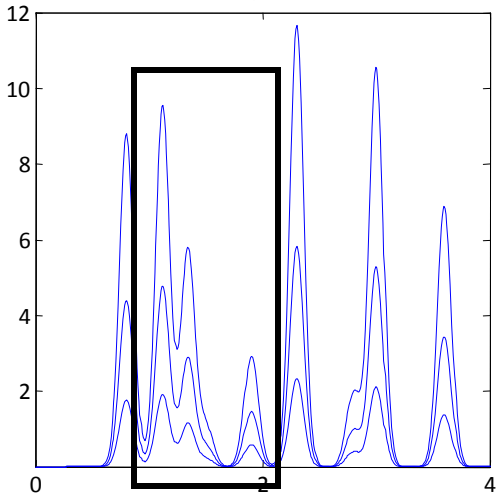
Data Matrix = Analyte 1 + Analyte 2 + ... Noise



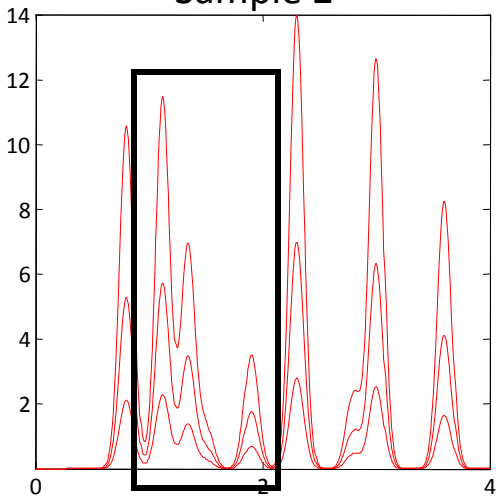


PARAFAC for GC-MS Data

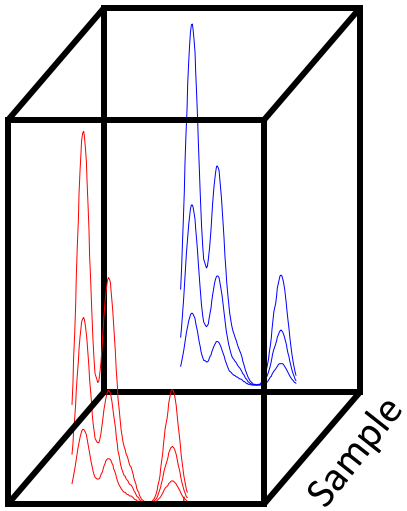
Sample 1



Sample 2

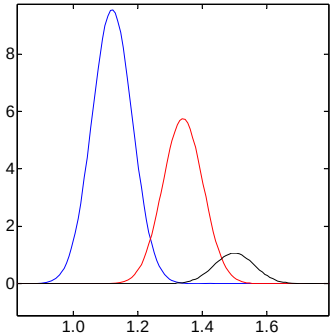


