Aalto University

Chemometric data analysis, fundamental methods (II)

Advanced crystallization and characterization techniques

June 1-5, 2020

Francesco Corona

Chemical and Metallurgical Engineering
School of Chemical Engineering

Classical calibration

Calibration is the problem of understanding the relation between **X** and **Y** blocks

- A model of how responses **y** vary as a function of explanatory variables **x**
- The practical goal is to use the model to predict **y**, given a value of **x**

Simple linear regression considers the case where both **x** and **y** are scalar variables

⇝ Estimate on concentration (or another property) from one spectral band

---

For vectors, firstly we consider two classical and fundamental calibration methods

- **Multiple linear regression**, **MLR**
- **Classical least-squares**, **CLS**

They are prototypical for all main calibration methods

- A good understanding is of essential importance
- They extend the simple linear regression

Multiple linear regression

Chemometric data analysis

Multiple linear regression, MLR

We considered the simple linear regression model, $y_n = (c + bx_n) + \varepsilon_n$ $(n = 1, 2, \ldots, N)$

- A single x-variable (input) is linked to a single y-variable (output)

We will develop further this method, by first letting the x-variable be a vector $\mathbf{x}$

- As we have multiple absorption bands, this is a more realistic scenario

The framework can then be further extended to account for vector $\mathbf{y}$-variables

- Again more realistic, as we measure multiple concentrations/properties

Multiple linear regression, MLR (cont.)

A **multiple linear regression model** is an extension of the linear regression model

$$y_n = c + b_1 x_{n,1} + b_2 x_{n,2} + \cdots + b_K x_{n,K} + \varepsilon_n, \quad (n = 1, 2, \cdots, N) \tag{1}$$

Sample size N, the number of data points, $(\mathbf{x}_n, y_n)$ pairs[1], $\mathbf{x}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,K})$

- $\mathbf{x}_n$, n-th value of the explanatory variables (K absorbances)
- y_n, n-th value of the response (concentration) variable
- ε_n, n-th error term (independent, zero mean, variance σ^2)

The model assumes that concentration is linearly related to absorbances, up to errors

- Each y_n is equal to a constant, plus a weighted sum of explanatory variables
- The weights can be collected in a weighting vector $\mathbf{b} = (b_1, b_2, \ldots, b_K)$

[1] The minimum number of observations is required to at least equal to $K + 1$.

Multiple linear regression, MLR (cont.)

$$y_n = \left(c + b_1 x_{n,1} + b_2 x_{n,2} + \cdots + b_K x_{n,K}\right) + \varepsilon_n, \quad (n = 1, 2, \cdots, N)$$

The model has a number of parameters that need to be calibrated/estimated from data

- $\rightsquigarrow$ c is the intercept of the regression model (an hyper-plane)
- $\rightsquigarrow$ b_k are the regression coefficients of the k-th x-variable

We implicitly assumed that ε_n the n-th noise term, the error, is somehow known

- It is assumed to be independent between the samples
- Assumed to have zero mean and common variance σ^2
- $\rightsquigarrow$ Thus, only estimation of σ^2 would be needed

---

Unknown parameters to be estimated from data

- $(c, b_1, b_2, \ldots, b_K)$, regression coefficients
- σ^2, residual variance

$$\boldsymbol{\theta} = (c, \underbrace{b_1, b_2, \ldots, b_K}_{\mathbf{b}}, \sigma^2)$$

Multiple linear regression, MLR (cont.)

Multiple linear regression in matrix-vector form

$$y_n = \left(c + \underbrace{b_1 x_{n,1} + b_2 x_{n,2} + \cdots + b_K x_{n,K}} \right) + \varepsilon_n, \quad (n = 1, 2, \cdots, N)$$

Consider the sum $b_1 x_{n,1} + b_2 x_{n,2} + \cdots + b_K x_{n,K}$, it is the inner product of two vectors

- $\rightsquigarrow$ One $(1 \times K)$ row-vector

$$\mathbf{x}_n = \begin{bmatrix} x_{n,1} & x_{n,2} & \cdots & x_{n,K} \end{bmatrix}$$

- $\rightsquigarrow$ One $(K \times 1)$ column-vector

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

$$\mathbf{x}_n \mathbf{b} = \begin{bmatrix} x_{n,1} & x_{n,2} & \cdots & x_{n,K} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} = b_1 x_{n,1} + b_2 x_{n,2} + \cdots + b_K x_{n,K}$$

We can rewrite the multiple regression model for a single observation in vector form

$$y_n = c + \mathbf{x}_n \mathbf{b} + \varepsilon_n, \quad (n = 1, 2, \cdots, N) \tag{3}$$

Multiple linear regression, MLR (cont.)

