Aalto University

Chemometric data analysis, fundamental methods (I)

Advanced crystallization and characterization techniques
June 1-5, 2020

Francesco Corona

Chemical and Metallurgical Engineering
School of Chemical Engineering

Fundamental chemometric data analysis methods

The term **chemometrics** was introduced in 1972 by Svante Wold (Swedish chemist)

- From slow and specialised wet-lab chemistry methods
- To general instrument- and model-based methods

How to use spectroscopy to determine concentrations in samples of various constituents

- Constituents absorb light in overlapping frequency regions

Idea: When the constituents do not absorb light in separated frequency regions, one must utilise a combination of many spectral frequencies to estimate the concentrations

> ⇝ **Multivariate calibration**[1]
>
> The problem of how to combine absorptions at several frequencies (or other chemical and physical sensor measurements) to approximate a measured set of concentrations (or other properties of the material under study) is called **multivariate calibration**

[1] Wold S, Martens H and Wold H (1983) The multivariate calibration problem in chemistry solved by the PLS method.

Fundamental chemometric data analysis methods (cont.)

Suppose that we are interested in the content of protein and water in grain (original)

- ⇝ Protein analysis by applying the Kejdahl method
- ⇝ Water by weighing normal and dried samples

In the spectroscopic method, we first lead infrared light through a number of samples

- ⇝ We measure the light absorption at a number of frequencies, for all samples
- ⇝ We also measure their protein and water concentration by wet-chemistry

Then, we use data to reconstruct the relation between absorbances and concentration

- This model is then used to estimate the concentration of unknown samples

---

Spectroscopy is fast, non-destructive and often it does not require sample preparation

- We need to couple the instrument/computer system to collect data
- We need to learn appropriate models using statistical tools

Fundamental chemometric data analysis methods (cont.)

We give an introduction to the fundamental techniques of chemometric data analysis

- A basic knowledge of matrix algebra and elementary statistics is needed

We shall focus on multiple and multivariate regression techniques

- ⇝ **Calibration data**, **overview**
- ⇝ **Classical least-squares**, **CLS**
- ⇝ **Multiple linear regression**, **MLR**
- ⇝ **Principal component regression**, **PCR**
- ⇝ **(Partial least-squares regression**, **PLSR)**

Fundamental chemometric data analysis methods (cont.)

We will mostly refer to multivariate data from NIR spectroscopy, and concentrations

- The methods can be used on multivariate data from other sensor technologies
- (Mass spectroscopy, Raman spectroscopy, chemical imaging, ...)

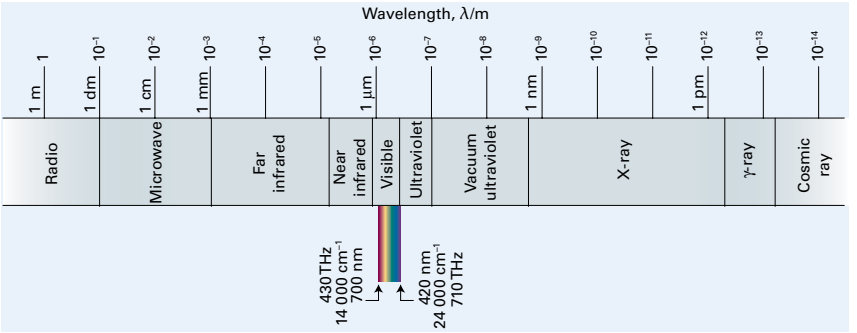
The same generality applies to other fundamental variables in crystallisation processes

- (Crystal size distribution, crystal shape, polymorphic form, ...)

Chemometric data analysis can be used for many types of multivariate data

The NIR frequency band

The electromagnetic spectrum [Atkin's, Physical chemistry] with wavelengths in [nm]



- Ultraviolet light, $1 - 400$ [nm]
- Visible light, $400 - 750$ [nm]
- **Infrared light** $750 - 10^6$ [nm]

⇝ **Near infrared light** $750 - 2.5K$ [nm]
⇝ **Mid infrared light** $2.5K - 1.6K$ [nm]
⇝ **Far infrared light** $1.6K - 1M$ [nm]

The NIR frequency band (cont.)

Spectroscopic absorption originates from molecular vibrations at different frequencies

- Fundamental vibrations found in the MIR band (Raman spectroscopy)
- Overtones and combinations in the NIR band (NIR spectroscopy)

Bond vibration	Structure	Wavelength [nm]
C-H stretch (2nd)	Aromatic	1143
C-H stretch (2nd)	-CH_3	1152
C-H stretch (2nd)	-CH_2	1215
C-H stretch (2nd)	-CH	1225

Liquid materials

- Transmission spectroscopy
- $(750 - 1100$ [nm])

Powdered materials

- Reflection spectroscopy
- $(1100 - 2500$ [nm])

Spectra provide a complex fingerprint of the sample's molecular constituents

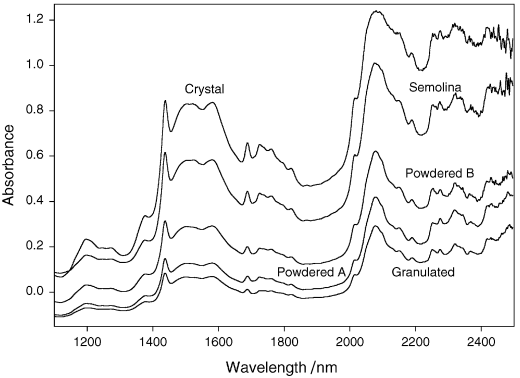
The NIR frequency band (cont.)

Chemical bonds in molecular structures are associated with characteristic wavelengths

- The different characteristics in the NIR range overlap
- (Different from a typical GC line spectrum)
- (No well defined features, spread peaks)

Spectra of different types of sucrose

- Analyst (2001)
- Blanco and Romero



Beer-Lambert's law

Suppose that we are given a sample, often a compound of several chemical components

The **absorbance of a single chemical component, at a particular wavelength**

$$x = -\log\left(\frac{I}{I_0}\right), \quad (x > 0)$$

I_0, original intensity of the incident light

- Before the sample is inserted

I, intensity of transmitted/reflected light

- After the sample is inserted

For transmission (reflection) spectroscopy $T = I_0/I$ is the transmittance (reflectance)

- Increasing concentration y will decrease transmittance or reflectance, T
- Then, also absorbance x will increase, according to some $x = g(y)$

Beer-Lambert's law (cont.)