Stack Samples

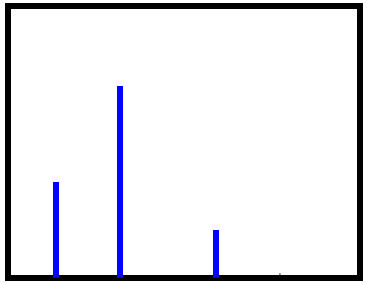


Run PARAFAC

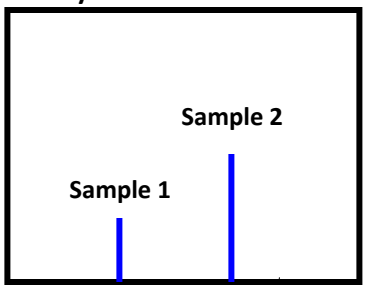
Peak Profile



Mass Spectrum



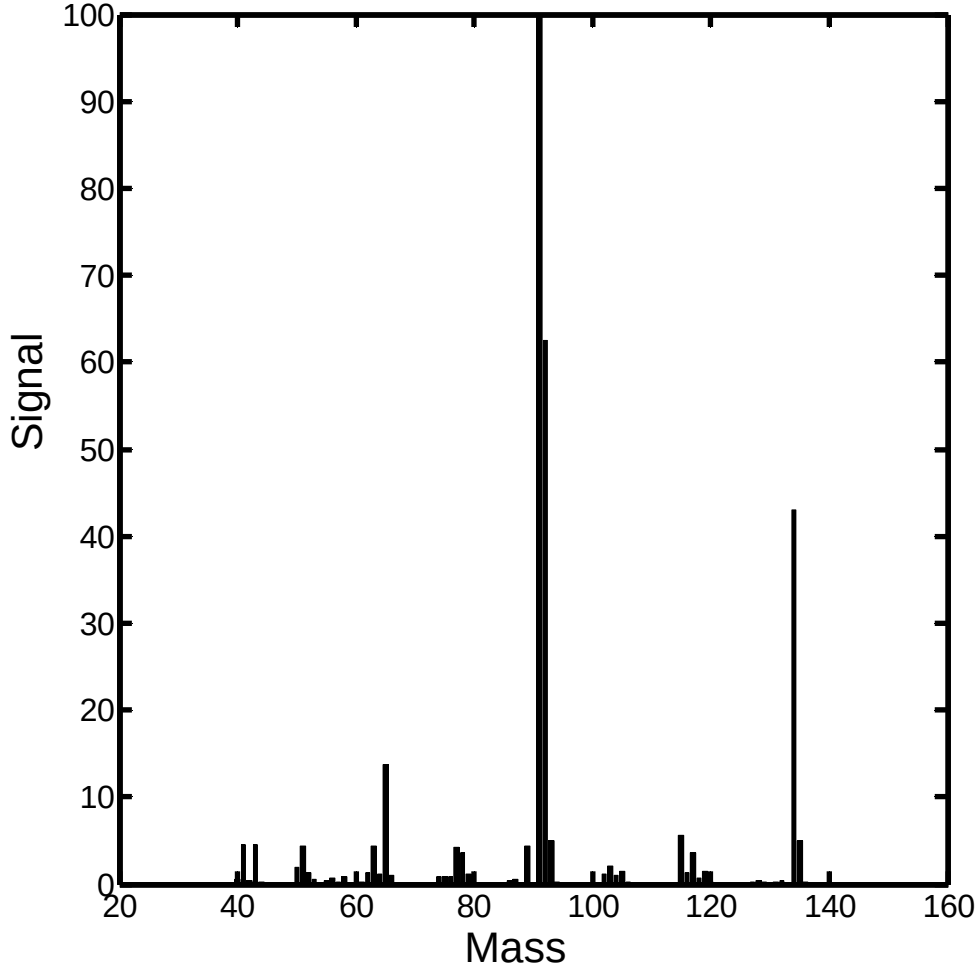
Analyte Concentration



