Data Processing in a Fast GC World

The case for unattended multivariate analysis

"We want reliable, intelligent answers driven by analytical technologies that are not perfectly reliable or intelligent."
- Lloyd Colegrove, Dow Chemical IFPAC 2013

Brian Rohrback, Infometrix, Inc, Bothell, WA

Gulf Coast Conference October 15, 2014

Infometrix

Two Ways to Use Chromatography

1. Quantitative Analysis
   - Provide a means of accurately quantitating a small number of compounds.
   - Predicting a physical property or system parameter
   - Unbundling a mixture
2. Qualitative Analysis
   - Evaluate a pattern of components to determine if the mixture is within specifications.

Another role for chemometrics

With the increase in speed, we need to automate the assessment of the chromatographic data such that samples behaving normally are accepted, but any problem is noted whether it be

- a raw material input deficiency,
- a process problem, or
- an instrument problem

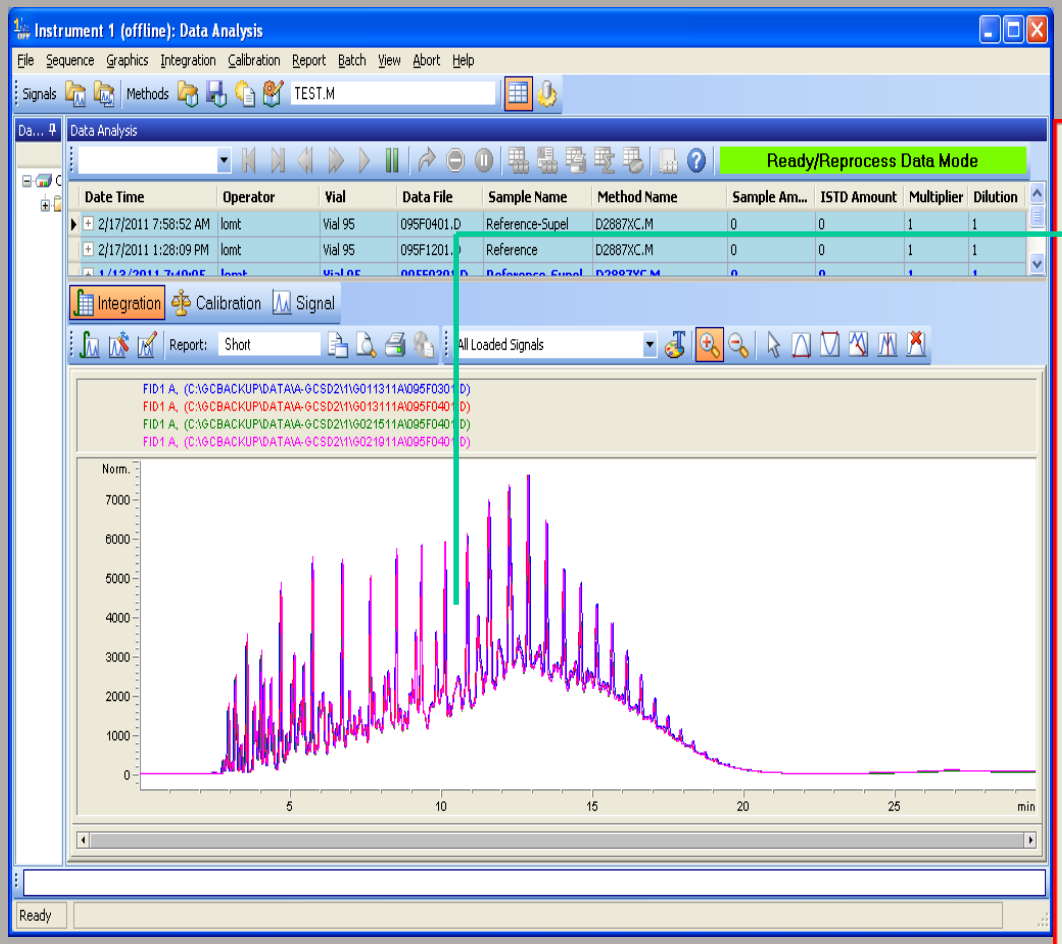
Re-thinking Chromatography

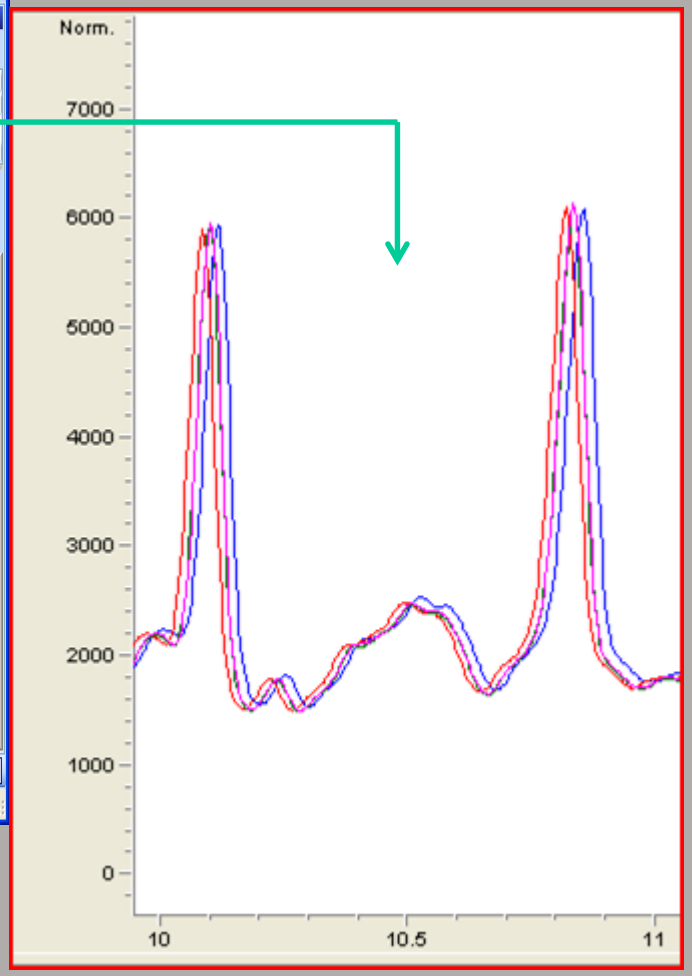
- 35 years ago, fused silica made GC more useful by quintupling the number of technicians that could be independent running chromatographic equipment in non-trivial applications.
- I believe chemometrics combined with fast GC is generating a fundamental change in how we deploy GCs.
  - *Simplifying Calibration*
  - *Process Control*
  - *Global Databases*

Processing Whole Chromatograms

- Chromatograms will show an x-axis (retention time) shift for a variety of reasons:
  - Changing columns
  - Aging columns
  - Different instruments
  - Degradation of the column over time
- We need to eliminated the retention time variability to improve the precision of the assessments.