Beer-Lambert law (Single component case)

$$x = y\,a + e \quad (1)$$

$a = \varepsilon\delta$ denotes the absorbance of the pure component

- ε, absorption coefficient (component specific)
- δ, path length of the incident light

Because the Beer-Lamber law may not hold exactly, therefore e is added as model error

---

Beer-Lambert law (Multiple component case)

$$x = y_1\,a_1 + y_2\,a_2 + \cdots + y_M\,a_M + e \quad (2a)$$

$$= \sum_{m=1}^{M} y_m\,a_m + e \quad (2b)$$

- y_m, concentration of component m
- a_m, absorbance of component m

$\rightsquigarrow$ M, number of components

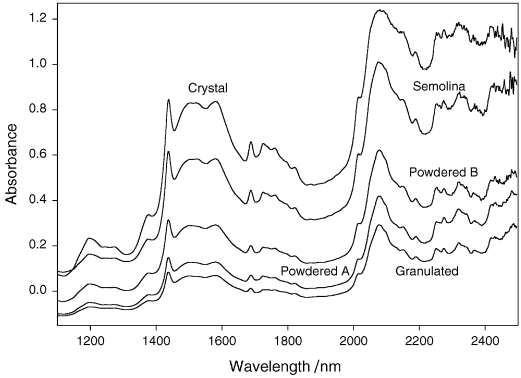
For closed systems where have that all components are analysed (known concentration)

- We have that $y_1 + y_2 + \cdots + y_M = 1$ (compactly, $\sum_{m=1}^{M} y_m = 1$)

Calibration data

Chemometric data analysis

Calibration data



Spectral data

Spectral data for an individual material sample amounts to a collection of absorbances

- At a number K of individual wavelengths
- $\{x_1, x_2, \ldots, x_k, \ldots, x_K\}$

Calibration data (cont.)

At each k-th wavelength, Beer-Lambert law for a system consisting of M components

$$x_k = y_1\, a_{1k} + y_2\, a_{2k} + \cdots + y_M\, a_{Mk} + e_k \tag{3a}$$

$$= \sum_{m=1}^{M} y_m\, a_{mk} + e_k, \quad (k = 1, \ldots, K) \tag{3b}$$

Considering the all the $k = 1, \ldots, K$ wavelengths, we have

$$\begin{aligned}
x_1 &= y_1\, a_{11} + y_2\, a_{21} + \cdots + y_M\, a_{M1} + e_1 \\
x_2 &= y_1\, a_{12} + y_2\, a_{22} + \cdots + y_M\, a_{M2} + e_2 \\
&\vdots \\
x_k &= y_1\, a_{1k} + y_2\, a_{2k} + \cdots + y_M\, a_{Mk} + e_k \\
&\vdots \\
x_K &= y_1\, a_{1K} + y_2\, a_{2K} + \cdots + y_M\, a_{MK} + e_K
\end{aligned}$$

Calibration data (cont.)

Beer-Lambert law (Multiple wavelength, multiple component case)

$$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{M1} \\ a_{12} & a_{22} & \cdots & a_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1K} & a_{2K} & \cdots & a_{MK} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}}_{\mathbf{y}} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_K \end{bmatrix}}_{\mathbf{e}} \tag{5}$$

- The measured spectrum, as a column vector

$$\mathbf{x} = (x_1, x_2, \ldots, x_K)'$$

- The spectra of the pure components, as column vectors

$$\mathbf{a}_m = (a_{m1}, a_{m2}, \ldots, a_{mK})', \quad (m = 1, \ldots, M)$$

- The concentrations of the components, as a column vector

$$\mathbf{y} = (y_1, y_2, \ldots, y_M)'$$

Calibration data

Often we have multiple samples, calibration data consists of an **X**-block and a **Y**-block

- N samples (absorption and concentrations)

Let $N = 4$ be the number of material's samples consisting of $M = 2$ be components

- Let $K = 3$ be the number of wavelengths

Y-block, M concentrations measured by some reference (wet) method

- One row for each sample

$$\mathbf{Y} = \underbrace{\begin{bmatrix} [y_{1,1} & y_{1,2}] \\ [y_{2,1} & y_{2,2}] \\ [y_{3,1} & y_{3,2}] \\ [y_{4,1} & y_{4,2}] \end{bmatrix}}_{4\times 2} = \begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \mathbf{y}_3' \\ \mathbf{y}_4' \end{bmatrix}$$

X-block, K-dimensional absorption spectra by a NIR instrument

- One row for each sample

$$\mathbf{X} = \underbrace{\begin{bmatrix} [x_{1,1} & x_{1,2} & x_{1,3}] \\ [x_{2,1} & x_{2,2} & x_{2,3}] \\ [x_{3,1} & x_{3,2} & x_{3,3}] \\ [x_{4,1} & x_{4,2} & x_{4,3}] \end{bmatrix}}_{4\times 3} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \mathbf{x}_3' \\ \mathbf{x}_4' \end{bmatrix}$$

Calibration data (cont.)

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{E}$$

We are interested in estimating how the $\mathbf{Y}$-block varies with the $\mathbf{X}$-block, function f

- (We are interested in estimating concentrations $\mathbf{y}$ from spectra $\mathbf{x}$)
- (We shall use only the given calibration data, blocks $\mathbf{X}$ and $\mathbf{Y}$)
- (We shall assume that function f is some unknown matrix, $\mathbf{B}$)

Notice how this is the inverse problem of what Beer-Lambert law models, $\mathbf{X} = g(\mathbf{Y}) + \mathbf{E}$

- (Spectra $\mathbf{x}$ from concentrations $\mathbf{y}$, and pure component spectra $\mathbf{a}$)
- (Beer-Lambert law assumes that function g is a matrix, $\mathbf{A}$)

Calibration data (cont.)

$$\mathbf{Y} = f(\mathbf{X}) + \mathbf{E}$$

Remember, not restricted to concentrations, other material's properties could be used

- Any property depending on concentration can be estimated from spectra
- (The property must be dependent of the sample type and composition)

To develop the treatment, we primarily use concentrations ($\mathbf{Y}$) and NIR spectra ($\mathbf{X}$)

Calibration data (cont.)