We can stack the N model equations

$$\begin{aligned} y_n &= c + \mathbf{x}_n\mathbf{b} + \varepsilon_n && \text{(4a)} \\ &= 1c + \mathbf{x}_n\mathbf{b} + \varepsilon_n && \text{(4b)} \end{aligned}$$

with $n = 1, 2, \ldots, N$

$$\rightsquigarrow \quad \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}}_{\mathbf{1}c} c + \underbrace{\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \\ \vdots \\ \mathbf{x}_{N-1} \\ \mathbf{x}_N \end{bmatrix}}_{\mathbf{X}\mathbf{b}} \mathbf{b} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \\ \vdots \\ \varepsilon_{N-1} \\ \varepsilon_N \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

We have defined the $(N \times K)$ matrix $\mathbf{X}$, the $\mathbf{X}$-block

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_{N-1} \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,K} \\ \vdots & \vdots & & \vdots \\ [x_{n,1} & x_{n,2} & \cdots & x_{n,K}] \\ \vdots & \vdots & & \vdots \\ x_{N-1,1} & x_{N-1,2} & \cdots & x_{N-1,K} \\ x_{N,1} & x_{N,2} & \cdots & x_{N,K} \end{bmatrix}$$

Multiple linear regression, MLR (cont.)

We can rewrite the multiple linear regression model in matrix form

$$\mathbf{y} = \mathbf{1}c + \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \tag{5}$$

Or, equivalently, in terms of the centred $\mathbf{X}$-block

$$\mathbf{y} = \mathbf{1}c + \underbrace{(\dot{\mathbf{X}} + \mathbf{1}\overline{\mathbf{x}})}_{\text{centring}}\mathbf{b} + \boldsymbol{\varepsilon} \tag{6a}$$

$$= \mathbf{1}b_0 + \dot{\mathbf{X}}\mathbf{b} + \boldsymbol{\varepsilon} \tag{6b}$$

with $b_0 = c + (\overline{\mathbf{x}}\mathbf{b})$ a constant term after centring $\mathbf{X}$

We also used $\overline{\mathbf{x}} = \begin{bmatrix} \overline{x_1} & \overline{x_2} & \cdots & \overline{x_k} & \cdots & \overline{x_K} \end{bmatrix}^T$

Multiple linear regression, MLR (cont.)

MLR, estimation

$$\mathbf{y} = \mathbf{1}c + \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$$

The least-squares estimators of the parameters of the multiple linear regression model

$$\widehat{b}_0 = \overline{y} \tag{7a}$$

$$\widehat{\mathbf{b}} = \underbrace{\left(\dot{\mathbf{X}}^T\dot{\mathbf{X}}\right)^{-1}\dot{\mathbf{X}}^T}_{\dot{\mathbf{X}}^\dagger}\dot{\mathbf{y}} = \mathbf{V}_X^{-1}\mathbf{V}_{Xy} \tag{7b}$$

The estimation of $\mathbf{b}$ depends upon the non-singularity of matrix $\dot{\mathbf{X}}^T\dot{\mathbf{X}}$ (invertible)

- $\mathbf{V}_X$, the variance-covariance matrix of (the variable in) $\mathbf{X}$
- $\mathbf{V}_{Xy}$, the covariance between (the variable in) $\mathbf{X}$ and $\mathbf{y}$

Alternatively, we can write the estimator of $\mathbf{b}$ using the left pseudo-inverse $\dot{\mathbf{X}}^\dagger$ of $\dot{\mathbf{X}}$

$$\rightsquigarrow \widehat{\mathbf{b}} = \dot{\mathbf{X}}^\dagger\dot{\mathbf{y}} \tag{8}$$

Multiple linear regression (cont.)

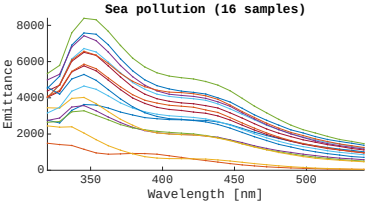
Example

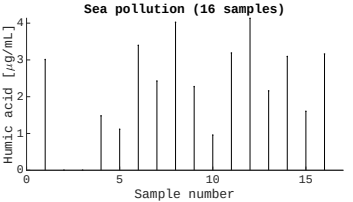
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Estimate the concentration of humic acid using absorbance at two given wavelengths

- We selected as band 345 [nm] and band 430 [nm] ($K = 2$)
- Multiple linear regression ($M = 1$)

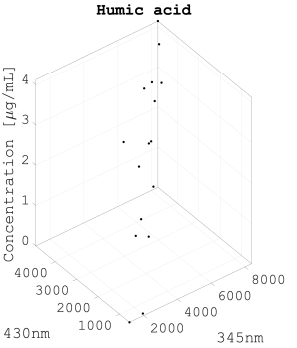
$$y_n = (c + b_1 x_{n,1} + b_2 x_{n,2}) + \varepsilon_n, \quad (n = 1, 2, \cdots, N) \tag{9}$$

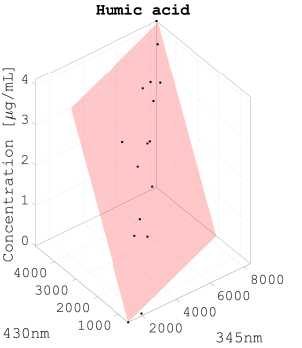




Multiple linear regression, MLR (cont.)

Case I: Original **x**-variable and original y_2-variable





The estimated regression parameters

- $b_1 = -0.4250$ and $b_2 = -0.0003$
- $c = 0.0015$

Multiple linear regression, MLR (cont.)