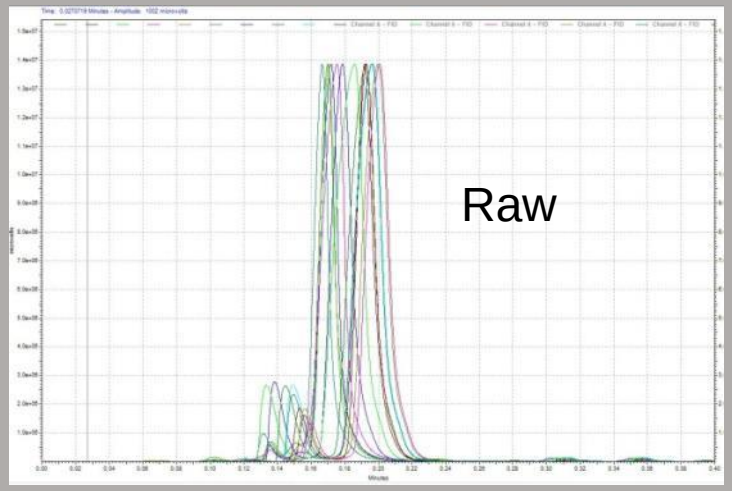
Retention Time Misalignment is a General Chromatography Problem

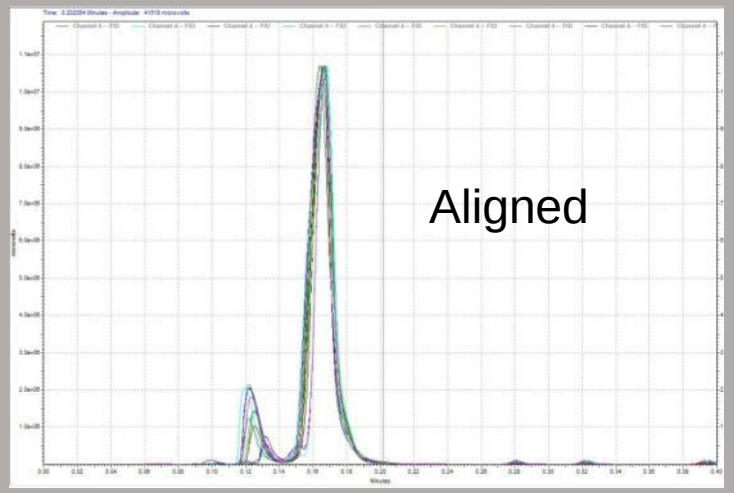


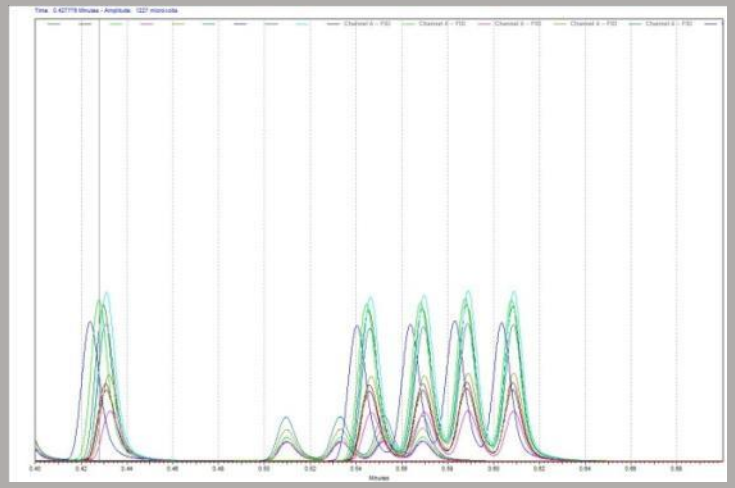


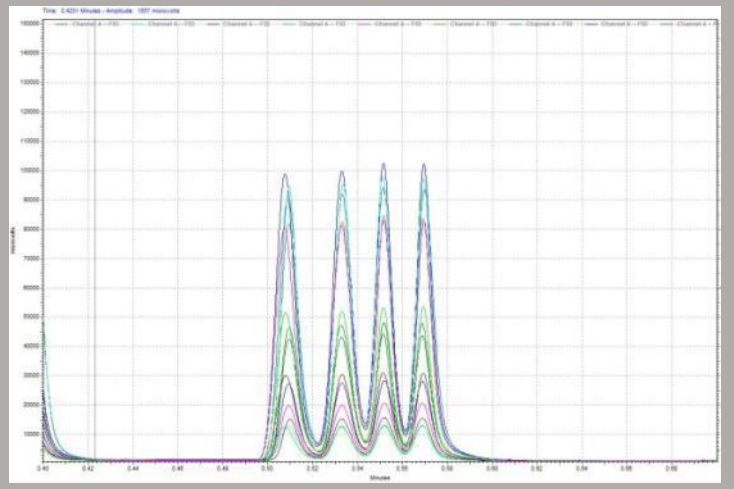
Infometrix

Gas Chromatography: 2 Instruments





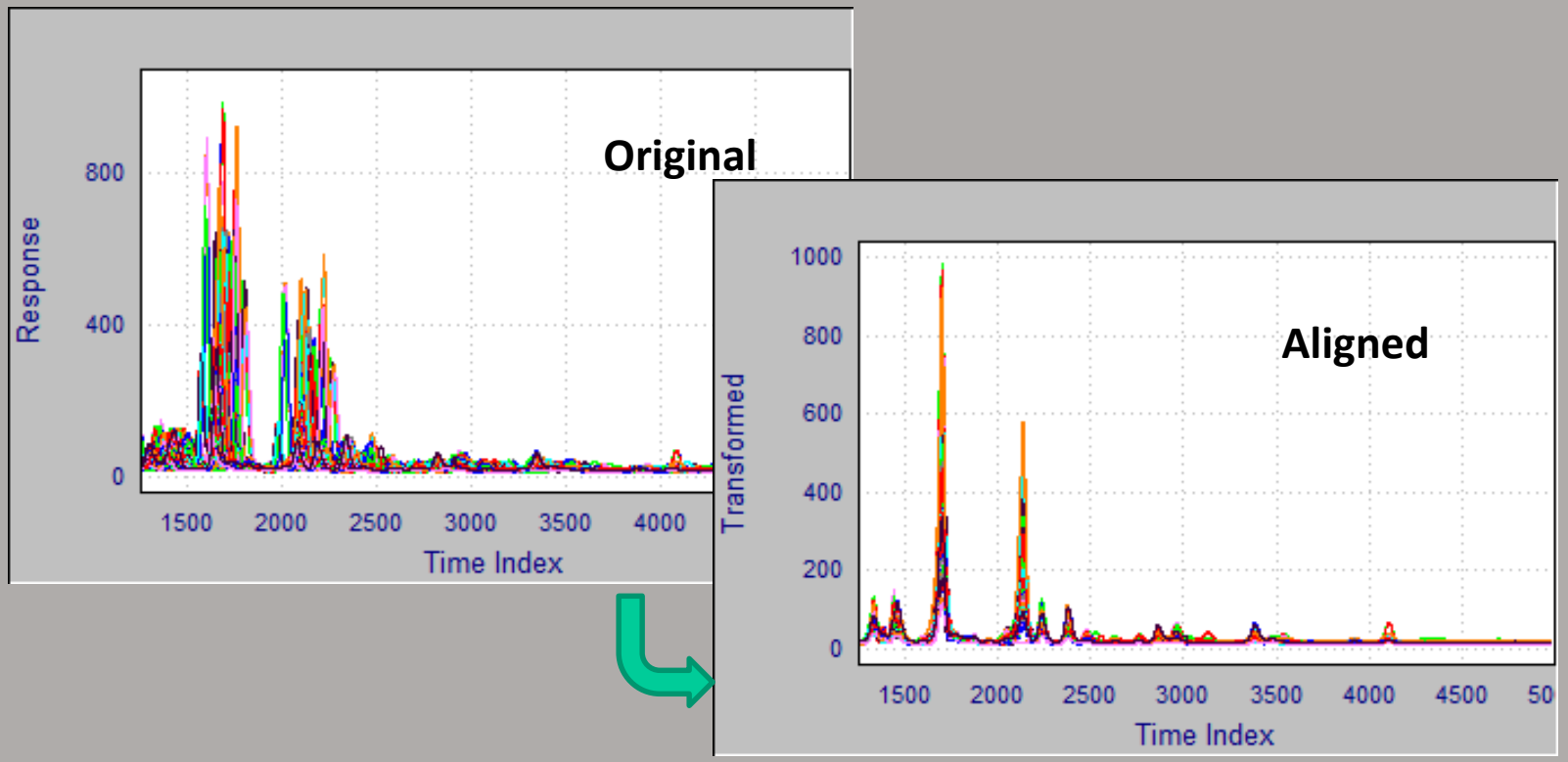




Infometrix

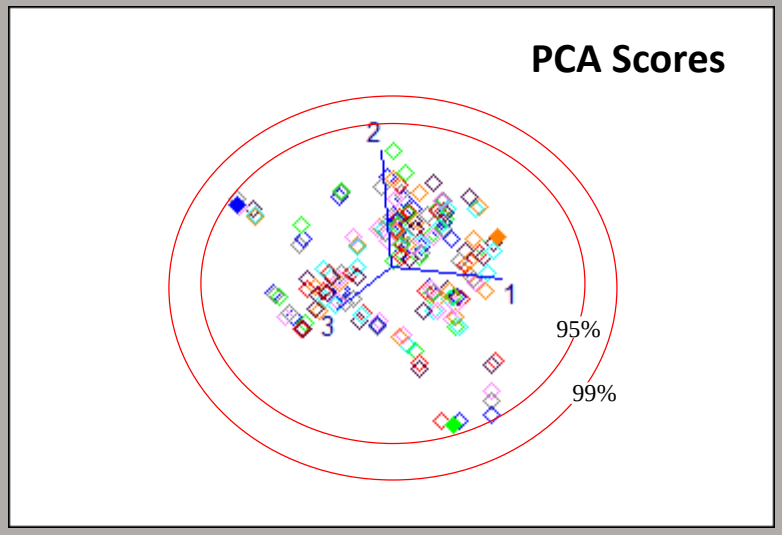
Alignment via Software

- Original chromatograms often show large variation in retention pattern; aligned chromatograms do not



PCA model from aligned profiles

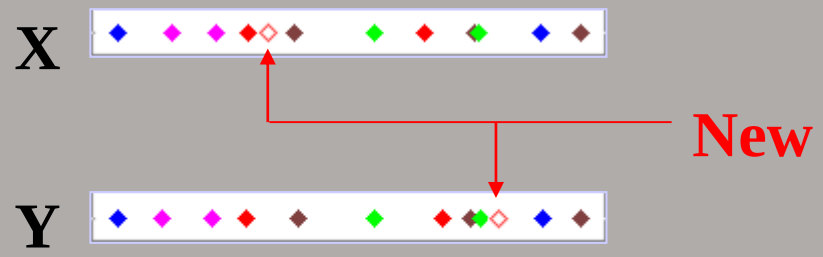
- Each evaluation will have a two qualifiers, one for in-model consistency and one to monitor out-of-model variation. These can be used to construct, for example, a warning limit (95% confidence interval) and an failure limit (99% confidence interval)

