Chemometrics

Application of mathematical, statistical, graphical or symbolic methods to maximize chemical information.

-However, this definition can be expanded to include:

biology (biometrics),

environmental science (environmetrics),

economics (econometrics)

-Two lines of development:

- experimental design: planning and performing experiments in a way that the resulting data contains the maximum information about stated questions.
- multivariate data analysis: utilizing all available data in the best possible way.

Chemometrics

1. Basic Statistics
2. Pattern Recognition and Classification
3. Optimization and Experimental Design
4. Multivariate Calibration Techniques
5. Quality Assurance and Good Laboratory Practice

Descriptive Statistics, Errors and Calibration

1. Basic Statistics → Descriptive

Spectrophotometric measurement (Abs) of a sample solution from 15 replicate measurements.

Measurement	Value	Measurement	Value
1	0.3410	9	0.3430
2	0.3350	10	0.3420
3	0.3470	11	0.3560
4	0.3590	12	0.3500
5	0.3530	13	0.3630
6	0.3460	14	0.3530
7	0.3470	15	0.3480
8	0.3460		

Descriptive Statistics, Errors and Calibration

Descriptive statistics for the spectrophotometric measurements.

Parameter	**Value**
Sample #, n	15
Mean	0.3486
Median	0.347
Std Dev	0.00731
RSD %	2.096
Std error	0.00189
Max value	0.363
Min value	0.335

Statistical Tests – Student's t-test, F-test, tests for outliers

Descriptive Statistics, Errors and Calibration

-Types of Errors

-Significant Figures

-Mean, Standard Deviation, RSD, etc…

-Gaussian Curve

-Confidence Intervals

-Comparison of Standard Deviations with the F test

-Calibration Issues

Comparison of Means (Student's t-test)

Rejection of Data (Q-test)

Descriptive Statistics, Errors and Calibration

Determining Concentration from Calibration Curve

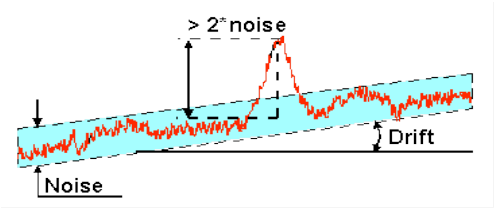
Basic steps:
(1) Make a series of dilutions of known concentration for the analyte.
(2) Analyze the known samples and record the results.
(3) Determine if the data is linear.
(4) Draw a line through the data and determine the line's slope and intercept.
(5) Test the unknown sample in duplicate or triplicate. Use the line equation to determine the concentration of the analyte: y = mx + b