MLR, prediction

We obtain another observation $\mathbf{z}$ of the explanatory variables, but not the response

- The system that generates this observation is assumed to be the same

We want to predict the value of the response (as if we had measured it), given $\mathbf{z}$

- This is easily done by substituting $\mathbf{z}$ for $\mathbf{x}$ in the learned model
- (Equivalent to 'reading' it from the plot of the regression plane)

$$\rightsquigarrow \quad \widehat{y} = \underbrace{\overline{y}}_{b_0} + (\mathbf{z} - \overline{\mathbf{x}})\,\widehat{\mathbf{b}} \quad \text{(predicted composition)} \tag{10}$$

Prediction $\widehat{y}$ depends on the learning data

$\rightsquigarrow$ Via $\overline{\mathbf{x}}$, $\overline{y}$ and $\widehat{\mathbf{b}}$

Multiple linear regression, MLR (cont.)

Applicability of multiple linear regression

Condition for existence of the least-squares solution to the MLR calibration problem

- The $(K \times K)$ matrix $\dot{\mathbf{X}}^T\dot{\mathbf{X}}$ must be non-singular
- (Equivalently, $\mathbf{V}_X$ is non-singular)

This requires that a number of practical conditions on the calibration data are satisfied

- $N > K$, there must be more samples than explanatory variables
- The columns of matrix $\begin{bmatrix}\mathbf{1} & \dot{\mathbf{X}}\end{bmatrix}$ must be linearly independent

---

As a result, a number of limitations to the use of plain MLR must be first considered

- In NIR spectroscopy, often K is in the order of hundreds or thousands

⇝ Therefore, a compatibly large number of samples must be collect

- In NIR spectroscopy, the K absorbances are highly collinear (correlated)

⇝ This leads to ill-conditioning of matrix $\dot{\mathbf{X}}^T\dot{\mathbf{X}}$ (whatever the sample size)

Multivariate multiple linear regression, mMLR

Multivariate multiple linear regression is a straightforward extension of MLR

- Not only multiple explanatory variables and multiple response variables
- $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix}^T$, size $(N \times K)$
- $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_N \end{bmatrix}^T$, size $(N \times M)$

The model in centred form,

$$\dot{\mathbf{y}}_n = \dot{\mathbf{x}}_n \mathbf{B} + \boldsymbol{\varepsilon}_n, \quad (n = 1, 2, \cdots, N) \tag{11}$$

Sample size N, the number of available data points, the $(\mathbf{x}_n, \mathbf{y}_n)$ pairs[2]

- $\mathbf{x}_n$, n-th value of the explanatory variables (K absorbances)

$$\mathbf{x}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,K})$$

- $\mathbf{y}_n$, n-th value of the response variables (M concentrations)

$$\mathbf{y}_n = (y_{n,1}, y_{n,2}, \ldots, y_{n,M})$$

- $\boldsymbol{\varepsilon}_n$, n-th error term (independent, zero mean, covariance $\boldsymbol{\Sigma}$)

[2] The minimum number of observations is required to at least equal to $K + 1$.

Multivariate multiple linear regression, mMLR (cont.)

$$\dot{\mathbf{y}}_n = \dot{\mathbf{x}}_n \underbrace{\begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_M\end{bmatrix}}_{\mathbf{B}} + \boldsymbol{\varepsilon}_n, \quad (n = 1, 2, \cdots, N)$$

The model assumes that concentrations are linearly related to absorbances, up to errors

- Each $y_{m,n}$ is a constant, plus the m-th weighted sum of explanatory variables
- The weighting vector is $\mathbf{b}_k = (b_{m,1}, b_{m,2}, \ldots, b_{m,K})$

Multiple linear regression in matrix-matrix form

We can again stack the N model equations $\dot{\mathbf{y}}_n = \dot{\mathbf{x}}_n\mathbf{B} + \boldsymbol{\varepsilon}_n$, with $n = 1, 2, \ldots, N$

$$\rightsquigarrow \underbrace{\begin{bmatrix} \dot{y}_{1,1} & \dot{y}_{1,2} & \cdots & \dot{y}_{1,M} \\ \dot{y}_{2,1} & \dot{y}_{2,2} & \cdots & \dot{y}_{2,M} \\ \vdots & \vdots & & \vdots \\ \dot{y}_{n,1} & \dot{y}_{n,2} & \cdots & \dot{y}_{n,M} \\ \vdots & \vdots & & \vdots \\ \dot{y}_{N-1,1} & \dot{y}_{N-1,2} & \cdots & \dot{y}_{N-1,M} \\ \dot{y}_{N,1} & \dot{y}_{N,2} & \cdots & \dot{y}_{N,M} \end{bmatrix}}_{\dot{\mathbf{Y}}} = \underbrace{\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \\ \vdots \\ \dot{\mathbf{x}}_n \\ \vdots \\ \dot{\mathbf{x}}_{N-1} \\ \dot{\mathbf{x}}_N \end{bmatrix} \begin{bmatrix}\mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_M\end{bmatrix}}_{\dot{\mathbf{X}}\mathbf{B}} + \underbrace{\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \\ \vdots \\ \boldsymbol{\varepsilon}_{N-1} \\ \boldsymbol{\varepsilon}_N \end{bmatrix}}_{\mathbf{E}}$$

Multivariate multiple linear regression, mMLR (cont.)