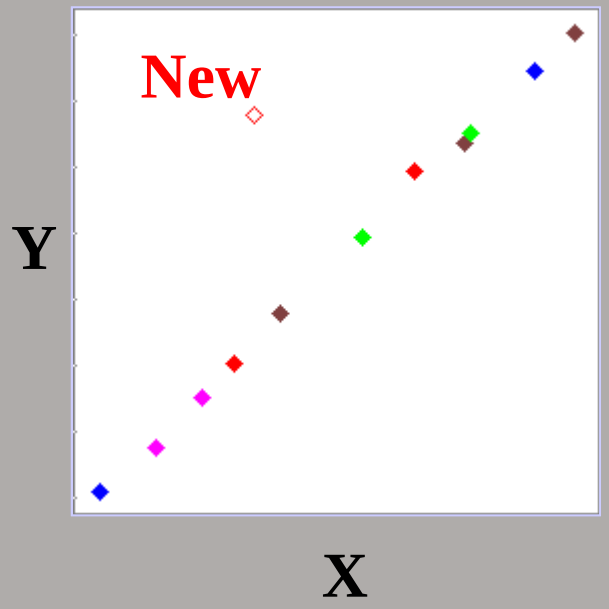


Infometrix

Why Chemometrics?

	X	Y
1	0.230	0.390
2	0.218	0.340
3	0.223	0.359
4	0.217	0.335
5	0.229	0.385
6	0.220	0.348
7	0.225	0.370
8	0.226	0.375
9	0.216	0.328
10	0.214	0.321
11	0.226	0.374
New	0.219	0.378





Is this sample good?

Delivering Information

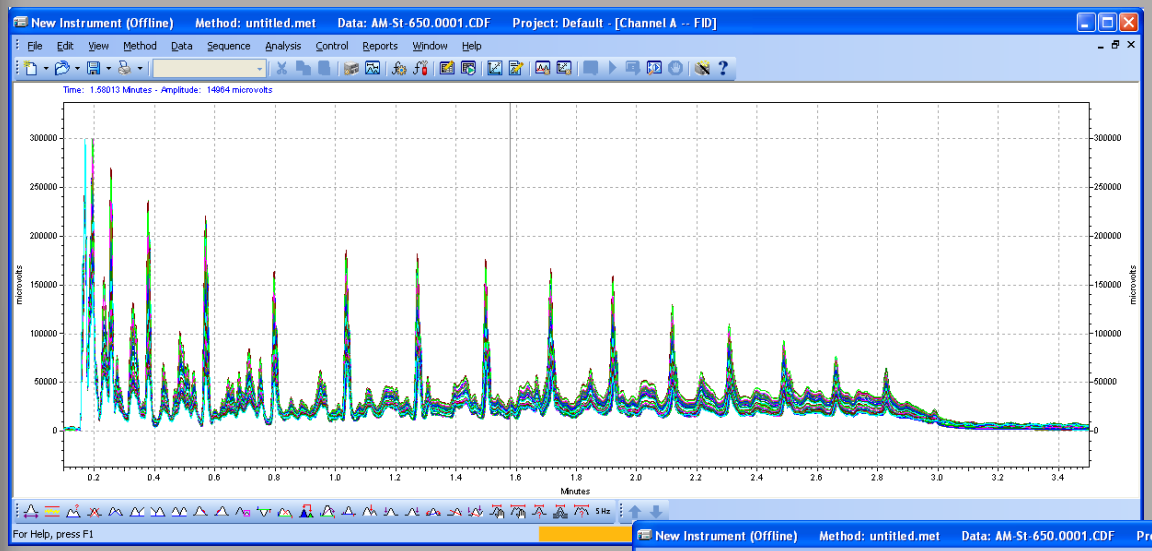
Just having the measurements does not translate into control

- There are not enough skilled technicians to handle even the current workload.
- Chemometrics solves the information processing problem with two technologies:
  - Alignment enables us to sell instruments that have vastly-lower calibration requirements.
  - Interpretation algorithms automates the generation and the qualification of the information derived from the raw data.