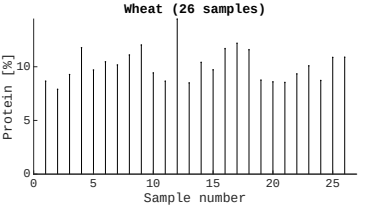
Example

Concentration of proteins and water in grain samples, from NIR spectra

Two concentrations ($M = 2$), 5-wavelength ($K = 5$) spectra, $N = 26$ samples

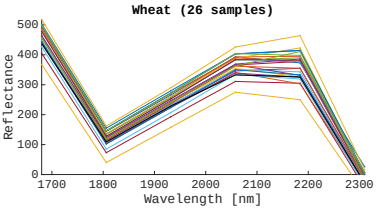
$\mathbf{Y}$-block, $M = 2$ concentrations

$$\mathbf{Y} = \underbrace{\begin{bmatrix} y_{1,1} & y_{1,2} \\ y_{2,1} & y_{2,2} \\ \vdots & \vdots \\ y_{26,1} & y_{26,2} \end{bmatrix}}_{26 \times 2}$$



$\mathbf{X}$-block, $K = 5$ absorption bands

$$\mathbf{X} = \underbrace{\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,5} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,5} \\ \vdots & \vdots & \ddots & \vdots \\ x_{26,1} & x_{26,2} & \cdots & x_{26,5} \end{bmatrix}}_{26 \times 5}$$



Calibration data (cont.)

Y-block, $M = 2$ concentrations

$$\mathbf{Y} = \underbrace{\begin{bmatrix} y_{1,1} & y_{1,2} \\ [y_{2,1} & y_{2,2}] \\ \vdots & \vdots \\ y_{26,1} & y_{26,2} \end{bmatrix}}_{26 \times 2}$$

The concentration of protein and water in the second sample

$$\mathbf{y}_2 = \begin{bmatrix} y_{2,1} & y_{2,2} \end{bmatrix}$$

X-block, $K = 5$ absorption spectra

$$\mathbf{X} = \underbrace{\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,5} \\ [x_{2,1} & x_{2,2} & \cdots & y_{2,3}] \\ \vdots & \vdots & \ddots & \vdots \\ x_{26,1} & x_{26,2} & \cdots & x_{26,5} \end{bmatrix}}_{26 \times 5}$$

The spectrum of the second sample

$$\mathbf{x}_2 = \begin{bmatrix} x_{21} & x_{22} & \cdots & x_{25} \end{bmatrix}$$

Test data

Test data consists of one or more spectra of samples of unknown composition

$$\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_K \end{bmatrix}$$

The calibration model can be used to predict the unknown concentrations

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 & \widehat{y}_2 & \cdots & \widehat{y}_m \end{bmatrix}$$

$\rightsquigarrow$ 'Hats' are used to denote predictions and estimates

We consider models that make predictions $\widehat{\mathbf{y}}$ of the form

$$\underbrace{\widehat{\mathbf{y}}}_{(M\times 1)} = \underbrace{\widehat{\mathbf{B}}}_{(M\times K)} \underbrace{\mathbf{z}}_{(K\times 1)} \tag{6}$$

- $\widehat{\mathbf{B}}$ is a matrix of regression coefficients
- It is learned from calibration data

Test data (cont.)

Example

Concentration of proteins and water in grain samples, from NIR spectra

Two concentrations ($M = 2$), 5-wavelength ($K = 5$) spectra, $N = 26$ samples

$\mathbf{X}$-block, $K = 3$ frequency spectra

$$\mathbf{X} = \underbrace{\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,5} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,5} \\ \vdots & \vdots & \ddots & \vdots \\ x_{26,1} & x_{26,2} & \cdots & x_{26,5} \end{bmatrix}}_{26\times 5}$$

$\mathbf{Y}$-block, $M = 2$ concentrations

$$\mathbf{Y} = \underbrace{\begin{bmatrix} y_{1,1} & y_{1,2} \\ y_{2,1} & y_{2,2} \\ \vdots & \vdots \\ y_{26,1} & y_{26,2} \end{bmatrix}}_{26\times 2}$$

The test data, only absorbances at each of the five frequencies are given

$$\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_5 \end{bmatrix}$$

The concentrations are unknown, must be estimated by the model

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 \end{bmatrix}$$

Basic statistics

Consider the data matrices $\mathbf{X}$ and $\mathbf{Y}$, we shall assume that there are no missing data

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,K} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,K} \end{bmatrix}$$

Explanatory (input) variables

- N data points, the samples (rows)
- K easy-to-measure variables (absorbances)

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,M} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,M} \end{bmatrix}$$

Response (output) variables

- N data points, the samples (rows)
- M hard-to-measure variables (concentrations)

The columns of $\mathbf{X}$ and $\mathbf{Y}$ will be denoted as variables, their rows are the observations

- $\mathbf{x}$ ($N \times 1$), the columns of $\mathbf{X}$ (absorbance of all samples at some wavelength)
- $\mathbf{y}$ ($N \times 1$), the columns of $\mathbf{Y}$ (concentration of all samples of some component)

For each variable and sample, we plot them and then compute descriptive statistics

- `min`, `max`, `mean`, `standard deviation`, `variance`, ...

Basic statistics (cont.)

Plot of the explanatory variables (spectral plot)

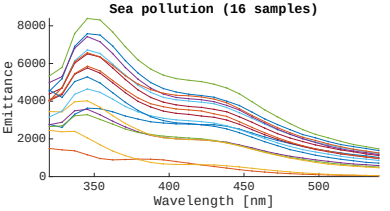
- Each row of $\mathbf{X}$ is plotted as function of the column variable k

Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Emission intensity spectra of the collected seawater samples

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,K} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,K} \end{bmatrix}$$



- $K = 27$ wavelengths
- $N = 16$ samples

Each emission intensity must be non-negative and must behave reasonably (smooth)

Basic statistics (cont.)

Plot of the response variables (composition plot)

⇝ Each row of $\mathbf{Y}$ is plotted as function of the column variable m

Example

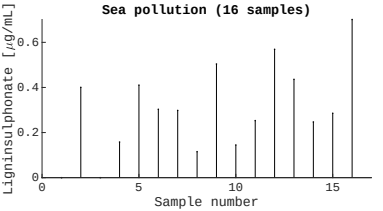
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

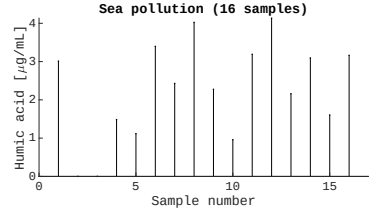
Humic acid, Ligninsulphonate, and also Detergents are found

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,M} \\ \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,M} \end{bmatrix}$$

- $M = 3(2)$ concentrations
- $N = 16$ samples

Concentrations must take on non-negative values only





Basic statistics (cont.)