mMLR, estimation

$$\mathbf{y}_n = \mathbf{1c} + \mathbf{x}_n \underbrace{\left[\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_M\right]}_{\mathbf{B}} + \boldsymbol{\varepsilon}_n$$

The least-squares estimators of the parameters of the multiple linear regression model

$$\widehat{\mathbf{b}}_0 = \overline{\mathbf{y}} \tag{12a}$$

$$\widehat{\mathbf{B}} = \underbrace{\left(\dot{\mathbf{X}}^T\dot{\mathbf{X}}\right)^{-1}\dot{\mathbf{X}}^T}_{\dot{\mathbf{X}}^\dagger}\dot{\mathbf{Y}} = \mathbf{V}_X^{-1}\mathbf{V}_{XY} \tag{12b}$$

The estimation of $\mathbf{B}$ depends upon the non-singularity of matrix $\dot{\mathbf{X}}^T\dot{\mathbf{X}}$ (invertible)

- $\mathbf{V}_Y$, the variance-covariance matrix of (the variables in) $\mathbf{Y}$
- $\mathbf{V}_{XY}$, the covariance between (the variables in) $\mathbf{X}$ and $\mathbf{Y}$

Alternatively, we can write the estimator of $\mathbf{B}$ using the left pseudo-inverse $\dot{\mathbf{X}}^\dagger$ of $\dot{\mathbf{X}}$

$$\rightsquigarrow \quad \widehat{\mathbf{B}} = \dot{\mathbf{X}}^\dagger\dot{\mathbf{Y}} \tag{13}$$

Multivariate multiple linear regression, mMLR (cont.)

The calibration of a multivariate multiple linear regression mMLR model is equivalent to performing multiple linear regression MLR on each of the columns of $\mathbf{Y}$ separately

$$\widehat{\mathbf{B}} = \underbrace{\left(\dot{\mathbf{X}}^T\dot{\mathbf{X}}\right)^{-1}\dot{\mathbf{X}}^T}_{\dot{\mathbf{X}}^\dagger}\dot{\mathbf{Y}} = \mathbf{V}_X^{-1}\mathbf{V}_{XY}$$

---

mMLR, prediction

We obtain another observation $\mathbf{z}$ of the explanatory variables, but not the response

- The system that generates this observation is assumed to be the same

We want to predict the value of the responses (as if we had measured them), given $\mathbf{z}$

- This is easily done by substituting $\mathbf{z}$ for $\mathbf{x}$ in the learned model
- (Equivalent to 'reading' it from the plot of the regression plane)

$$\widehat{\mathbf{y}} = \underbrace{\overline{\mathbf{y}}}_{\mathbf{b}_0} + (\mathbf{z} - \overline{\mathbf{x}})\,\widehat{\mathbf{B}} \tag{14}$$

Prediction $\widehat{\mathbf{y}}$ depends on data via $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ and $\widehat{\mathbf{B}}$

Multiple linear regression (cont.)

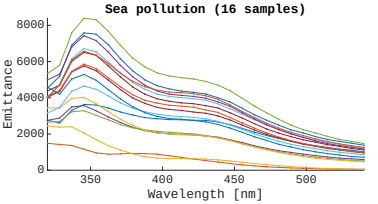
Example

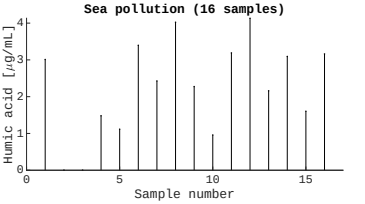
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

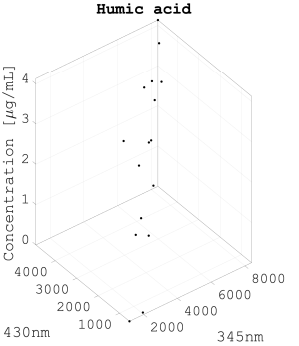
Estimate the concentration of all components using absorbance at two wavelengths

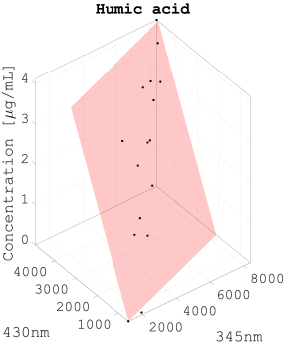
- We selected as band 345 [nm] and band 430 [nm] ($K = 2$)
- Multivariate multiple linear regression ($M = 3$)

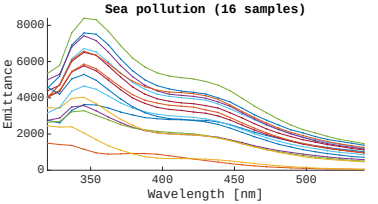
Case I: Original **x**-variable and original **y**-variable

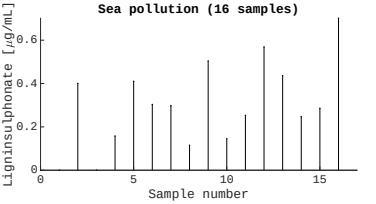

Sea pollution (16 samples)
Emittance
Wavelength [nm]
8000
6000
4000
2000
0
350
400
450
500