And if we can make all of our instruments look as much alike as possible.

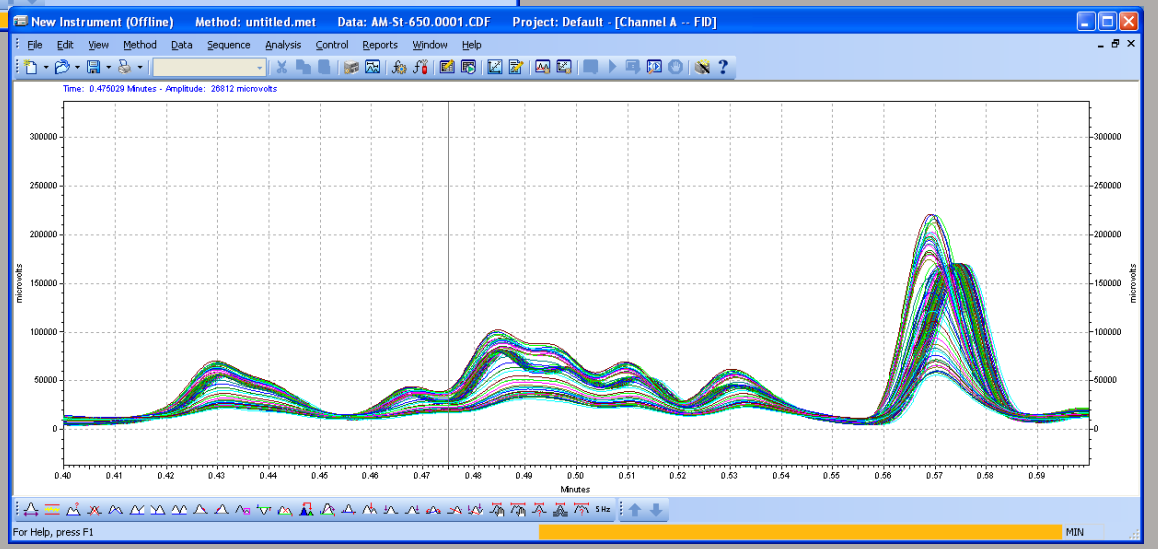
Interchangeability
Common interpretive base

Repeatability of a Process-Based Micro GC



Start to 650ºF Sample of a Conventional Crude Oil

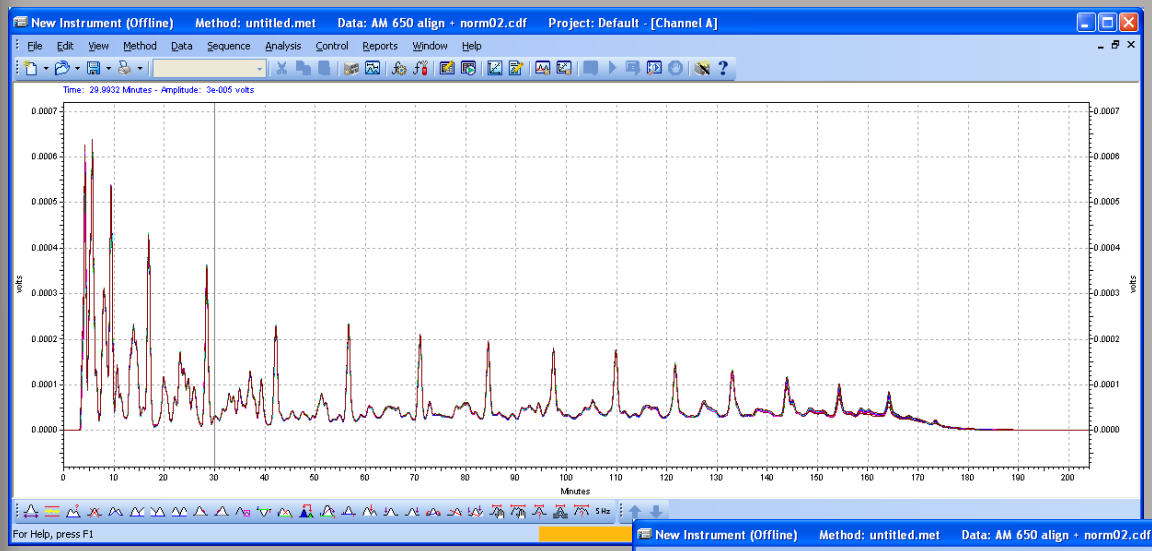
Above: 3.5 min run



Right: expansion from 0.4 to 0.6 min.