$$Conc_{analyte} = \frac{reading - intercept}{slope}$$

Calibration Cont…

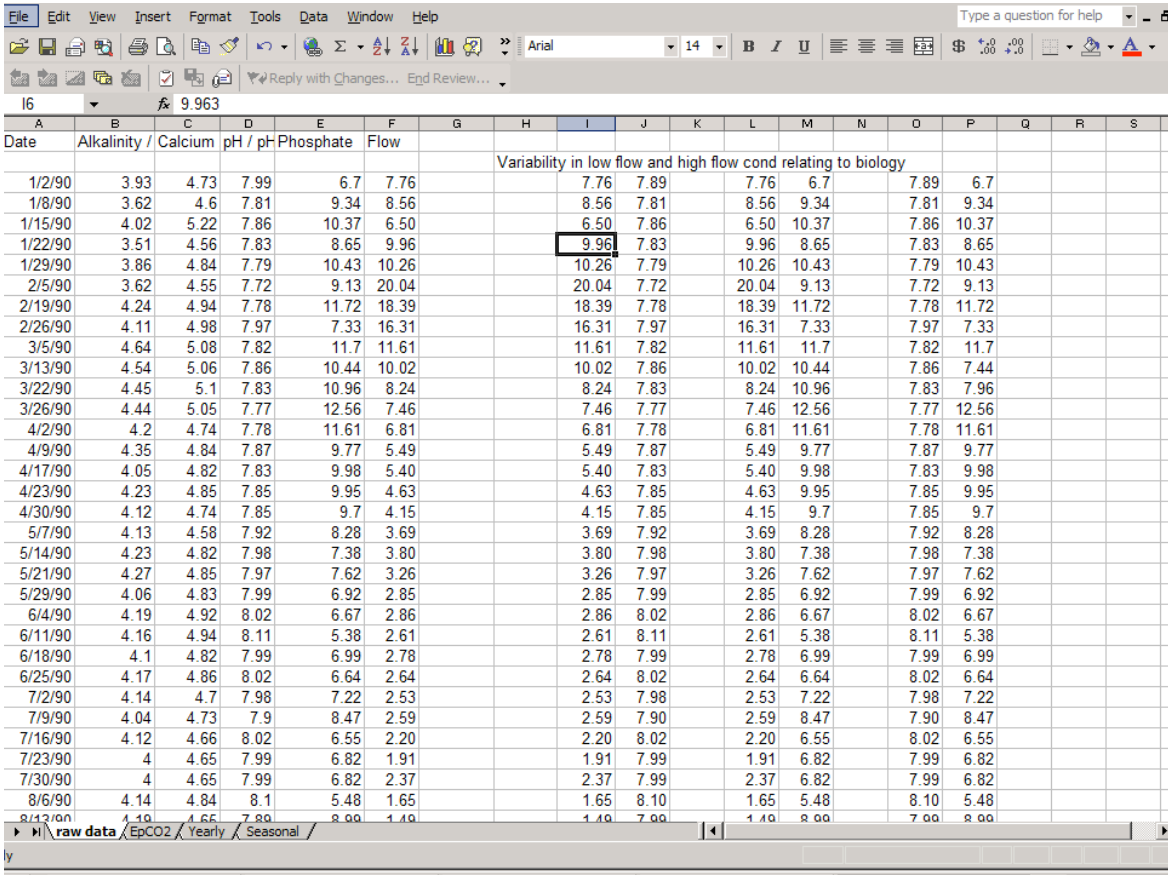
Limit of Detection (LOD) - lowest amount of analyte in a sample which can be detected but not necessarily quantitated as an exact value.

- mean of the blank sample plus 2 or 3 times the SD obtained on the blank sample (i.e., LOD = $\text{mean}_{blk} + Zs_{blk}$)



LOD calculation - alternative
Data required:
(1) calibration sensitivity = slope of line through the signals of the concentration standards including blank solution
(2) standard deviation for the analytical signal given by the blank solution