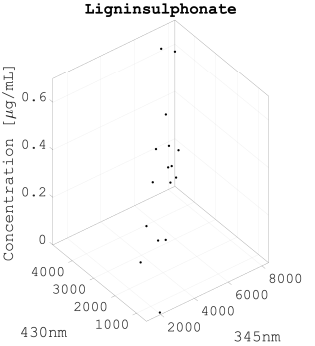

Sea pollution (16 samples)
Humic acid [μg/mL]
Sample number
4
3
2
1
0
5
10
15

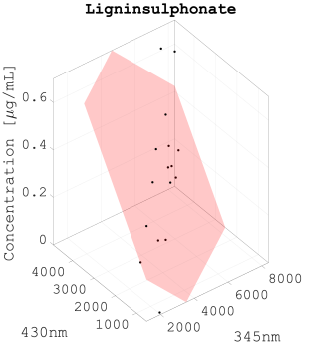

Humic acid
Concentration [μg/mL]
4
3
2
1
0
4000
3000
2000
1000
430nm
8000
6000
4000
2000
345nm

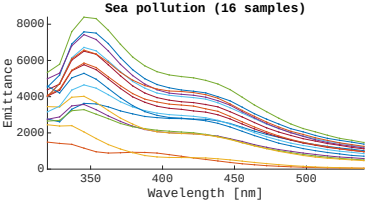

Humic acid
Concentration [μg/mL]
4
3
2
1
0
4000
3000
2000
1000
430nm
8000
6000
4000
2000
345nm

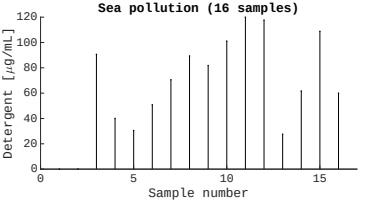

Sea pollution (16 samples)
Emittance
Wavelength [nm]
8000
6000
4000
2000
0
350
400
450
500

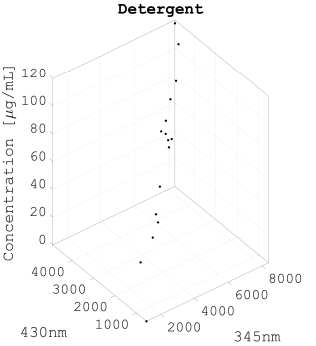

Sea pollution (16 samples)
Ligninsulphonate [µg/mL]
Sample number
0.6
0.4
0.2
0
0
5
10
15

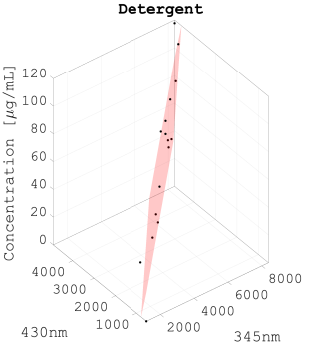

Ligninsulphonate
Concentration [µg/mL]
0.6
0.4
0.2
0
4000
3000
2000
1000
430nm
2000
4000
6000
8000
345nm


Ligninsulphonate
Concentration [µg/mL]
0.6
0.4
0.2
0
4000
3000
2000
1000
430nm
2000
4000
6000
8000
345nm


Sea pollution (16 samples)
Emittance
Wavelength [nm]


Sea pollution (16 samples)
Detergent [μg/mL]
Sample number


Detergent
Concentration [μg/mL]
430nm
345nm


Detergent
Concentration [μg/mL]
430nm
345nm

Multiple linear regression (cont.)

Example

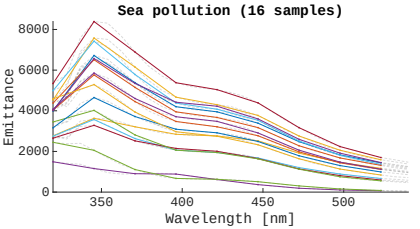
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

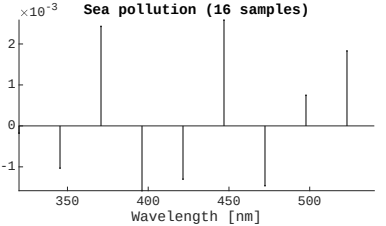
Estimate the concentration of all components using absorbance at two wavelengths

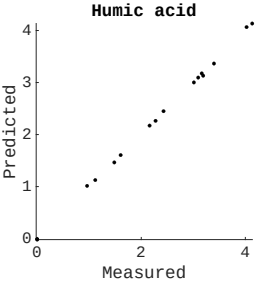
- We selected every third band in the given range ($K = 9$)
- Multivariate multiple linear regression ($M = 3$)

Case I: Original **x**-variable and original **y**-variable

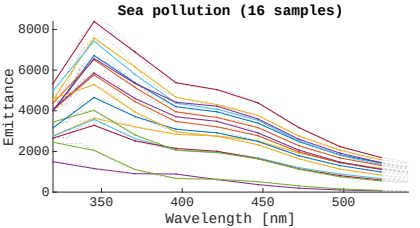
Multiple linear regression, MLR (cont.)