Empirical (statistical) quantities are properties of the N observations (the samples)

- For a given variable (column of either block)

Let $\mathbf{x}$ $(N \times 1)$ be a column of $\mathbf{X}$ (absorbances at a specific wavelength, all samples)

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,k} & \cdots & x_{1,K} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,k} & \cdots & x_{n,K} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,k} & \cdots & x_{N,K} \end{bmatrix} \rightsquigarrow \mathbf{x} = [x_{1,k}, x_{2,k}, \ldots, x_{N,k}] = [x_1, x_2, \ldots, x_N]$$

Let $\mathbf{y}$ $(N \times 1)$ be a column of $\mathbf{Y}$ (absorbances of a specific component, all samples)

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,k} & \cdots & y_{1,M} \\ \vdots & \ddots & \vdots & & \vdots \\ y_{n,1} & \cdots & y_{n,m} & \cdots & y_{n,M} \\ \vdots & & \vdots & \ddots & \vdots \\ y_{N,1} & \cdots & y_{N,m} & \cdots & y_{N,M} \end{bmatrix} \rightsquigarrow \mathbf{y} = [y_{1,m}, y_{2,m}, \ldots, y_{N,m}] = [y_1, y_2, \ldots, y_N]$$

Basic statistics (cont.)

The sample mean of x: (An estimate of the) Expected value of x, its average

$$\overline{x} = \frac{1}{N}(x_1 + x_2 + \cdots + x_N) \tag{7a}$$

$$\rightsquigarrow \frac{1}{N}\mathbf{1}^T\mathbf{x} \tag{7b}$$

- The vector of means of $\mathbf{X}$, a collection of averages

$$\overline{\mathbf{x}} = \begin{bmatrix}\overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_k & \cdots & \overline{x}_K\end{bmatrix}, \quad \rightsquigarrow \frac{1}{N}\mathbf{1}^T\mathbf{X} \tag{8}$$

The sample mean of y: (An estimate of the) Expected value of y, its average

$$\overline{y} = \frac{1}{N}(y_1 + y_2 + \cdots + y_N) \tag{9a}$$

$$\rightsquigarrow \frac{1}{N}\mathbf{1}^T\mathbf{y} \tag{9b}$$

- The vector of means of $\mathbf{Y}$, a collection of averages

$$\overline{\mathbf{y}} = \begin{bmatrix}\overline{y}_1 & \overline{y}_2 & \cdots & \overline{y}_m & \cdots & \overline{y}_M\end{bmatrix}, \quad \rightsquigarrow \frac{1}{N}\mathbf{1}^T\mathbf{Y} \tag{10}$$

$\mathbf{1}$ $(N \times 1)$, a column-vector of ones

Basic statistics (cont.)

The sample variance of x: Expected squared deviation of x from its mean $\overline{x}$

$$s_x^2 = \frac{1}{N-1}\left[(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + \cdots + (x_N-\overline{x})^2\right] \tag{11a}$$

$$\rightsquigarrow \frac{1}{N-1}(\mathbf{x}-\overline{x}\mathbf{1})^T(\mathbf{x}-\overline{x}\mathbf{1}) = \frac{1}{N-1}||\mathbf{x}-\overline{x}\mathbf{1}||^2 \tag{11b}$$

- The sample standard deviation of $\mathbf{x}$, $s_x = \sqrt{s_x^2}$

The sample variance of y: Expected squared deviation of y from its mean $\overline{y}$

$$s_y^2 = \frac{1}{N-1}\left[(y_1-\overline{y})^2 + (y_2-\overline{y})^2 + \cdots + (y_N-\overline{y})^2\right] \tag{12a}$$

$$\rightsquigarrow \frac{1}{N-1}(\mathbf{y}-\overline{y}\mathbf{1})^T(\mathbf{y}-\overline{y}\mathbf{1}) = \frac{1}{N-1}||\mathbf{y}-\overline{y}\mathbf{1}||^2 \tag{12b}$$

- The sample standard deviation of $\mathbf{y}$, $s_y = \sqrt{s_y^2}$

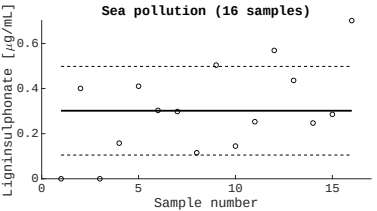
Basic statistics (cont.)

Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

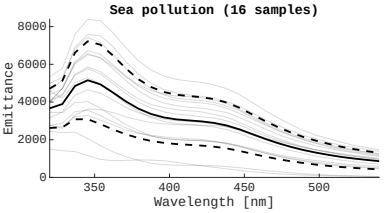
The mean concentration $\overline{y}_2$ $(-)$

- Ligninsulphonate
- $\overline{y_2} \pm s_{y2}$ $(--)$



The mean spectrum $\overline{\mathbf{x}}$ $(-)$

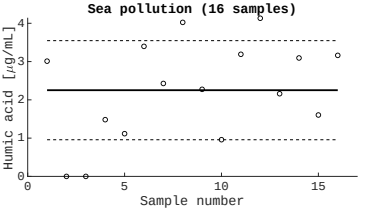
- $\overline{\mathbf{x}} \pm \mathbf{1} s_x$ $(--)$



Basic statistics (cont.)

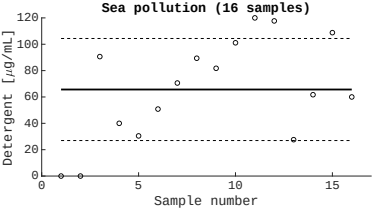
The mean concentration $\overline{y}_1$ $(-)$

- $\overline{y_1} \pm s_{y1}$ $(--)$
- Humic acid



The mean concentration $\overline{y}_3$ $(-)$

- $\overline{y_3} \pm s_{y3}$ $(--)$
- Detergent



Basic statistics (cont.)

The sample covariance of x and y: Expected product of deviations from means

$$v_{xy} = \frac{1}{N-1}[(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_N - \overline{x})(y_N - \overline{y})] \quad (13a)$$

$$= \frac{1}{N-1}(\mathbf{x} - \overline{x}\mathbf{1})^T(\mathbf{y} - \overline{y}\mathbf{1}) \quad (13b)$$

The sample covariance of y and x: Expected product of deviations from means

$$v_{yx} = \frac{1}{N-1}[(y_1 - \overline{y})(x_1 - \overline{x}) + (y_2 - \overline{y})(x_2 - \overline{x}) + \cdots + (y_N - \overline{y})(x_N - \overline{x})] \quad (14a)$$

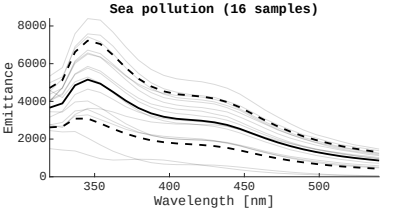
$$= \frac{1}{N-1}(\mathbf{y} - \overline{y}\mathbf{1})^T(\mathbf{x} - \overline{x}\mathbf{1}) \quad (14b)$$

Clearly, we have that $v_{xy} = v_{yx}$, and that $v_{xx} = s_x^2$ and $v_{yy} = s_y^2$

Basic statistics (cont.)