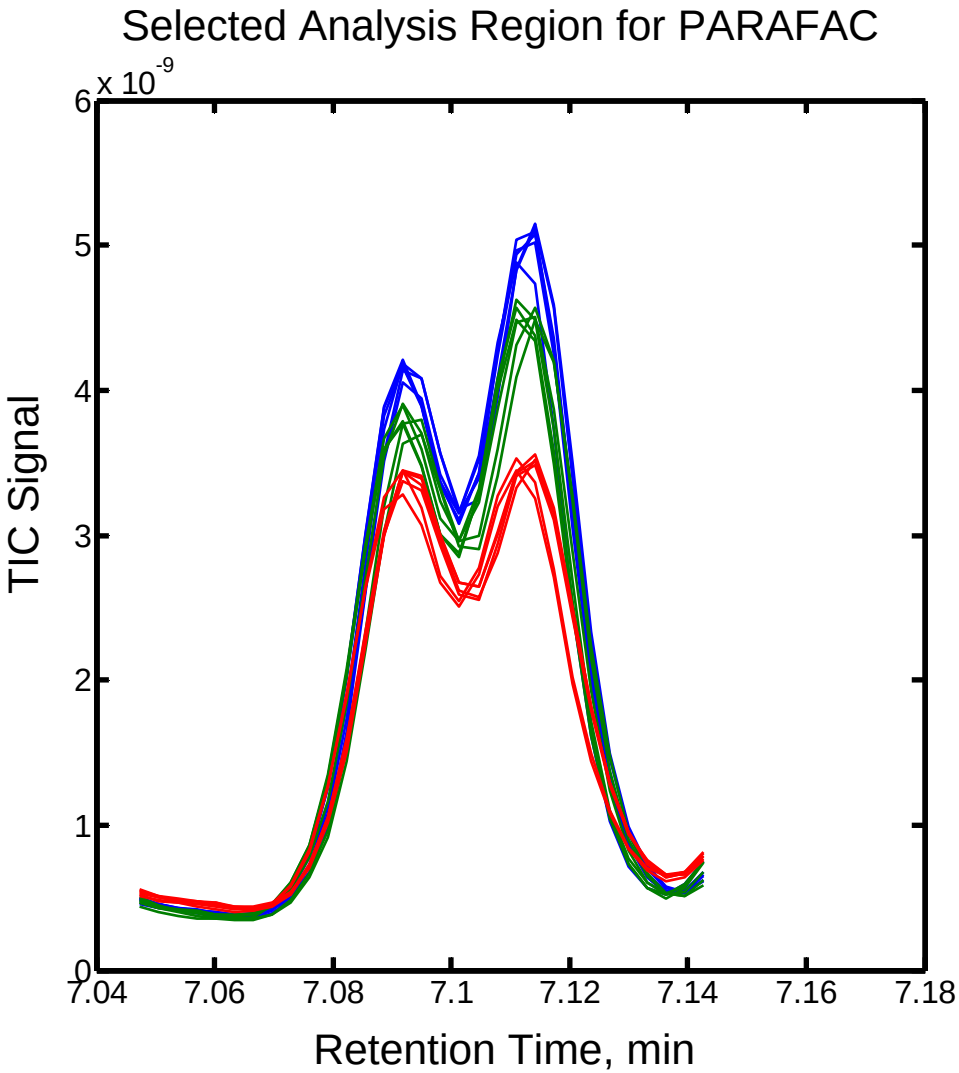
Experimental

- GC-MS Separation
- 3 gasoline sample types
- 15 minute separation
- 6 replicate injections over 2 days
- Misalignment is due to day to day instrument variation
- Looking for isobutyl benzene in the gasoline