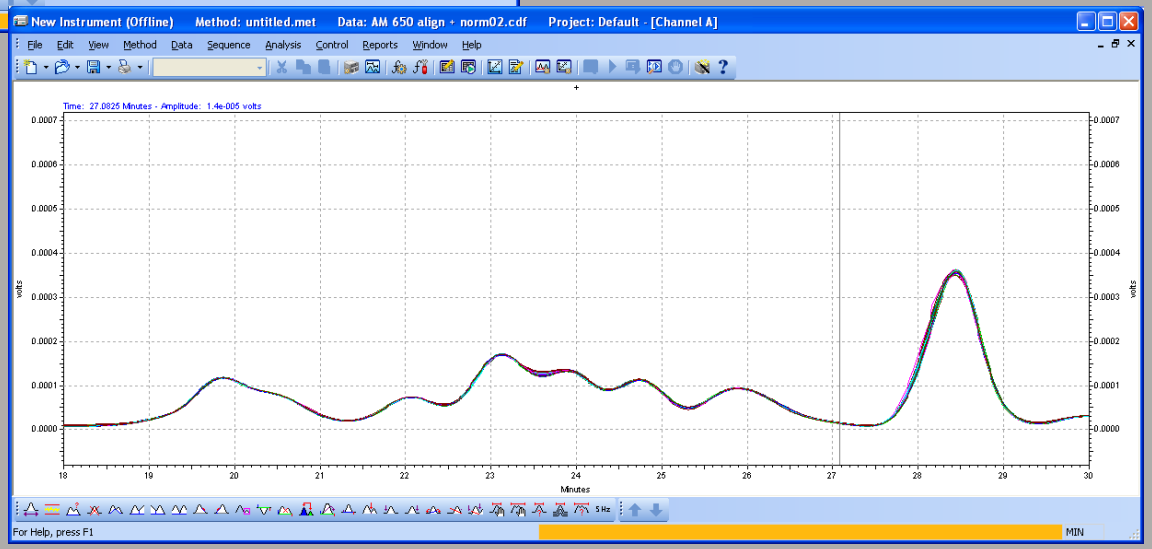
Infometrix

Repeatability of a Process-Based Micro GC



Substantial Reduction in Sample to Sample Variability

Above: 3.5 min run



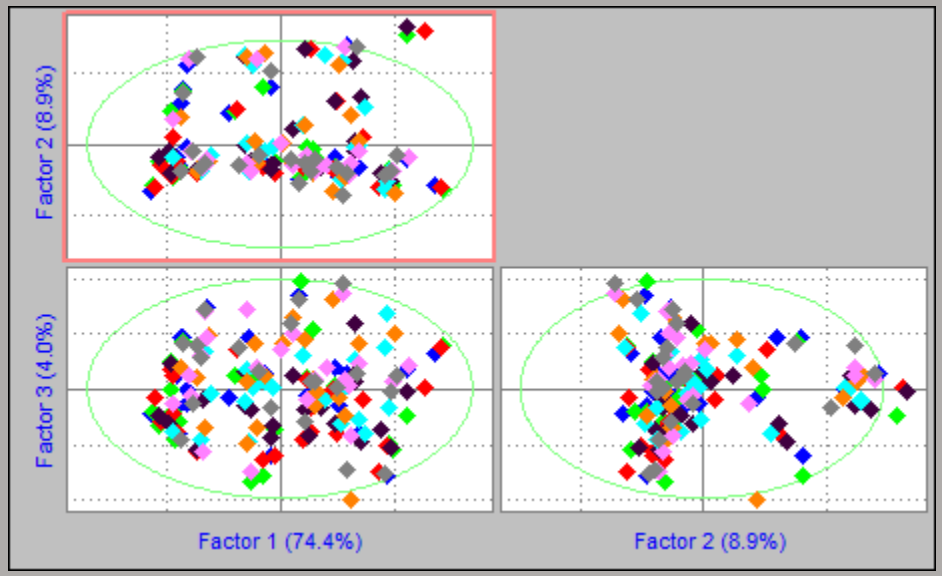
Right: expansion from 0.4 to 0.6 min.

Infometrix

Building a PCA Model

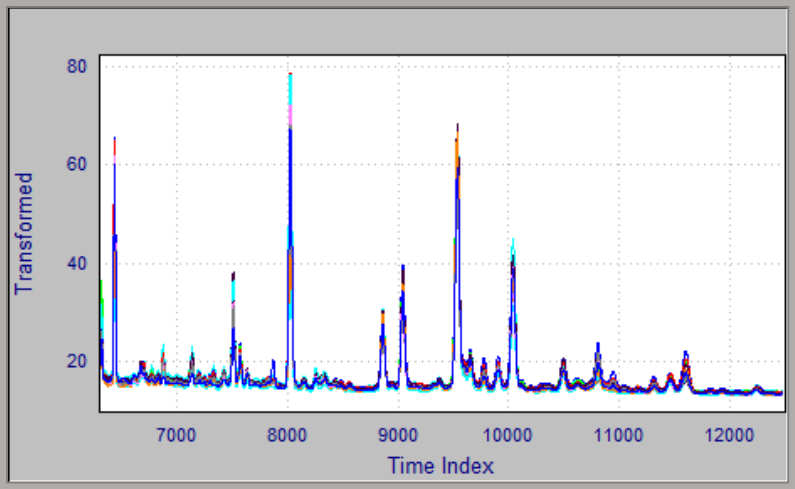
- For each fraction, a subset of training data was created using Kennard & Stone algorithm, then PCA was run
- PCA Scores from aligned profiles indicate a reasonably homogeneous data set

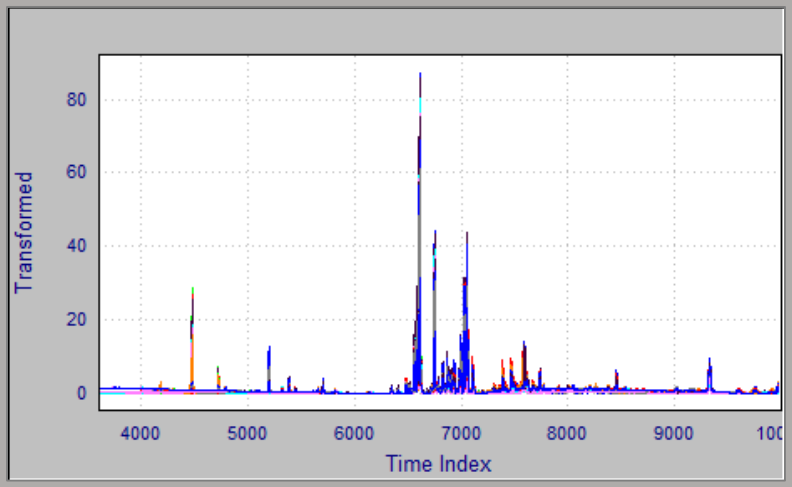
	Variance	Percent	Cumulative
Factor1	0.202	74.389	74.389
Factor2	0.024	8.864	83.253
Factor3	0.011	4.014	87.267
Factor4	0.008	2.950	90.217
Factor5	0.006	2.180	92.397
Factor6	0.004	1.435	93.832
Factor7	0.003	1.191	95.023
Factor8	0.003	0.971	95.994
Factor9	0.002	0.722	96.716
Factor10	0.001	0.529	97.246