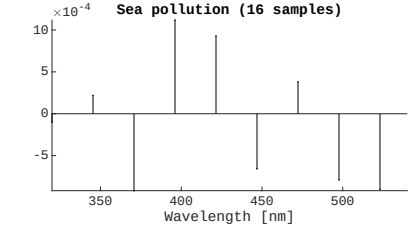


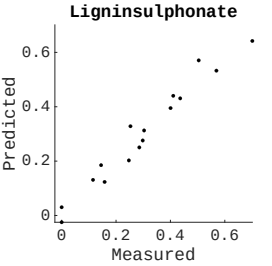




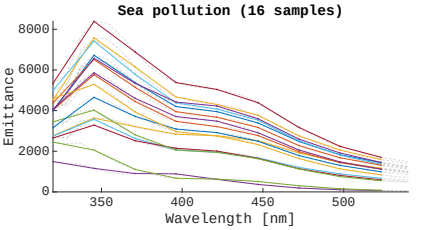
Multiple linear regression, MLR (cont.)

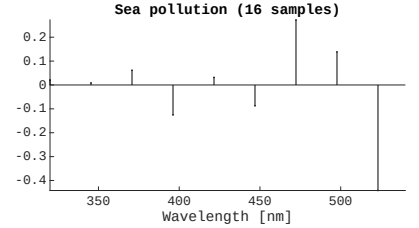


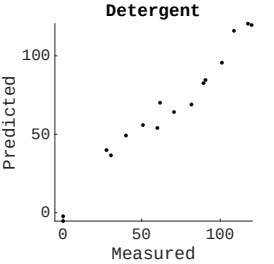




Multiple linear regression, MLR (cont.)







Classical least squares

Chemometric data analysis

Classical least-squares, CLS

MLR and mMLR models assume that the y-variables are the dependent variables

- The models assume that concentrations/properties depend on absorbances

This is the inverse relationship of what the Beer-Lambert law seems to suggest

- The absorbance $x_{n,k}$ of sample n depends (linearly) on concentration

$$\begin{aligned} x_{n,k} &= \left(y_{n,1}\, a_{1,k} + y_{n,2}\, a_{2,k} + \cdots + y_{n,M}\, a_{M,k}\right) + \varepsilon_{n,k} \\ &= \left(\sum_{m=1}^{M} y_{n,m}\, a_{m,k}\right) + \varepsilon_{n,k} \end{aligned}$$

Classical least-squares, CLS (cont.)

It is crucial to study what happens when physical principles are reinstated in the model

$\rightsquigarrow$ **Classical least-squares** aims at generalising the Beer-Lambert law

$$x_{n,k} = \alpha_k + \underbrace{\left(\sum_{m=1}^{M} y_{n,m} a_{m,k} \right)}_{\text{Beer-Lambert}} + \varepsilon_{n,k}, \quad n = 1, \ldots, N \text{ and } k = 1, \ldots K \tag{16}$$

- $y_{n,m}$, m-th concentration of the n-th sample, the (n, m)-th entry in the **Y**-block
- $x_{n,k}$, k-th absorbance of the n-th sample, the (n, k)-th entry in the **X**-block
- $a_{m,k}$, k-th absorbance of the pure m-th component (absorptivities)
- α_k, k-th offset term (what is added to the Beer-Lambert law)

Moreover, $e_{n,k}$ is the (n, k)-th error term, assumed independent across the samples

- It is assumed to have zero mean and common variance

Classical least-squares, CLS (cont.)

Classical least squares in matrix-matrix form

$$x_{n,k} = \left(\alpha_k + \underbrace{\sum_{m=1}^{M} y_{n,m}\, a_{m,k}}_{\mathbf{y}_n \mathbf{a}} \right) + \varepsilon_{n,k}, \quad (n = 1, \ldots, N \text{ and } k = 1, \ldots K)$$

Consider the sum $y_{n1}\, a_{1,k} + y_{n,2}\, a_{2,k} + \cdots + y_{n,M}\, a_{M,k}$, the inner product of two vectors

- $\rightsquigarrow$ One $(1 \times M)$ row-vector

$$\mathbf{y}_n = \begin{bmatrix} y_{n,1} & y_{n,2} & \cdots & y_{n,M} \end{bmatrix}$$

- $\rightsquigarrow$ One $(M \times 1)$ column-vector

$$\mathbf{a}_k = \begin{bmatrix} a_{1,k} \\ a_{2,k} \\ \vdots \\ a_{M,k} \end{bmatrix}$$

$$\mathbf{y}_n \mathbf{a}_k = \begin{bmatrix} y_{n,1} & y_{n,2} & \cdots & y_{n,M} \end{bmatrix} \begin{bmatrix} a_{1,k} \\ a_{2,k} \\ \vdots \\ a_{M,k} \end{bmatrix} = y_{n,1}\, a_{1,k} + y_{n,2}\, a_{2,k} + \cdots + y_{n,M}\, a_{M,k}$$

Rewrite the expression for a single sample n and a single wavelength k in vector form

$$x_{n,k} = \alpha_k + \mathbf{y}_n \mathbf{a}_k + \varepsilon_{n,k}, \quad (n = 1, 2, \cdots, N \text{ and } k = 1, \ldots, K) \tag{18}$$

Classical least-squares, CLS (cont.)