$$\text{LOD} = \frac{3x\,SD\,blank\,signals}{\text{slope of signals for stds}}$$



File Edit View Insert Format Tools Data Window Help — Type a question for help

Arial 14 — Reply with Changes… End Review…

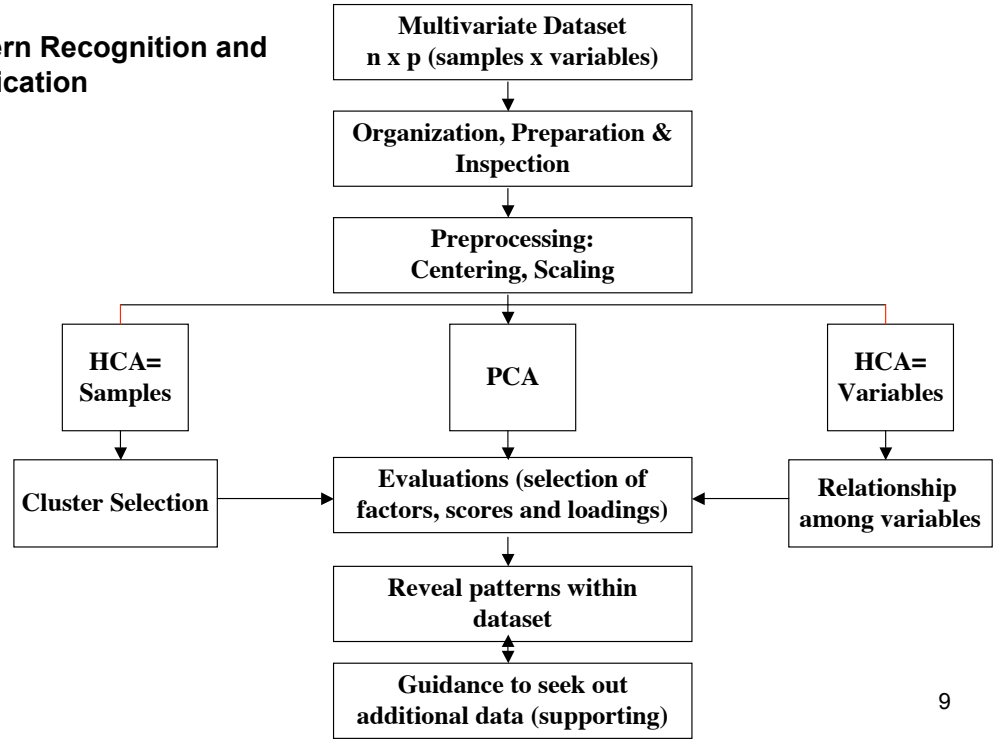
I6 9.963

Date	Alkalinity /	Calcium	pH / pH	Phosphate	Flow											
							Variability in low flow and high flow cond relating to biology									
1/2/90	3.93	4.73	7.99	6.7	7.76		7.76	7.89		7.76	6.7		7.89	6.7		
1/8/90	3.62	4.6	7.81	9.34	8.56		8.56	7.81		8.56	9.34		7.81	9.34		
1/15/90	4.02	5.22	7.86	10.37	6.50		6.50	7.86		6.50	10.37		7.86	10.37		
1/22/90	3.51	4.56	7.83	8.65	9.96		9.96	7.83		9.96	8.65		7.83	8.65		
1/29/90	3.86	4.84	7.79	10.43	10.26		10.26	7.79		10.26	10.43		7.79	10.43		
2/5/90	3.62	4.55	7.72	9.13	20.04		20.04	7.72		20.04	9.13		7.72	9.13		
2/19/90	4.24	4.94	7.78	11.72	18.39		18.39	7.78		18.39	11.72		7.78	11.72		
2/26/90	4.11	4.98	7.97	7.33	16.31		16.31	7.97		16.31	7.33		7.97	7.33		
3/5/90	4.64	5.08	7.82	11.7	11.61		11.61	7.82		11.61	11.7		7.82	11.7		
3/13/90	4.54	5.06	7.86	10.44	10.02		10.02	7.86		10.02	10.44		7.86	7.44		
3/22/90	4.45	5.1	7.83	10.96	8.24		8.24	7.83		8.24	10.96		7.83	7.96		
3/26/90	4.44	5.05	7.77	12.56	7.46		7.46	7.77		7.46	12.56		7.77	12.56		
4/2/90	4.2	4.74	7.78	11.61	6.81		6.81	7.78		6.81	11.61		7.78	11.61		
4/9/90	4.35	4.84	7.87	9.77	5.49		5.49	7.87		5.49	9.77		7.87	9.77		
4/17/90	4.05	4.82	7.83	9.98	5.40		5.40	7.83		5.40	9.98		7.83	9.98		
4/23/90	4.23	4.85	7.85	9.95	4.63		4.63	7.85		4.63	9.95		7.85	9.95		
4/30/90	4.12	4.74	7.85	9.7	4.15		4.15	7.85		4.15	9.7		7.85	9.7		
5/7/90	4.13	4.58	7.92	8.28	3.69		3.69	7.92		3.69	8.28		7.92	8.28		
5/14/90	4.23	4.82	7.98	7.38	3.80		3.80	7.98		3.80	7.38		7.98	7.38		
5/21/90	4.27	4.85	7.97	7.62	3.26		3.26	7.97		3.26	7.62		7.97	7.62		
5/29/90	4.06	4.83	7.99	6.92	2.85		2.85	7.99		2.85	6.92		7.99	6.92		
6/4/90	4.19	4.92	8.02	6.67	2.86		2.86	8.02		2.86	6.67		8.02	6.67		
6/11/90	4.16	4.94	8.11	5.38	2.61		2.61	8.11		2.61	5.38		8.11	5.38		
6/18/90	4.1	4.82	7.99	6.99	2.78		2.78	7.99		2.78	6.99		7.99	6.99		
6/25/90	4.17	4.86	8.02	6.64	2.64		2.64	8.02		2.64	6.64		8.02	6.64		
7/2/90	4.14	4.7	7.98	7.22	2.53		2.53	7.98		2.53	7.22		7.98	7.22		
7/9/90	4.04	4.73	7.9	8.47	2.59		2.59	7.90		2.59	8.47		7.90	8.47		
7/16/90	4.12	4.66	8.02	6.55	2.20		2.20	8.02		2.20	6.55		8.02	6.55		
7/23/90	4	4.65	7.99	6.82	1.91		1.91	7.99		1.91	6.82		7.99	6.82		
7/30/90	4	4.65	7.99	6.82	2.37		2.37	7.99		2.37	6.82		7.99	6.82		
8/6/90	4.14	4.84	8.1	5.48	1.65		1.65	8.10		1.65	5.48		8.10	5.48		
8/13/90	4.19	4.65	7.89	8.99	1.49		1.49	7.99		1.49	8.99		7.99	8.99		