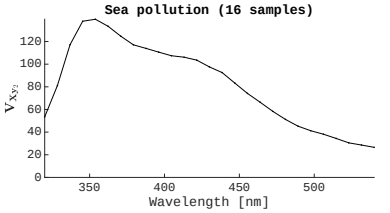
Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)



A vector of covariances v_{xy_2}

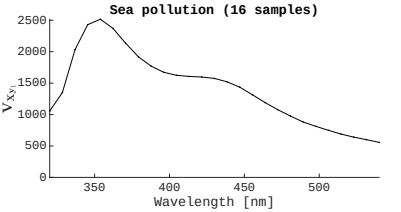
- Ligninsulphonate
- v_{xy_2} $(-)$



Basic statistics (cont.)

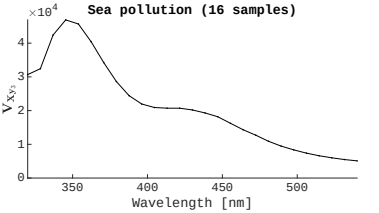
A vector of covariances v_{xy_1}

- Humic acid
- v_{xy_1} $(-)$



A vector of covariances v_{xy_3}

- Detergent
- v_{xy_3} $(-)$



Basic statistics (cont.)

Data standardisation

Centering: We center $(N \times 1)$ vectors $\mathbf{x}$ and $\mathbf{y}$, by subtracting their mean $\overline{x}$ and $\overline{y}$

We get two new $(N \times 1)$ vectors $\dot{\mathbf{x}}$ and $\dot{\mathbf{y}}$

$$\dot{\mathbf{x}} = \mathbf{x} - \mathbf{1}\overline{x} \tag{15a}$$

$$\rightsquigarrow \mathbf{x} = \dot{\mathbf{x}} + \mathbf{1}\overline{x} \tag{15b}$$

$\rightsquigarrow$ Their mean is equal zero

- $\overline{\dot{\mathbf{x}}} = 0$ and $\overline{\dot{\mathbf{y}}} = 0$

$\rightsquigarrow$ Variance is unchanged

- $s^2_{\dot{x}} = \dfrac{1}{N-1}\dot{\mathbf{x}}^T\dot{\mathbf{x}}$
- $s^2_{\dot{y}} = \dfrac{1}{N-1}\dot{\mathbf{y}}^T\dot{\mathbf{y}}$

$$\dot{\mathbf{y}} = \mathbf{y} - \mathbf{1}\overline{y} \tag{16a}$$

$$\rightsquigarrow \mathbf{y} = \dot{\mathbf{y}} + \mathbf{1}\overline{y} \tag{16b}$$

The corresponding centred matrices $\dot{\mathbf{X}}$ and $\dot{\mathbf{Y}}$, size $(N \times K)$ and $(N \times M)$ respectively

$$\dot{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}} \tag{17a}$$

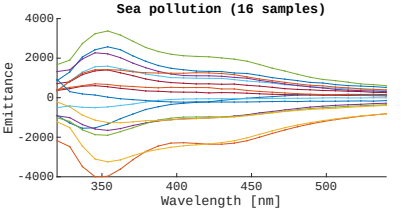
$$\dot{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\overline{\mathbf{y}} \tag{17b}$$

Basic statistics (cont.)

Example

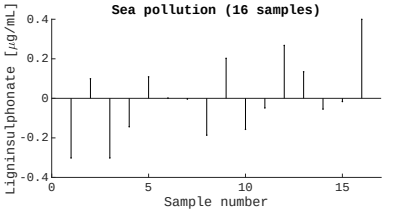
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Each row of $\dot{\mathbf{X}}$ is plotted as function of the variable k



The centred concentrations $\dot{\mathbf{y}}_2$

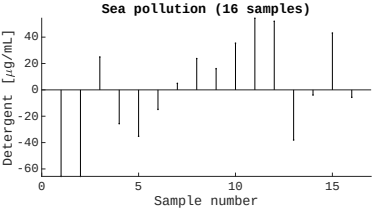
- Ligninsulfonate



Basic statistics (cont.)

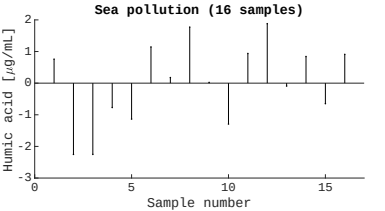
The centred concentrations $\dot{\mathbf{y}}_1$

- Humic acid



The centred concentrations $\dot{\mathbf{y}}_3$

- Detergent



Basic statistics (cont.)

Scaling: The idea is to make the columns of $\mathbf{Y}$ have the same standard deviation s_y

- ⇝ This is only needed when the measurements units of $\mathbf{y}$s are different
- ⇝ (Spectral variables $\mathbf{x}$ need not be scaled, same absorbance units)

We can replace each column $\mathbf{y}$ by $\mathbf{y}_{\text{scaled}}$

⇝ $s_{y_{\text{scaled}}} = 1$ ⇝ $\mathbf{y}_{\text{scaled}} = \frac{1}{s_y}\mathbf{y}$

Scaling: Make the columns of $\mathbf{Y}$ have zero mean and the same standard deviation s_y

We can replace each column $\mathbf{y}$ by $\mathbf{y}_{\text{autoscaled}}$

⇝ $\overline{y}_{\text{autoscaled}} = 0$

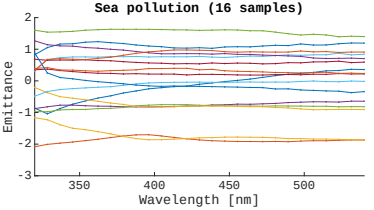
⇝ $s_{y_{\text{autoscaled}}} = 1$ ⇝ $\mathbf{y}_{\text{autoscaled}} = \frac{1}{s_y}\dot{\mathbf{y}}$

Basic statistics (cont.)