We can also rewrite the classical least-squares calibration model in matrix form to get

- Now, we must stack the $N \times K$ expressions $x_{n,k} = \alpha_k + y_n \mathbf{a}_k + \varepsilon_{n,k}$

$$\underbrace{\begin{bmatrix} x_{1,1} & \cdots & x_{1,K} \\ \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,K} \\ \vdots & & \vdots \\ x_{N,1} & \cdots & x_{N,K} \end{bmatrix}}_{\mathbf{X}} = \underbrace{\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_n \\ \vdots \\ \boldsymbol{\alpha}_N \end{bmatrix}}_{\mathbf{1}\boldsymbol{\alpha}} + \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \\ y_N \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_k & \cdots & \mathbf{a}_K \end{bmatrix}}_{\mathbf{Y}\mathbf{A}} + \underbrace{\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_n \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{bmatrix}}_{\mathbf{E}} \tag{19a}$$

$$\mathbf{X} = \mathbf{1}\boldsymbol{\alpha} + \mathbf{Y}\mathbf{A} + \mathbf{E} \tag{19b}$$

- $\mathbf{A}$ is the $(M \times K)$ matrix of absorbances for pure components
- $\mathbf{X}$ is the $(N \times K)$ $\mathbf{X}$-block and $\mathbf{Y}$ is the $(N \times M)$ $\mathbf{Y}$-block
- $\mathbf{E}$ is the $(N \times K)$ matrix of random noise terms
- $\boldsymbol{\alpha}$ is the $(1 \times K)$ row vector of offsets
- $\mathbf{1}$ denotes a $(N \times 1)$ vector of ones

$\boldsymbol{\alpha}$ and $\mathbf{A}$ (and the residual variance) are the unknown parameters to be estimated

Classical least-squares, CLS (cont.)

$$\mathbf{X} = \mathbf{1}\boldsymbol{\alpha} + \mathbf{Y}\mathbf{A} + \mathbf{E}$$

We firstly introduce the centred version of the $\mathbf{Y}$-block, $\dot{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\overline{\mathbf{y}}$, then we get

$$\rightsquigarrow \quad \mathbf{Y} = \dot{\mathbf{Y}} + \mathbf{1}\overline{\mathbf{y}}$$

We used $\overline{\mathbf{y}} = \begin{bmatrix} \overline{y}_1 & \cdots & \overline{y}_m & \cdots & \overline{y}_M \end{bmatrix}$ denoting a $(1 \times M)$ vector of means of $\mathbf{Y}$

After substituting $\mathbf{Y}$ in the model, we can write

$$\mathbf{X} = \mathbf{1}\boldsymbol{\alpha} + \left(\dot{\mathbf{Y}} + \mathbf{1}\overline{\mathbf{y}}\right)\mathbf{A} + \mathbf{E} \tag{20a}$$

$$= \mathbf{1}\boldsymbol{\alpha}_0 + \dot{\mathbf{Y}}\mathbf{A} + \mathbf{E} \tag{20b}$$

We used $\boldsymbol{\alpha}_0 = \mathbf{1}\boldsymbol{\alpha} + \overline{\mathbf{y}}\mathbf{A}$

Classical least-squares, CLS (cont.)

CLS, estimation

The least squares estimators of the unknown parameters $\boldsymbol{\alpha}_0$ and $\mathbf{A}$

$$\widehat{\boldsymbol{\alpha}}_0 = \overline{\mathbf{x}} \tag{21a}$$

$$\hat{\mathbf{A}} = \underbrace{\left(\dot{\mathbf{Y}}^T\dot{\mathbf{Y}}\right)^{-1}\dot{\mathbf{Y}}^T}_{\dot{\mathbf{Y}}^\dagger}\dot{\mathbf{X}} = \mathbf{V}_Y^{-1}\mathbf{V}_{YX} \tag{21b}$$

The estimation of $\mathbf{A}$ depends upon the non-singularity of matrix $\dot{\mathbf{Y}}^T\dot{\mathbf{Y}}$ (invertible)

- $\mathbf{V}_X$, the variance-covariance matrix of $\mathbf{Y}$
- $\mathbf{V}_{YX}$, the covariance between $\mathbf{Y}$ and $\mathbf{X}$

Alternatively, we can rewrite the estimator of $\mathbf{A}$ using the left pseudo-inverse of $\dot{\mathbf{Y}}$

$$\rightsquigarrow \quad \widehat{\mathbf{A}} = \dot{\mathbf{Y}}^\dagger\dot{\mathbf{X}} \tag{22}$$

Classical least-squares, CLS (cont.)

Prediction is classical least-squares

For prediction, we start by considering the centred classical least-squares model

$$\dot{\mathbf{X}} = \dot{\mathbf{Y}}\mathbf{A} \quad \text{(error-free case)}$$

We solve this equation with respect to $\dot{\mathbf{Y}}$, by right-multiplying both sides by $\mathbf{A}^\dagger$

$$\mathbf{A}^\dagger = \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T\right)^{-1}$$

Assuming that matrix $\mathbf{A}\mathbf{A}^T$ of dimension $(M \times M)$ in non-singular, we write

$$\dot{\mathbf{X}}\mathbf{A}^\dagger = \dot{\mathbf{Y}}\mathbf{A}\mathbf{A}^\dagger = \dot{\mathbf{Y}} \tag{23a}$$

$$\rightsquigarrow \quad \dot{\mathbf{Y}} = \dot{\mathbf{X}}\mathbf{A}^\dagger \tag{23b}$$

Classical least-squares, CLS (cont.)