raw data | EpCO2 | Yearly | Seasonal

Chemometrics

2. Pattern Recognition and Classification



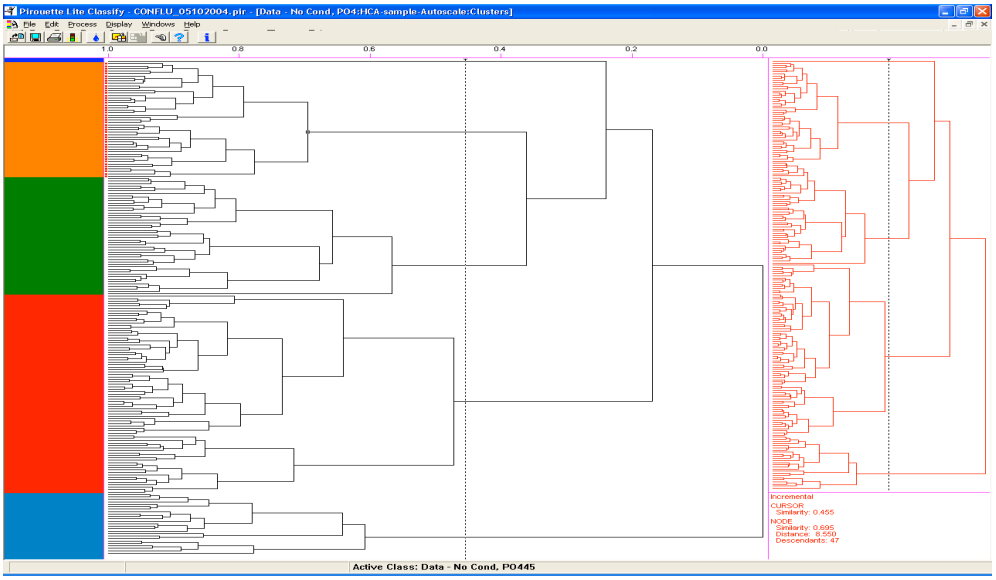
Chemometrics

<u>Hierarchical Cluster Analysis</u> (HCA) – interrelationships between samples presented in the form of a dendogram.

Each sample is initially considered as an individual <u>cluster</u>, and the clusters are progressively combined using a measure of <u>similarity</u>.

The use of HCA is to emphasize <u>natural groupings</u>. The length of the branches connecting two clusters is inversely related to their similarity: the longer the branch, the less the similarity.