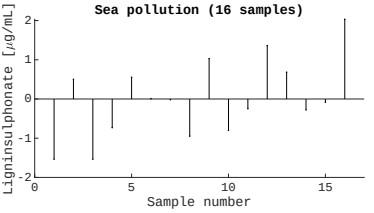
Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Each row of $\mathbf{X}_{\text{scaled}}$ is plotted as function of the variable k

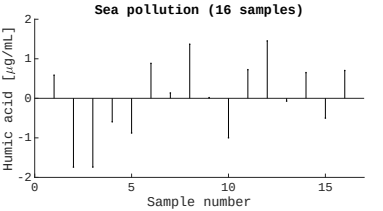


The autoscaled concentrations of ligninsulpfonate, $\mathbf{y}_{2,\text{autoscaled}}$

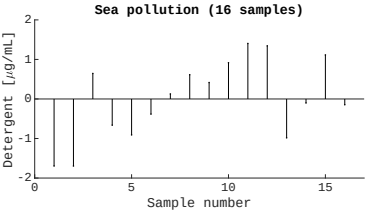


Basic statistics (cont.)

The autoscaled concentrations of humic acid, $\mathbf{y}_{1,\text{autoscaled}}$



The autoscaled concentrations of detergent, $\mathbf{y}_{3,\text{autoscaled}}$



Basic statistics (cont.)

We can estimate the sample (**variance-**)**covariance** matrices of variables X and Y

Explanatory variables

- Size $(K \times K)$
- $\mathbf{V}_X = \mathbf{V}_X{}^T$

$$\mathbf{V}_X = \frac{1}{n-1}\dot{\mathbf{X}}^T\dot{\mathbf{X}} \tag{18a}$$

$$= \begin{bmatrix} s^2_{x1} & v_{x1,x2} & \cdots & v_{x1,xK} \\ v_{x2,x1} & s^2_{x2} & \cdots & v_{x2,xK} \\ \vdots & \vdots & \ddots & \vdots \\ v_{xK,x1} & v_{xK,x2} & \cdots & v_{xK,xK} \end{bmatrix} \tag{18b}$$

Response variables

- Size $(M \times M)$
- $\mathbf{V}_Y = \mathbf{V}_Y{}^T$

$$\mathbf{V}_Y = \frac{1}{n-1}\dot{\mathbf{Y}}^T\dot{\mathbf{Y}} \tag{19a}$$

$$= \begin{bmatrix} s^2_{y1} & v_{y1,y2} & \cdots & v_{y1,yM} \\ v_{y2,y1} & s^2_{y2} & \cdots & v_{y2,yM} \\ \vdots & \vdots & \ddots & \vdots \\ v_{yK,y1} & v_{yK,y2} & \cdots & v_{yK,yK} \end{bmatrix} \tag{19b}$$

Basic statistics (cont.)

We can also estimate the sample (**variance-**)**covariance** matrix between X and Y

Explanatory and response variables

- Size $(K \times M)$
- $\mathbf{V}_{X\,Y} = \mathbf{V}_{Y\,X}^T$

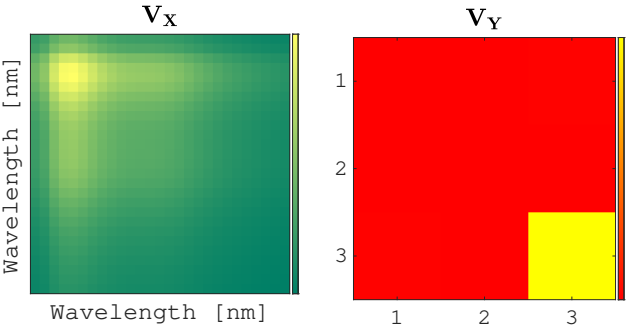
$$\mathbf{V}_{X\,Y} = \frac{1}{n-1}\dot{\mathbf{X}}^T\dot{\mathbf{Y}} \tag{20a}$$

$$= \begin{bmatrix} v_{x1,y1} & v_{x1,y2} & \cdots & v_{x1,yM} \\ v_{x2,y1} & v_{x2,y2} & \cdots & v_{x2,yM} \\ \vdots & \vdots & \ddots & \vdots \\ v_{xK,y1} & v_{xK,y2} & \cdots & v_{xK,yM} \end{bmatrix} \tag{20b}$$

Basic statistics (cont.)

Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

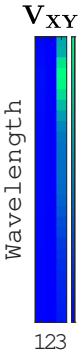


The individual variance-covariance matrices of the **X**- and the **Y**-block, respectively

- $\mathbf{V_Y}$ is dimension $(M \times M)$
- $\mathbf{V_X}$ is dimension $(K \times K)$

Notation for the **y** variables: 1) Humic acid; 2) Lignisupfonate; and 3) Detergent

Basic statistics (cont.)



How variation in $\mathbf{y}$ vars is explained by variation in $\mathbf{x}$ vars

- The variance-covariance matrix between blocks
- $\mathbf{V_X Y}$ is dimension $(K \times M)$

Notation:

1. Humic acid
2. Lignisupfonate
3. Detergent

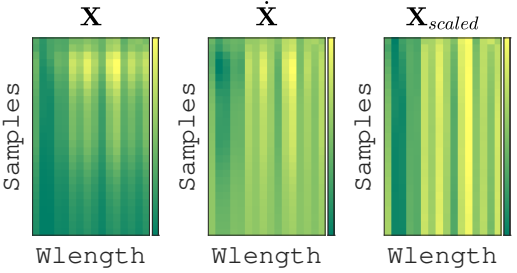
Basic statistics (cont.)

Example

Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Preprocessing of the $\mathbf{X}$-block

$$\mathbf{X} \rightsquigarrow \dot{\mathbf{X}} \rightsquigarrow \mathbf{X}_{\text{scaled}}$$

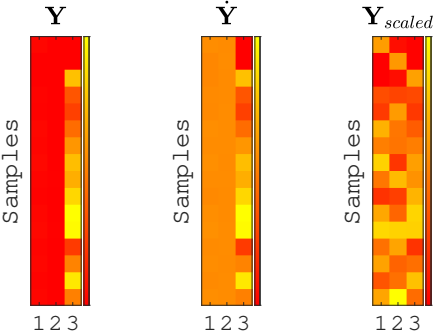


Basic statistics (cont.)