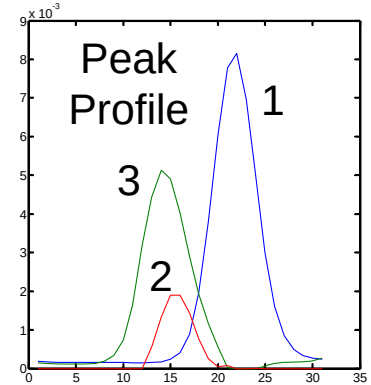
Isobutyl Benzene Spectrum



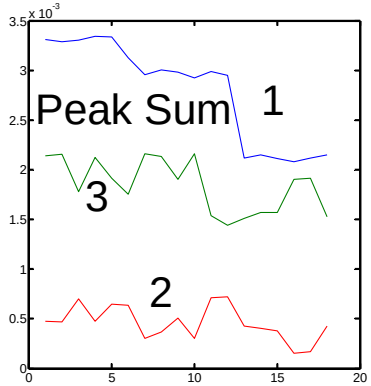
PARAFAC Results for GC-MS Data



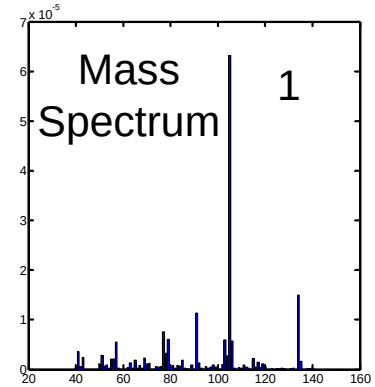
Qualitative Info



Quantitative Info



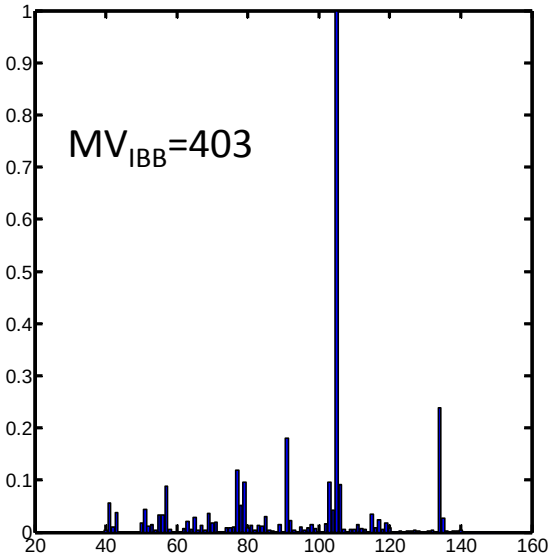
Qualitative Info



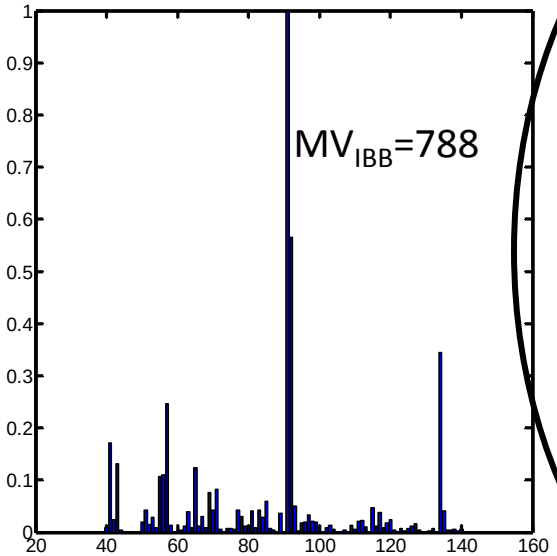
PARAFAC Results for GC-MS Data

Best match to isobutyl benzene

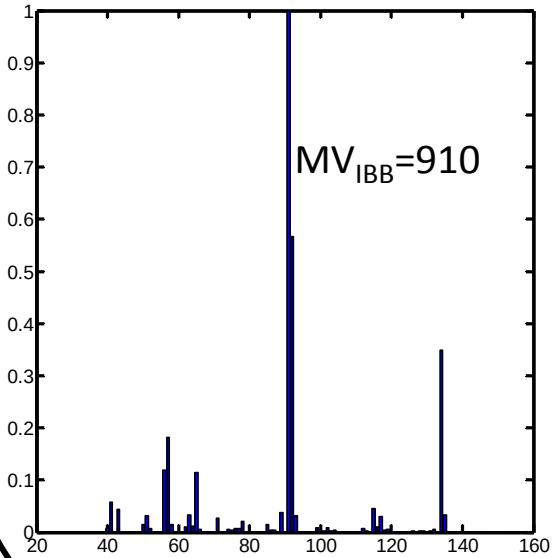
Component 1



Component 2



Component 3



Summary

- Mass spectral chromatographic data should be aligned
- An available alignment algorithm is able to align 1-D and 2-D data
- F-Ratio methods can be used to improve classification
- PARAFAC can be effectively used to separate overlapped analytes