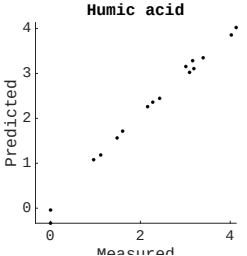
Example

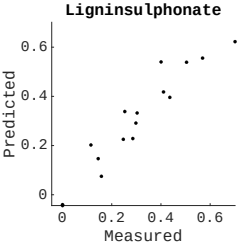
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

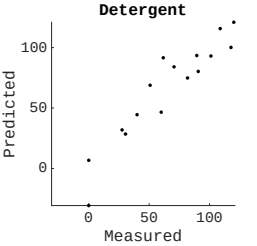
Estimate the concentration of all components using absorbance at two wavelengths

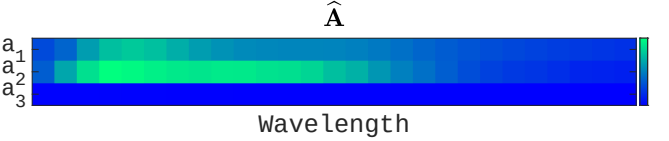
- We selected all bands in the given range ($K = 27$)
- Classical least-squares regression ($M = 3$)

Classical least-squares, CLS (cont.)









Classical least-squares, CLS (cont.)

$$\dot{\mathbf{Y}} = \dot{\mathbf{X}}\mathbf{A}^{\dagger}$$

The equation is equivalent to the centred MLR model, matrix $\mathbf{A}^{\dagger}$ is used instead of $\mathbf{B}$

- Thus, we can use the same expression for prediction

For test sample with spectrum $\mathbf{z} = [z_1, z_2, \ldots, z_K]$ and unknown composition, we have

$$\rightsquigarrow \quad \widehat{\mathbf{y}} = \overline{\mathbf{y}} + (\mathbf{z} - \overline{\mathbf{x}})\,\widehat{\mathbf{A}}^{\dagger} \quad \text{(predicted composition)} \tag{24}$$

We used $\widehat{\mathbf{A}}^{\dagger} = \mathbf{A}^{\dagger\,T}\left(\mathbf{A}\,\dagger\,\mathbf{A}^{\dagger\,T}\right)^{-1}$, the right pseudo-inverse of $\widehat{\mathbf{A}}$ (instead of $\widehat{\mathbf{B}}$)

The calibration of each y-variable involves information from all of the other y-variables

- Because in classical least-squares all components are evaluated simultaneously

Classical least-squares, CLS (cont.)

Applicability of classical least-squares calibration

Condition for existence of the least-squares solution to the a CLS calibration problem

- The $M \times M$ matrix $\dot{\mathbf{Y}}^T\dot{\mathbf{Y}}$ must be non-singular

$\rightsquigarrow$ (Equivalently, $\mathbf{V}_Y$ is non-singular)

- The $M \times M$ matrix $\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T$ must be non singular

$\rightsquigarrow$ (Needed to compute $\widehat{\mathbf{A}}^\dagger$)

This requires that a number of practical conditions on the calibration data are satisfied

- $K \geq M$, there are at least as many wavelengths as constituents, $\dot{\mathbf{A}}$ is non-singular
- $N > M$, there must be more sample than components, $\dot{\mathbf{Y}}$ is non-singular

Classical least-squares, CLS (cont.)

Advantages of classical least-squares

By definition, CLS calibration has a direct relation with the Beer-Lamber law

- The model has no limit in the number of wavelengths
- It can handle well multicollinearity in the spectra
- (The entire **X**-block can be used, also noisy parts)

Drawbacks of classical least-squares

By definition, CLS requires that all components in the sample are accounted for

- This requirement may be practically impossible to satisfy
- (This is an important practical difference from MLR)

Calibration models, so far

Chemometric data analysis

Calibration models, so far

Multivariate multiple linear regression (MLR)

- Model: $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$
- Estimation: $\widehat{\mathbf{B}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$
- Prediction: $\widehat{\mathbf{y}} = \overline{\mathbf{y}} + (\mathbf{z} - \overline{\mathbf{x}})\,\widehat{\mathbf{B}}$

Classical least-squares (CLS)

- Model: $\mathbf{X} = \mathbf{Y}\mathbf{A} + \mathbf{E}$
- Estimation: $\widehat{\mathbf{A}} = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{X}$
- Prediction: $\widehat{\mathbf{y}} = \overline{\mathbf{y}} + (\mathbf{z} - \overline{\mathbf{x}})\,\widehat{\mathbf{A}}^{\dagger}$

Principal components and partial least-squares regression (PCR and PLSR)

- Model: $\mathbf{Y} = \mathbf{T}\mathbf{B} + \mathbf{E}$
- Estimation: $\widehat{\mathbf{B}} = \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{Y}$
- Prediction: $\widehat{\mathbf{y}} = \overline{\mathbf{y}} + \widehat{\mathbf{t}}\widehat{\mathbf{B}}$