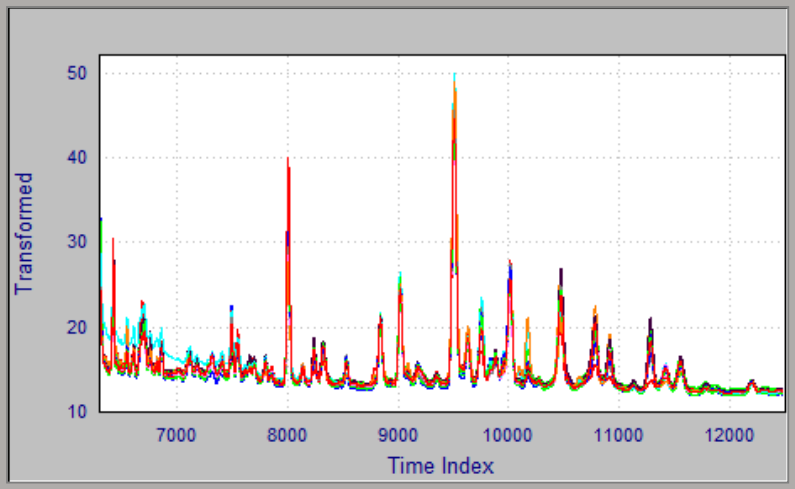


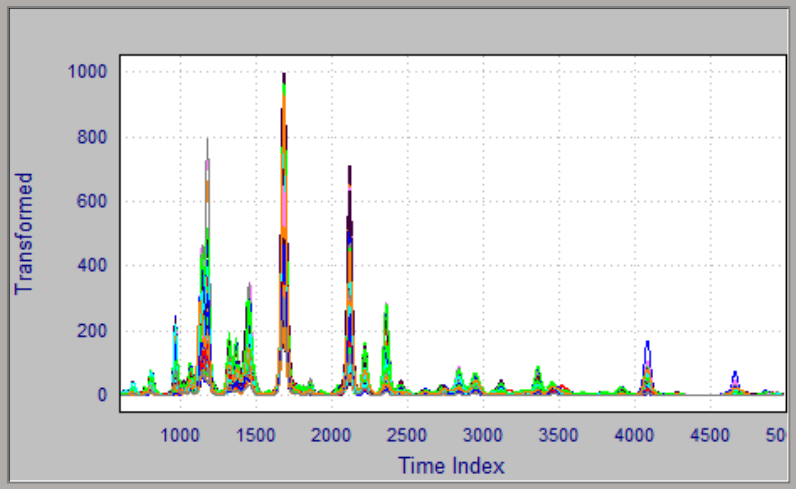
Infometrix

Product Fingerprints





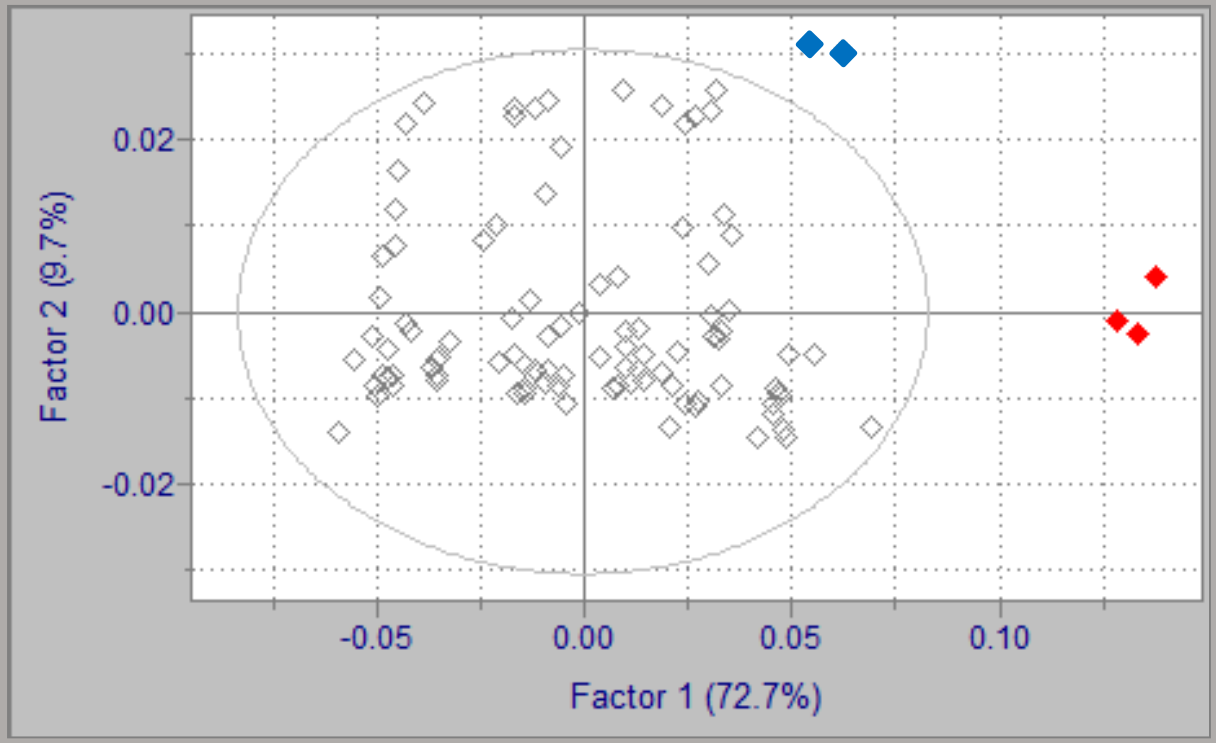




Infometrix

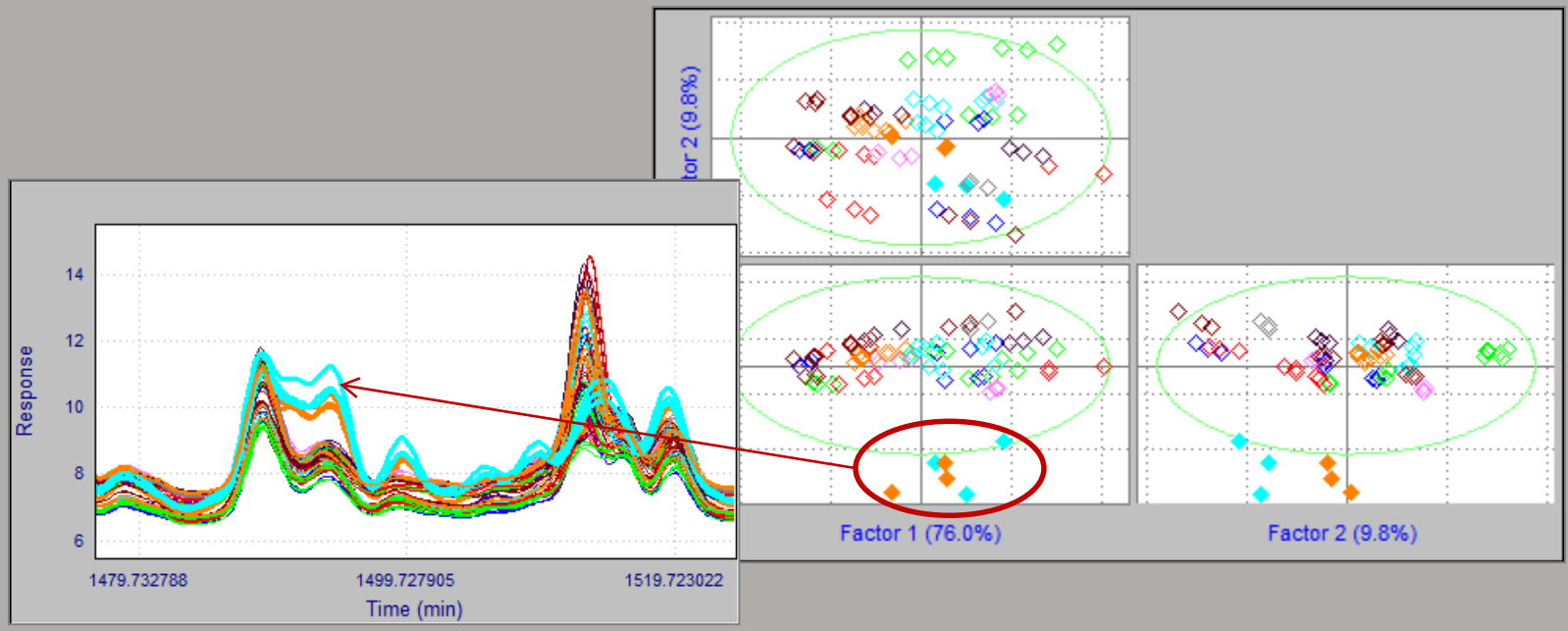
Control Plot Showing Outliers

- If there are deviations from the expected range, we instantly identify the variables at fault
- A plot of training set with aberrant samples overlaid can show multiple causes simultaneously