Chemometrics

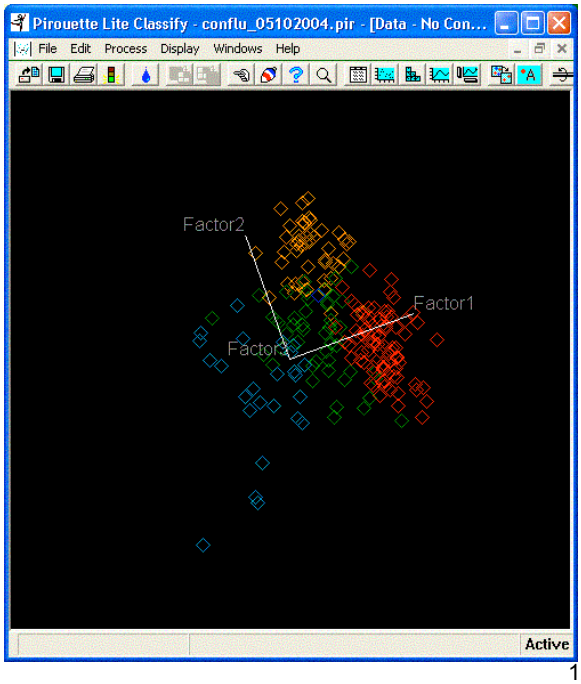


At a chosen similarity value (0.46 here), we can define and color 4 clusters – blue, red, green and orange - in the samples. These clusters are also clearly visible in the projections of samples generated by Principal Component Analysis.

Chemometrics

Principle Component Analysis (PCA) – a mathematical manipulation of a data matrix where the goal is to represent the variation present in many samples and/or variables using a small number of "factors."

The samples are plotted in a 3D space, with the first three components (or factors) defining three axes, and sample points are color-coded according to the grouping in a cluster analysis.



Outliers

Treatment of Outliers

1. Re-examine for Gross Errors
2. Estimate Precision to be Expected
3. Repeat Analysis if Time and Sufficient Sample is Available
4. If Analysis can not be Repeated, Perform a Q-Test
5. If Q-Test Indicates Retention of Value, Consider Reporting the Median

Chemometrics

3. Optimization and Experimental Design

Allow the experimenter to better understand and evaluate the factors that influence a particular system by means of statistical approaches.

The relationship between the various factors and response within a given system can be shown mathematically as follows:

$$y = f(x_1, x_2, x_3 \ldots, x_k)$$

where y is the response of interest in the system and x are the factors that affect the response when their values change.

Applications:
Screening- the factors that influence the experiment are identified.
Optimization- the optimal settings or conditions for an experiment are found.

Chemometrics

Experimental Designs

1. **Full Factorial Designs (two levels per factor)**
2. Fractional Factorial Design
3. Latin Squares
4. Greco-Latin Squares
5. Response Surface Designs (more than two levels for one or more factors)
6. Box-Behnken Designs
7. **Mixture Designs**

In general, the following types of factors can be distinguished: 1) continuous, e.g. temperature; and, 2) discrete, e.g. experimenter.

Factors are considered to be independent if there is no relationship between them and dependent if a relationship exists.

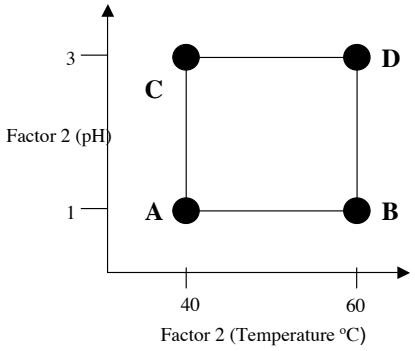
Chemometrics

Full Factorial Designs (two levels per factor)

Ex: The effects of reaction temperature and pH in determining the spectrophotometric response (absorbance) of a standard analyte solution.

Figure shows a graphical definition of the experimental domain.