Preprocessing of the $\mathbf{Y}$-block

$$\mathbf{Y} \rightsquigarrow \dot{\mathbf{Y}} \rightsquigarrow \mathbf{Y}_{\text{scaled}}$$

1. Humic acid
2. Lignisupfonate
3. Detergent



Simple linear regression

Chemometric data analysis

Simple linear regression

Suppose that we are interested in estimating the concentration of a single component

- Also, suppose that we want to use the absorbance at a single wavelength

$\rightsquigarrow$ One explanatory input and one response output variable only

---

For the task, we are given data and we consider the **simple linear regression model**

$$y_n = (c + b x_n) + \varepsilon_n, \quad (n = 1, 2, \ldots, N) \tag{21}$$

N is the sample size, the number of available data, (x_n, y_n) pairs[2]

- x_n, n-th value of the explanatory (absorbance) variable
- 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 absorbance, up to errors

[2] The minimum number of observations is required to at least equal to 2.

Simple linear regression (cont.)

Linear regression in vector-scalar form

$$y_n = (c + bx_n) + \varepsilon_n, \quad (n = 1, 2, \ldots, N)$$

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

- c, the intercept of the regression model (a line)
- b, the slope of the regression model (a line)

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), regression coefficients
- σ^2, residual variance

$$\theta = (c, b, \sigma^2)$$

Simple linear regression (cont.)

We can stack the N model equations

$$y_n = c + bx_n + \varepsilon_i \quad (22a)$$
$$= 1c + bx_n + \varepsilon_i \quad (22b)$$

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

$$\rightsquigarrow \quad \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}}_{\mathbf{1}c} c + \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ \vdots \\ x_{N-1} \\ x_N \end{bmatrix}}_{\mathbf{x}b} b + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \\ \vdots \\ \varepsilon_{N-1} \\ \varepsilon_N \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

We can rewrite the linear regression model in vector form

$$\mathbf{y} = \mathbf{1}c + \mathbf{x}b + \boldsymbol{\varepsilon} \quad (23a)$$
$$= \mathbf{1}c + \underbrace{(\dot{\mathbf{x}} + \mathbf{1}\overline{x})}_{\text{centring}} b + \boldsymbol{\varepsilon} \quad (23b)$$
$$= \mathbf{1}c + \dot{\mathbf{x}}b + \mathbf{1}(\overline{x}b) + \boldsymbol{\varepsilon} \quad (23c)$$
$$= \mathbf{1}b_0 + \dot{\mathbf{x}}b + \boldsymbol{\varepsilon} \quad (23d)$$

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

Simple linear regression (cont.)

LR, estimation

$$\mathbf{y} = \mathbf{1}b_0 + \dot{\mathbf{x}}b + \boldsymbol{\varepsilon}$$

The least-squares estimators for the regression parameters b_0 and b are the following

$$\widehat{b}_0 = \overline{y} \tag{24a}$$

$$\widehat{b} = \frac{\dot{\mathbf{x}}^T\dot{\mathbf{y}}}{\dot{\mathbf{x}}^T\dot{\mathbf{x}}} = \frac{v_{xy}}{s_x^2} \qquad (\text{with } \sigma_x > 0) \tag{24b}$$

The case $s_x^2 = 0$ is uninteresting as it corresponds to all absorbances being equal

- v_{xy}, the covariance between x and y
- (s_x^2, the variance of x)

Simple linear regression (cont.)

Consider the simple linear regression model in vector form $\mathbf{y} = \mathbf{1}b_0 + \dot{\mathbf{x}}b + \boldsymbol{\varepsilon}$

- In sample-by-sample form, $y_n = b_0 + \dot{x}_n b + \varepsilon_n = b_0 + (x_n - \overline{x})b + \varepsilon_n$

We are interested in the pair of values (b_0, b) that minimise the sum of squared errors

- **Residual sum of squares** (**RSS**) as cost function

$$\mathcal{J}(b_0, b) = \sum_{n=1}^{N} \left[\underbrace{y_n}_{\text{Measurement}} - \underbrace{(b_0 - \dot{x}_n b)}_{\text{Model prediction}} \right]^2 = \sum_{i=1}^{N} \varepsilon_i^2 \tag{25}$$

Necessary first-order optimality condition, the gradient of the cost function is zero

$$\nabla \mathcal{J}(b_0, b) = \begin{bmatrix} \dfrac{\partial \mathcal{J}(b_0, b)}{\partial b_0} \\ \dfrac{\partial \mathcal{J}(b_0, b)}{\partial b} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}$$

Simple linear regression (cont.)

We differentiate $\mathcal{J}(b_0, b)$ with respect to b_0 and set the partial to be equal to zero,

$$\frac{\partial \mathcal{J}(b_0, b)}{\partial b_0} = \frac{\partial}{\partial b_0}\left[\sum_{n=1}^{N}(y_n - b_0 - \dot{x}_n\, b)^2\right] = \sum_{n=1}^{N}\left[\frac{\partial}{\partial b_0}(y_n - b_0 - \dot{x}_n\, b)^2\right] \tag{26a}$$

$$= \sum_{n=1}^{N}\left[-2\,(y_n - b_0 - \dot{x}_n\, b)\right] = -2\sum_{n=1}^{N}(y_n - b_0 - \dot{x}_n\, b) \tag{26b}$$

$$= -2\left(\sum_{n=1}^{N} y_n - \sum_{n=1}^{N} b_0 - \sum_{n=1}^{N}\dot{x}_n\, b\right) = -2\left(\sum_{n=1}^{N} y_n - \underbrace{\sum_{n=1}^{N} b_0}_{Nb_0} - b\sum_{n=1}^{N}\underbrace{\dot{x}_n}_{x_n - \overline{x}}\right) \tag{26c}$$

$$= -2\left(\sum_{n=1}^{N} y_n - Nb_0 - 0\right) = 0 \tag{26d}$$

We get,

$$\rightsquigarrow \quad \widehat{b}_0 = \frac{1}{N}\sum_{n=1}^{N} y_n = \overline{y}$$

Simple linear regression (cont.)