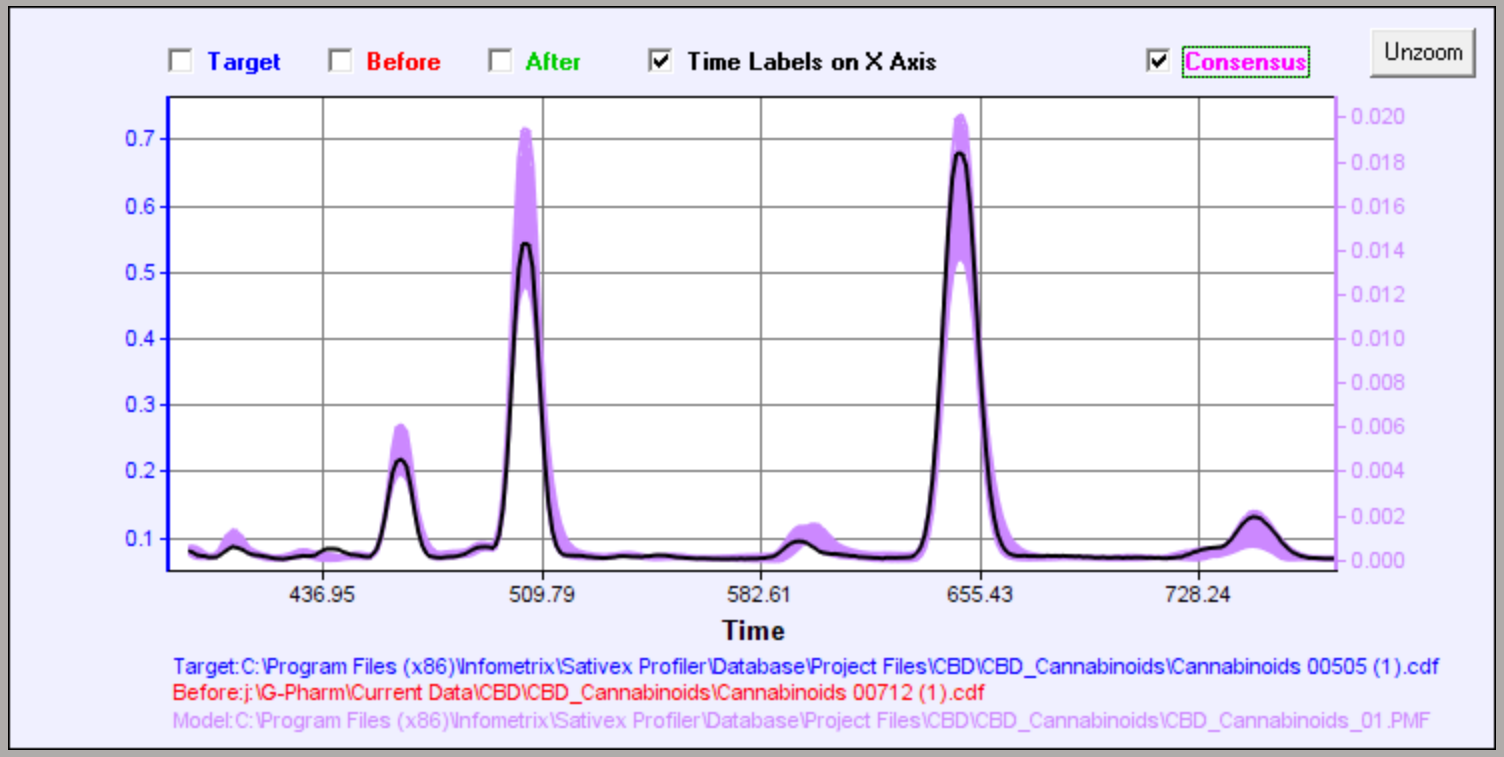


Validating the PCA Model

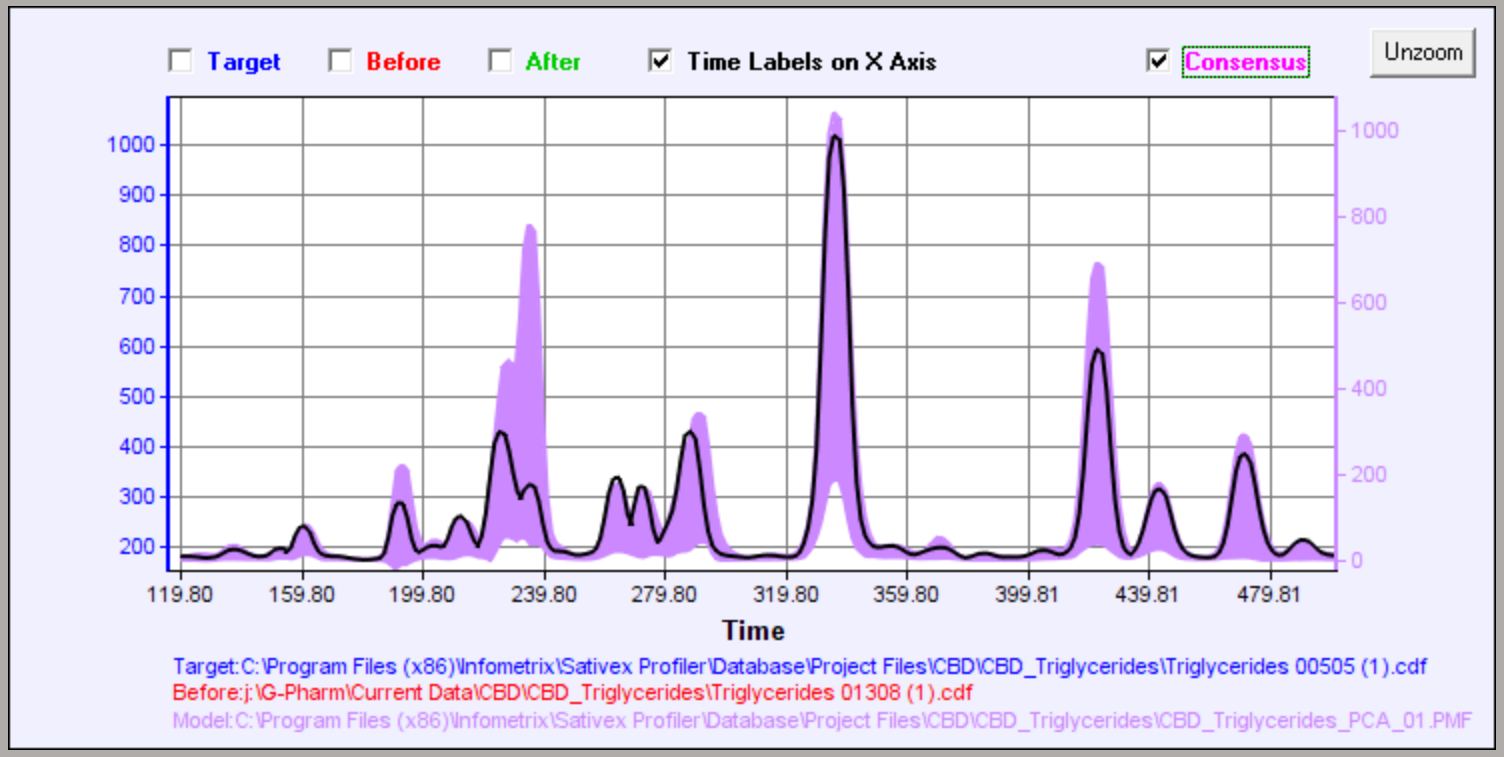
- Samples of bonafide and adulterated samples were projected into PCA models for each fraction
- Peaks responsible for aberrant profiles were discovered with contribution plots