The best experimental points in the domain are located in the corners A, B, C and D as follows:
A (40°C, pH 1); B (60 °C, pH 1); C (40°C, pH 3); D (60 °C, pH 3)



Chemometrics

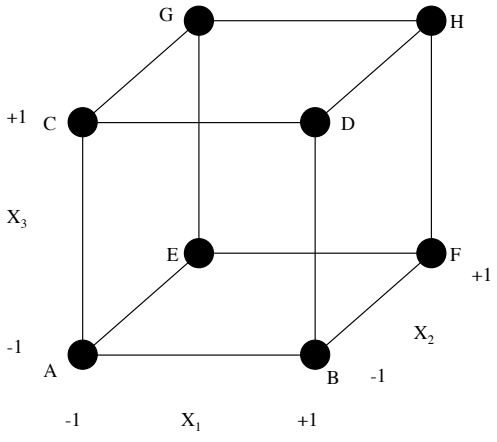
The four trials of experimental matrix are shown in the Table, with the results of each experiment indicated in the response column and the factor levels in the rows below the experimental matrix.

Expt. number	Temperature	pH	Response
A	-1	-1	y_1
B	+1	-1	y_2
C	-1	+1	y_3
D	+1	+1	y_4
Factor levels			
(-)	40°C	pH 1	
(+)	60°C	pH 3	

Note that -1 is used for the low level of each factor and +1 for the high level.

Chemometrics

If we introduce another variable (e.g. reagent concentration) in the experiment, it is then possible to represent the factors as faces on one or more cubes with the responses at the points.



Chemometrics

Mixture Designs

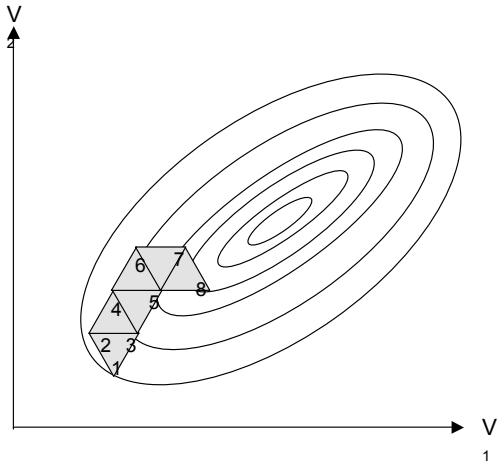
The independent factors are proportions of different components of a blend and often measured by their portions, which sum to 100% or normalized to 1, i.e.

$$\sum_{i=1}^{N} x_i = 1 \text{ for } x_i \geq 0$$

The design region for mixture proportions is termed a simplex. Simplex is a simple optimization algorithm seeking the vector of parameters corresponding to the global extreme (maximum or minimum) of any n-dimensional function $F(x_1, x_2,..,x_n)$.

Chemometrics

Ex: If the optimization of two factors occurs, the simplex will be a triangle. Points labeled 1, 2 and 3 define the first simplex with the worst response found at point 3.



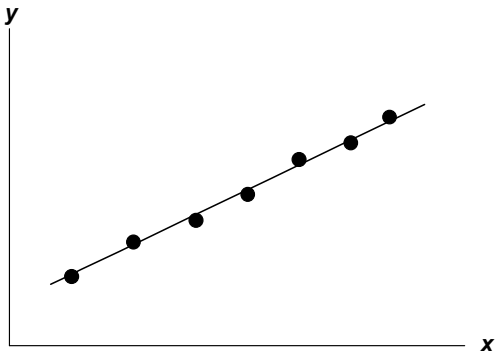
As the simplex continues along the path of the surface, points 1, 2 and 4 form a new simplex and the response is measured for the combination of factor levels given by 4.

Chemometrics

Multivariate Calibration Techniques

Traditional univariate calibration techniques involve the use of a single instrumental measurement to determine a single analyte.

In an ideal chemical measurement using high-precision instrumentation, an experimenter may obtain measurements linearly related to analyte concentration:



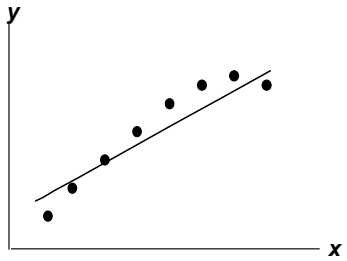
Chemometrics

Limitations of Univariate Techniques

-Very sensitive to the presence of outlier points.