We differentiate $\mathcal{J}(b_0, b)$ with respect to b and set the partial to be equal to zero,

$$\frac{\partial \mathcal{J}(b_0, b)}{\partial b_0} = \frac{\partial}{\partial b}\left[\sum_{n=1}^{N} (y_n - b_0 - \dot{x}_n b)^2\right] = \sum_{n=1}^{N}\left[\frac{\partial}{\partial b}(y_n - b_0 - \dot{x}_n b)^2\right] \tag{27a}$$

$$= \sum_{n=1}^{N} \left[-2\dot{x}_n \left(y_n - b_0 - \dot{x}_n b\right)\right] = -2\sum_{n=1}^{N} \dot{x}_n \left(y_n - b_0 - \dot{x}_n b\right) \tag{27b}$$

$$= -2\sum_{n=1}^{N} \left(\dot{x_n} y_n - \dot{x_n} b_0 - \dot{x}_n^2 b\right) = -2\left(\underbrace{\sum_{n=1}^{N} \dot{x_n} y_n}_{\dot{\mathbf{x}}^T\mathbf{y}} - b_0 \sum_{n=1}^{N} \underbrace{\dot{x_n}}_{x_n - \overline{x}} - b\underbrace{\sum_{n=1}^{N} \dot{x}_n^2}_{\dot{\mathbf{x}}^T\dot{\mathbf{x}}}\right) \tag{27c}$$

$$= -2\left(\dot{\mathbf{x}}^T\mathbf{y} - 0 - b\dot{\mathbf{x}}^T\dot{\mathbf{x}}\right) \tag{27d}$$

We get,

$$\rightsquigarrow \quad \hat{b} = \frac{\dot{\mathbf{x}}^T\mathbf{y}}{\dot{\mathbf{x}}^T\dot{\mathbf{x}}} = \frac{v_{xy}}{s_x^2} \quad (\text{with } s_x^2 > 0)$$

Simple linear regression (cont.)

Sufficient second-order optimality condition, Hessian of the cost function is PD

$$\nabla^2 \mathcal{J}(b_0, b) = \begin{bmatrix} \dfrac{\partial^2 \mathcal{J}(b_0, b)}{\partial {b_0}^2} & \dfrac{\partial^2 \mathcal{J}(b_0, b)}{\partial b_0 \partial b} \\ \dfrac{\partial^2 \mathcal{J}(b_0, b)}{\partial b \partial b_0} & \dfrac{\partial^2 \mathcal{J}(b_0, b)}{\partial b^2} \end{bmatrix} \tag{28a}$$

$$= \underbrace{\begin{bmatrix} 2N & 0 \\ 0 & \dot{\mathbf{x}}^T \dot{\mathbf{x}} \end{bmatrix}}_{\text{Positive definite}} \succ 0 \tag{28b}$$

This is always true provided that $N > 0$ (trivial, pointless) and that $\dot{\mathbf{x}}^T \dot{\mathbf{x}} > 0$

- The second condition corresponds to a positive sample variance $s_x^2 > 0$

$$s_x^2 = \frac{1}{1 - N} \left[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2 \right]$$

The least-squares estimators for the parameters of the linear regression model are

$$\widehat{b} = \frac{\dot{\mathbf{x}}^T \mathbf{y}}{\dot{\mathbf{x}}^T \dot{\mathbf{x}}}$$

$$\widehat{b_0} = \overline{y}$$

Simple linear regression (cont.)

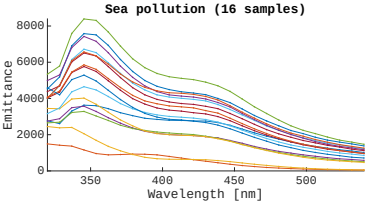
Example

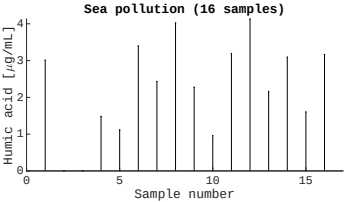
Ligninsulfonate in seawater, fluorescence spectroscopy (emission spectra)

Estimate the concentration of humic acid using absorbance at a single wavelength

- We selected as single band 345 [nm] (Remember the highest spectral peak?)
- We selected it because it is also the one of highest covariance with y_2

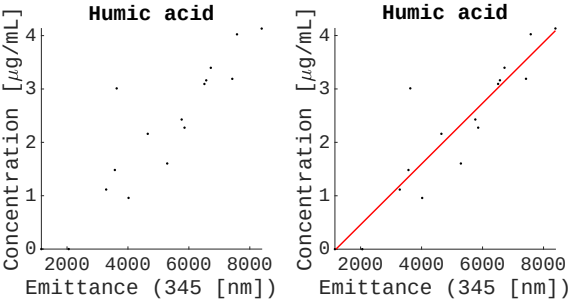
$$y_n = (c + bx_n) + \varepsilon_n, \quad (n = 1, 2, \ldots, N)$$





Simple linear regression (cont.)

Case I: Original (no centring, no scaling) **x**-variable and y_2-variable

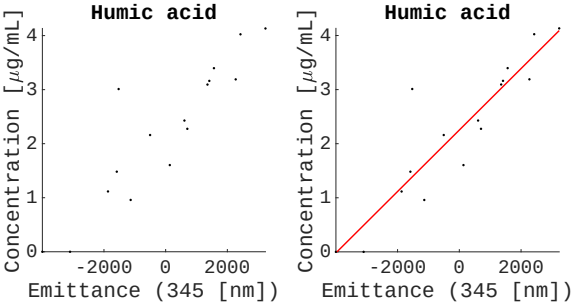


The estimated regression parameters

- Intercept, $c = -0.6775$
- Slope, $b = 0.0006$

Simple linear regression (cont.)

Case II: Centred **x**-variable and original y_2-variable

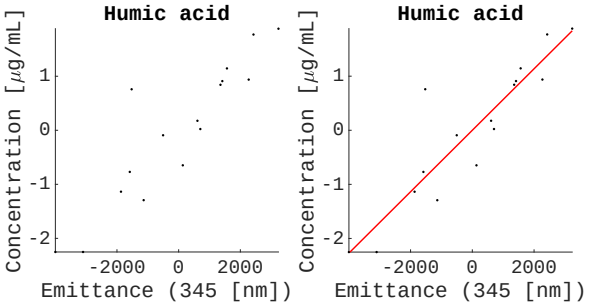


The estimated regression parameters

- Intercept, $c = 2.2521$
- Slope, $b = 0.0006$

Simple linear regression (cont.)

Case III: Centred **x**-variable and centred y_2-variable



The estimated regression parameters

- Intercept, $c = 0$
- Slope, $b = 0.0006$

Simple linear regression (cont.)

LR, prediction

We obtain another observation z of the explanatory variable, but not the response

- The system that generates this observation is the same

We want to predict the value of the response (as if we had measured it), given z

- This is easily done by substituting z for x in the learned model
- (Equivalent to reading it from the plot of the regression line)

$$\widehat{y} = \underbrace{\overline{y}}_{b_0} + (z - \overline{x})\,\widehat{b} \quad \text{(predicted composition)} \tag{30}$$

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

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