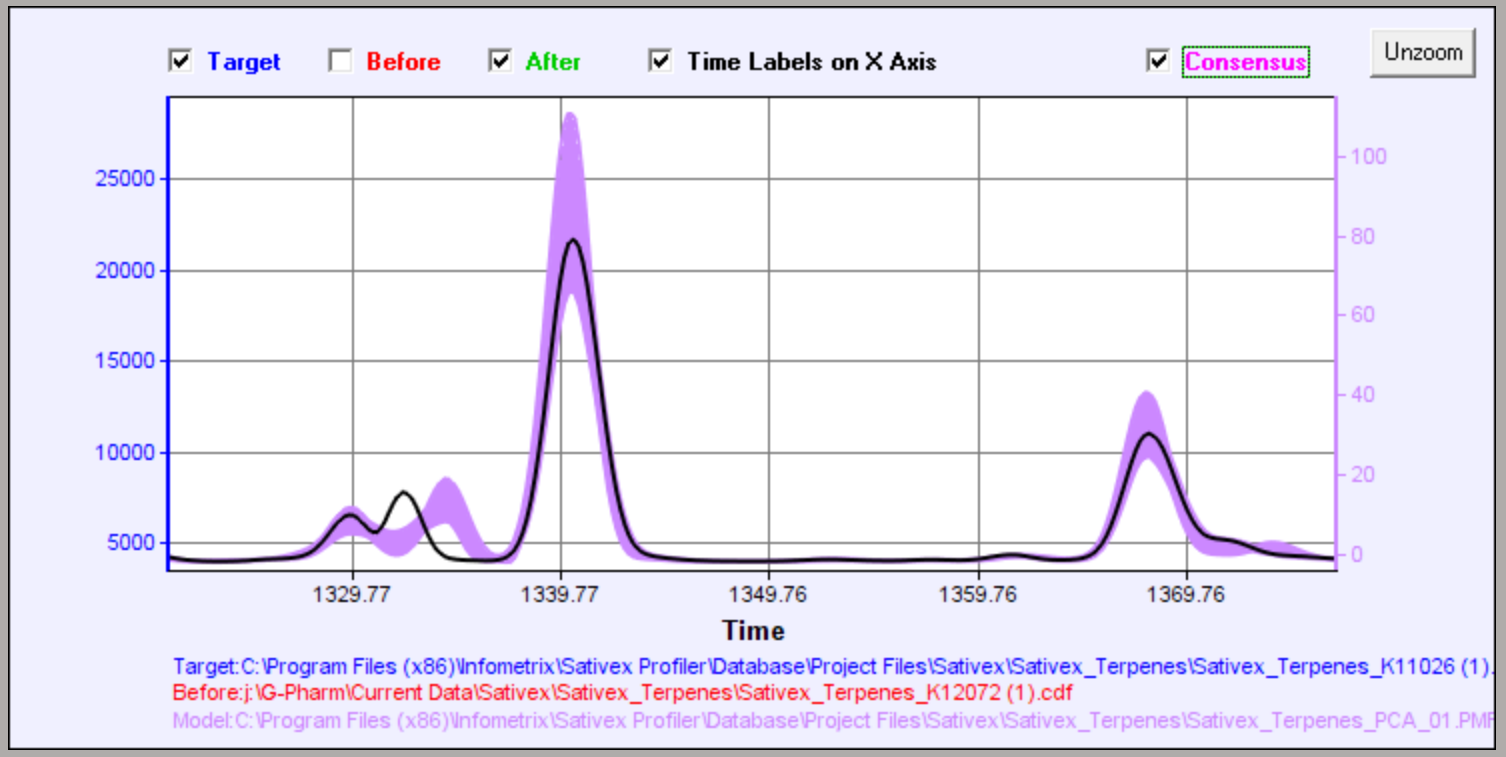
Good Sample


Target
Before
After
Time Labels on X Axis
Consensus
Unzoom
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.020
0.018
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
436.95
509.79
582.61
655.43
728.24
Time
Target:C:\Program Files (x86)\Infometrix\Sativex Profiler\Database\Project Files\CBD\CBD_Cannabinoids\Cannabinoids 00505 (1).cdf
Before:j:\G-Pharm\Current Data\CBD\CBD_Cannabinoids\Cannabinoids 00712 (1).cdf
Model:C:\Program Files (x86)\Infometrix\Sativex Profiler\Database\Project Files\CBD\CBD_Cannabinoids\CBD_Cannabinoids_01.PMF

In-model Outlier



Out-of-model Outlier



Infometrix

Example: HRVOC Fence Line Analysis

Sample	Quality	TNMHC	Ethane	Ethylene	Propane	Propylene	n-Butane	1-Butene	1,3-Butadiene	Pentane	1-Pentene	Isoprene	Hexane	1-Hexene	Heptane	Benzene
10-27-2005 12-26-53 pm_microfast 01_032.dat	Event	1198	0	3	0	0	5	0	0	8	0	1	1	0	0	1218
10-27-2005 2-14-07 pm_microfast 01_004.dat	Event	473	0	0	18	13	30	12	10	27	13	18	50	16	22	32
10-27-2005 2-54-00 pm_microfast 01_008.dat	Calibration	12744	491	487	492	488	977	489	488	987	493	499	1293	533	544	519
10-27-2005 2-34-04 pm_microfast 01_006.dat	Event	1547	58	67	53	52	105	52	51	111	56	56	174	78	78	74
10-27-2005 2-24-05 pm_microfast 01_005.dat	Event	477	21	22	19	13	29	13	14	30	14	12	52	4	22	28
10-27-2005 2-44-02 pm_microfast 01_007.dat	Event	4046	141	160	148	141	295	148	142	307	146	152	447	183	195	170
10-27-2005 3-13-55 pm_microfast 01_010.dat	N/A	207	0	0	0	3	8	4	5	15	1	2	21	7	12	18
10-27-2005 1-31-07 pm_microfast 01_003.dat	Flame out	47	4	0	1	1	0	0	0	1	0	1	0	0	0	0
10-27-2005 1-21-06 pm_microfast 01_002.dat	Flame out	49	0	0	0	0	0	0	0	0	2	1	1	0	1	0
10-27-2005 3-03-57 pm_microfast 01_009.dat	Calibration	25447	1005	1003	1004	1007	2012	1005	1007	2005	1003	0	2460	974	964	984
10-27-2005 7-26-55 am_microfast 01_003.dat	Background	110	36	0	12	0	7	0	0	5	1	1	0	0	3	8
10-27-2005 7-16-55 am_microfast 01_002.dat	Background	84	50	0	0	0	7	3	1	0	0	2	0	0	0	10
10-27-2005 7-06-54 am_microfast 01_001.dat	Background	89	31	3	13	0	7	0	2	4	1	10	0	0	0	9
10-27-2005 6-56-54 am_microfast 01_071.dat	Background	115	22	7	0	0	7	1	0	3	2	2	1	0	1	10
10-27-2005 6-46-55 am_microfast 01_070.dat	Background	72	23	0	9	0	6	0	1	0	0	3	5	0	2	12

Continuous data interpretation *PLUS* validation of a multivariate instrument

We can correct retention times to match an application-specific relevant sample

You can use this to make all instruments performing a similar task to look identical (Plug and Play)

Common regression and classification algorithms can be applied automatically to infer physical properties or characteristics

This allows us to bring more complex analyses into on-line use and creates the ability to automate an application-specific, objective evaluation system