-Selectivity and interferences (chemical and physical) - causing some degree of non-linearity.

-Univariate techniques are not well suited to the multitude of data collected from instrumentation currently being used in the analytical sciences.



Chemometrics

CLASSICAL LEAST SQUARES (CLS)

CLS approach is best applied to systems where the concentration of every analyte in the sample is known and obeys a linear relationship with measurement vectors.

For a single wavelength and a single analyte, this relationship (Beer's Law) can be explained mathematically by the following equation:

$$A_\lambda = \varepsilon_\lambda bc$$

where A_λ = the absorbance at wavelength λ; ε_λ = *molar absorption coefficient at wavelength* λ *in L* mol^{-1} cm^{-1}; b = cell pathlength (cm); and c = concentration of the analyte (mol L^{-1}).

Chemometrics

INVERSE LEAST SQUARES (ILS)

The dependent variable (concentration) is solved by calculating a solution from multiple independent variables (responses at the selected wavelengths).

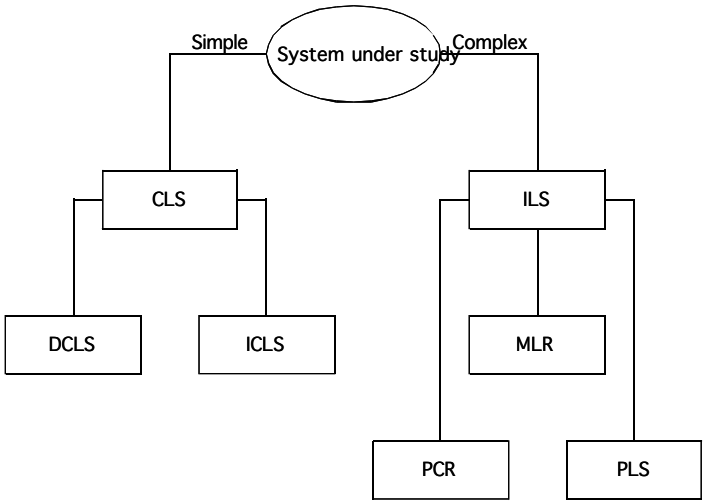
In ILS, we can combine the absorptivity coefficient (ε_λ) and cell pathlength (b) from Beer's Law to form a single constant relationship (matrix notation) with concentration:

$$c = P\, A_\lambda$$

where P = the matrix of coefficients.

ILS used in complex mixtures where it is not necessary to obtain the spectra of pure analytes present.

Chemometrics



Chemometrics

Approach	Advantages	Disadvantages
CLS	Used in estimating multivariate limits of detection, often based directly on Beer's Law	Not useful for mixtures with components that interact
	Used for Moderately complex mixtures	Requires knowledge of all components in calibration mixture
	Wavelength selection is not necessarily required for calibration	Susceptible to baseline effects
	Averaging effects make it less susceptible to noise	Interferences must be included in model
ILS	Used in estimating multivariate limits of detection, often based directly on Beer's Law	Wavelength selection can be difficult and time consuming
	Allows calibration of very complex mixtures	Accurate calibration often requires large numbers of samples
	Calculations are relatively fast	Number of wavelengths used in the model limited by the number of calibration samples

Laboratory Practice

Quality Assurance and Good Laboratory Practice

Validation – the assurance that an analytical procedure provides reproducible and secure results.

Validation Criteria:

1. Precision
2. Dynamic Range
3. Trueness
4. Selectivity
5. Limit of Detection
6. Limit of Determination
7. Robustness

Laboratory Practice

Internal Quality Assurance

→ Control Samples:

1. Standard Solutions
2. Blank Samples
3. Real Samples
4. Synthetic Samples
5. Certified Standard Reference Materials

Control samples should be analyzed at least once or twice in each series of analyses to monitor the accuracy of measurements.

Laboratory Practice

External Quality Assurance

→ Laboratory Intercomparison Studies:

1. Standardization of Analytical Procedures
2. Controlling the Analyses of a Laboratory
3. Preparation of Certified Reference Materials